University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Liu, Chunhua

Title:
Word Associations as a Source of Commonsense Knowledge

Date:
2023-12

Persistent Link:
https://hdl.handle.net/11343/341142

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

Word Associations as a Source of Commonsense Knowledge

by

Chunhua Liu

ORCID: 0009-0009-6778-0172

A thesis submitted in total fulfillment for the

degree of Doctor of Philosophy

in the

Faculty of Engineering and Information Technology

School of Computing and Information Systems

THE UNIVERSITY OF MELBOURNE

March 2024

THE UNIVERSITY OF MELBOURNE

Abstract

Faculty of Engineering and Information Technology
School of Computing and Information Systems

Doctor of Philosophy

by Chunhua Liu
ORCID: 0009-0009-6778-0172

Commonsense knowledge helps individuals naturally make sense of everyday situations and is important for AI systems to truly understand and interact with humans. However, acquiring such knowledge is difficult due to its implicit nature and sheer size, causing existing large-scale commonsense resources to suffer from a sparsity issue. This thesis addresses the challenge of acquiring commonsense knowledge by using word associations, a resource yet untapped for this purpose in natural language processing (NLP). Word associations are spontaneous connections between concepts that individuals make (e.g., *smile* $\rightarrow$ *happy*), reflecting the human mental lexicon. The aim of this thesis is to complement existing resources like commonsense knowledge graphs and pre-trained language models (PLMs), and enhance models' ability to reason in a more intuitive and human-like manner.

To achieve this aim, we explore three aspects of word associations: (1) understanding the relational knowledge they encode, (2) comparing the content and utility for NLP downstream tasks of large-scale word associations with widely-used commonsense knowledge resources, and (3) improving knowledge extraction from PLMs with word associations.

We introduce a crowd-sourced large-scale dataset of word association explanations, which is crucial for disambiguating multiple reasons behind word associations. This resource fills a gap in the cognitive psychology community by providing a dataset to study the rationales and structures underlying associations. By automating the process of labelling word associations

with relevant relations, we demonstrate that these explanations enhance the performance of relation extractors.

We conduct a comprehensive comparison between large-scale word association networks and the `ConceptNet` commonsense knowledge graph, analysing their structures, knowledge content, and benefits for commonsense reasoning tasks. Even though we identify systematic differences between the two resources, we find that they both show improvements when incorporated into NLP models.

Finally, we propose a diagnostic framework to understand the implicit knowledge encoded in PLMs and identify effective strategies for knowledge extraction. We show that word associations can enhance the quality of extracted knowledge from PLMs.

The contributions of this thesis highlight the value of word associations in acquiring commonsense knowledge, offering insights into their utility in cognitive psychology and NLP research.

Dedicated to my loving parents and brother . . .

Declaration of Authorship

I, Chunhua Liu, declare that this thesis titled, 'Word Associations as a Source of Commonsense Knowledge' and the work presented in it are my own. I confirm that:

- the thesis comprises only my original work towards the PhD degree except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Signed:

Date:

Preface

Chapter 3 is largely based on our work published in the following paper:

> Chunhua Liu, Trevor Cohn, Simon De Deyne and Lea Frermann. 2022. WAX: A New Dataset for Word Association eXplanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 106–120, Online only.

Chapter 4 is largely based on our work published in the following paper:

> Chunhua Liu, Trevor Cohn and Lea Frermann. 2021. Commonsense Knowledge in Word Associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CONLL 2021)*, pages 481–495, Online only.

Chapter 5 is largely based on our work published in the following paper:

> Chunhua Liu, Trevor Cohn, and Lea Frermann. 2023. Seeking Clozure: Robust Hypernym extraction from BERT with Anchored Prompts. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 193–206, Toronto, Canada. Association for Computational Linguistics.

Acknowledgements

I am immensely thankful to my supervisors, Trevor and Lea, whose wisdom, patience, and enthusiasm have been the cornerstones of my PhD journey. It is your continuous support and insightful feedback that have made this path both enjoyable and enlightening. Trevor's infinite ideas and zeal for research, coupled with his out-of-the-box thinking, have continually inspired me. Lea, a true model and compass, exudes care, energy, and dedication, deeply influencing me. Their combined guidance has indelibly shaped me in my research journey. Beyond that, they have exemplified how to lead a fulfilling life, tackle obstacles with optimism, and become a better person. Their influence has been significant, often prompting me to wonder, "What would Trevor and Lea do?" in various scenarios.

I am thankful to Simon De Deyne, co-author of Chapter 3, for his invaluable insights. Our stimulating conversations and his valuable feedback have been highly beneficial to my growth during my PhD. Working with such a kind and talented researcher, who introduced me to the fascinating intersection of cognitive psychology and NLP, has marked an important step in my academic journey.

I would like to extend my sincere thanks to my committee chair, Lars, for your pertinent and constructive suggestions at every milestone.

I am also grateful for the financial support provided by the China Scholarship Council and the University of Melbourne, as well as the Conference Travel Scholarship from the School of Computing and Information Systems.

Special thanks to Fei Liu, who has been more than a true friend to me—a guiding elder brother and a mentor. Fei, you introduced me to Trevor in 2018, sowing the seeds for my PhD journey. I am grateful for your constant presence during the lows of my life, your understanding, support, and guidance, which extended beyond research into showing me how to be a kind, helpful and thoughtful person.

A collective shout-out to the amazing group of PhDs, post-docs, and faculty members that I have been surrounded by during my PhD journey at UniMelb. This vibrant and talented group has brought joy, laughter, and indispensable support into my life. My profound appreciation goes to Aili, Biaoyan, Yulia, and Miao for the extensive conversations about

both the complexities of research and life, which have consistently influenced me. Thanks to Kemal, Shima, Yuxia, Zenan, Haonan, Xudong, Fajari, Yitong, Nitika, Thinh, Lyra, Gisela, Viktoria, Zheng Wei, Sukai, Takashi, Jinghui, Jingrui, Jun, Mel, Rui, Zhuohan, Jiyu, Brian, Yilin, Rongxin, Julie, and Kevin. Your willingness to help, to share moments, and to walk alongside me has immeasurably enriched my journey. Thank you, Tim Baldwin, Eduard Hovy, Charles Kemp, Qiongkai Xu, Ekaterina Vylomova, Jeyhan Lau, Daniel Beck, and Karin Verspoor, for shining like distant stars in my academic sky, influencing me and showing kindness throughout my journey. My mind floods with memories of all the moments we have shared together. They are beautiful!

I would like to extend my thanks to the researchers who helped provide resources during my journey: Tess Fitzpatrick, Yanlin Feng, and Peifeng Wang. My gratitude also goes to my old friends Bohan, Hainan, Zhaoyang, Gongbo, and Dong Yu for their invaluable assistance.

Thank you, Kabir, for infusing my life with love, laughter, and care. Our deep and joyful conversations have been a wellspring of comfort and happiness. I deeply cherish our every shared moment and eagerly look forward to a bright future together, holding our hands. Your presence is a priceless gift beyond words.

And, of course, to my family - my parents and my brother - your enduring love and support have been the foundation upon which I have built myself as a positive, kind, and enthusiastic person.

Last but not least, thanks to the universe, the air, the clouds, the sunshine, and the wonders of nature, especially the plants, trees, and water. Engaging in conversation with you fills me with energy and inspiration.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

As humans, we use intuition and commonsense to navigate in our daily world (Kahneman, 2012, Davis and Marcus, 2015). Consider a simple scenario:

> *Jordan wanted to tell Tracy a secret, so Jordan leaned towards Tracy. Why did Jordan do this?* (Sap et al., 2019b)

Intuitively, we understand Jordan's action is to make sure no one else could hear. It is an unspoken rule that secrets are whispered, not shouted, requiring quiet and hidden exchanges. This understanding stems from a rich repository of commonsense knowledge that we, as humans, possess (Liu and Singh, 2004).

Similarly, for AI systems to understand and accommodate human needs, they require commonsense reasoning ability (Lenat et al., 1985, Davis and Marcus, 2015, Davis, 2023). This necessitates the acquisition of commonsense knowledge, a shared understanding of everyday situations and events, such as "A secret is meant to be kept unknown by many people". Such knowledge is crucial for developing intelligent systems capable of human-like understanding and reasoning. In the field of natural language processing (NLP), commonsense knowledge is fundamental to various tasks like question answering (Gordon et al., 2012), reading comprehension (Norvig, 1987), sentiment analysis (Ma et al., 2018), and machine translation (Bar-Hillel, 1960, Nirenburg, 1989), facilitating the generation of coherent and contextually accurate responses.

However, acquiring commonsense knowledge is a challenge, owing to its inherent nature of being **big**, the estimated size of human commonsense ranges from 0.5×10^9 to 3.4×10^9 bits (von Neumann, 1958, Landauer, 1986); and **implicit**, as it is often shared, inherited and unnecessary to mention explicitly during conversations (Grice, 1975, Van Durme, 2009).

Currently, commonsense knowledge is primarily acquired from three sources: text corpora, pre-trained language models (PLMs), and human subjects. Text corpora, as accumulations of human experience, are considered as a source encoding commonsense knowledge. Researchers have tapped into these vast text collections, utilizing NLP methods to extract commonsense knowledge (Gordon et al., 2010, Tandon et al., 2014). However, due to the implicitness property, this approach is affected by *reporting bias* (Gordon and Durme, 2013). This bias is a discrepancy that arises between real-world knowledge and its representation in text, as more salient or frequent information is over-represented, while obvious facts are underrepresented. For instance, the action of "thinking", although frequent in human behaviour, is often too obvious to be explicitly mentioned in texts, while "dying", a singular event for every individual, appears quite frequently (Shwartz and Choi, 2020). This bias can cause models to struggle with implicit aspects of human understanding.

PLMs, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), are neural networks trained on massive text data and have achieved advancements in diverse NLP tasks. PLMs implicitly capture and represent language knowledge within their learned parameters, offering another resource of acquiring commonsense knowledge (Petroni et al., 2019, Davison et al., 2019). However, they too are subject to reporting bias (Shwartz and Choi, 2020), as their training is based on text corpora, and understanding the extent of knowledge they encode requires further exploration (Liu et al., 2023).

Commonsense knowledge in human subjects is often harnessed through a more direct method, typically involving the elicitation of their shared knowledge to compile it into commonsense knowledge graphs (Lenat et al., 1985, Miller, 1995, Singh et al., 2002, von Ahn et al., 2006). One of the most comprehensive and widely-used knowledge graphs is `ConceptNet` (Speer et al., 2017), which represents concepts (single words or phrases)[1] as

[1] In this thesis, we use 'word' and 'concept' interchangeably for ease of reading, acknowledging their distinct meanings and varied definitions (De Saussure et al., 1916, Pinker, 2003).

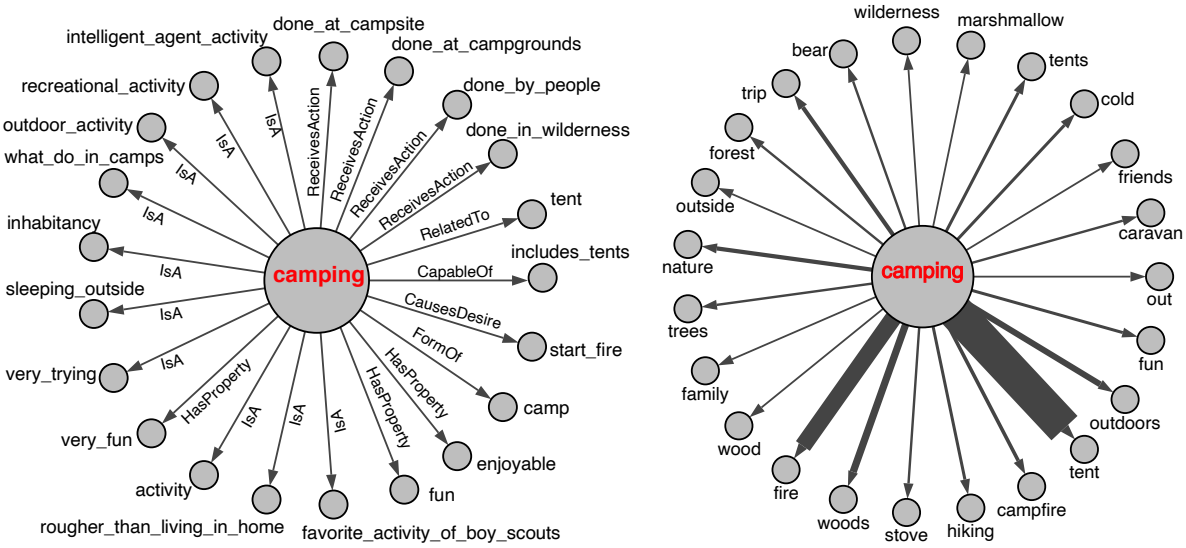


Figure 1.1 Concepts associated with "camping" from a commonsense knowledge graph (ConceptNet, left) and a word association network (SWOW, right). In contrast to ConceptNet, word associations are not labelled with relations, however, implicit connections like *camp* → *marshmallow* are elicited. Thicker lines denote stronger associations.

free text nodes and their relations as edges (e.g., camping, CausesDesire, start fire). See Figure 1.1 (left) for an illustration of the graph structure. Commonsense knowledge graphs like ConceptNet have the advantage of being highly structured and relationally interpretable and have been widely used to support commonsense understanding and reasoning (Lin et al., 2019a, Chen et al., 2019, Lv et al., 2020). However, even they suffer from high sparsity and incompleteness (Li et al., 2016, Storks et al., 2019, Malaviya et al., 2020) due to the difficulty of acquiring commonsense knowledge. Although subsequent efforts attempted to augment ConceptNet with more knowledge from text-corpora (Nguyen et al., 2023) or crowd-sourced knowledge (Sap et al., 2019a), the sparsity issue of commonsense knowledge graphs remains.

The constraints and biases present in current resources raise the question: How can we efficiently and effectively obtain large-scale commonsense knowledge? To overcome these limitations, we must seek methods that can capture the breadth and depth of human understanding more effectively. This thesis posits that word associations represent an untapped yet potent resource, capable of eliciting commonsense knowledge and augmenting existing NLP resources.

Word associations are spontaneous connections that humans form between words or concepts, rooted in cognitive psychology. Typically, a participant is presented with a cue word (e.g., *blink*) and asked "what is the first word that comes to your mind", to which they might respond with *eye* or *wink*. These associative connections are generated quickly by the brain, stemming from our intuition system (Kahneman, 2012). As an example presented in Figure 1.1 (right), these associations provide a window into the human mental lexicon, revealing meaningful connections grounded in human experience. Associations serve as a simple yet powerful method for extracting hidden or implicit knowledge, which is difficult to obtain through direct questioning or easily accessible sources like text corpora (Szalay and Deese, 1978, Mollin, 2009, De Deyne et al., 2016c). For example, people naturally associate *blink* with *quick*, despite their infrequent co-occurrence in text, uncovering underlying connections and knowledge that enrich our understanding of human cognition and concept comprehension.

With the emergence of crowd-sourcing, large-scale word associations (Nelson et al., 2004, De Deyne et al., 2019) are created by collecting associations from a large amount of participants, whose responses are aggregated, producing dense and diverse responses connected with cues. This is appealing, as responses elicited by each cue word encapsulate a comprehensive, experience-based understanding of the cue, stemming from a sizable and varied population. Consequently, the derived resource embodies large-scale collective and basic knowledge alongside implicit associations (e.g., needle and hurt), potentially augmenting the currently sparse commonsense knowledge graphs.

Large-scale word associations have the innate characteristic of eliciting implicit knowledge from human minds. Previous studies have demonstrated the value of word associations in predicting human similarity judgments (De Deyne et al., 2016d) and in capturing knowledge beyond text corpora, such as visual and affect information (De Deyne et al., 2021). However, unlike relation-labelled commonsense knowledge graphs such as `ConceptNet`, word associations do not inherently come with relational types (see Figure 1.1), posing a challenge to their interpretability and potentially limiting their direct utility in relation reasoning. This raises questions about the practical applications of incorporating word associations

into NLP models. The purpose of this thesis is to better understand word associations and explore their utility in NLP models and tasks.

1.1 Aim and Research Questions

This thesis studies the problem of commonsense knowledge acquisition via large-scale word associations to complement existing resources, such as commonsense knowledge graphs and pre-trained language models. Our aim is to understand the driving force of word associations and study its utility as a commonsense knowledge resource in NLP. The main hypothesis of this thesis is the following:

With the ability to effectively capture collective knowledge, large-scale word associations encode commonsense knowledge (§3), serve as an alternative to large-scale commonsense knowledge resources like ConceptNet (§4), and consequently improve the commonsense reasoning ability (§4) and knowledge extraction (§5) from pre-trained language models.

We evaluate the hypothesis with the following research questions.

- What relational knowledge is encoded in word associations? Existing large-scale word associations (Kiss et al., 1973, De Deyne et al., 2019) do not include relation labels to indicate why two words are associated, rendering a deeper understanding of what is driving word associations difficult. Identifying the relational knowledge within these associations is crucial for understanding their role in commonsense knowledge.

- Can large-scale word associations improve performance on downstream commonsense reasoning tasks? How does the knowledge they encode differ from the existing largest commonsense knowledge graph `ConceptNet`? These questions aim to evaluate the utility of large-scale word associations in commonsense required tasks, identifying knowledge overlaps and gaps with existing graphs and suggesting new approaches for commonsense knowledge acquisition.

- How to better understand and robustly extract implicit relational knowledge encoded in pre-trained language models? Can word associations contribute to the knowledge extraction? This inquiry shifts focus from knowledge within word associations to its dynamic with pre-trained language models, which have advanced many tasks in NLP by implicitly encoding commonsense knowledge. The key issues are the robust extraction of knowledge from pre-trained language models and understanding how this extracted knowledge differs from human mental representations. Exploring these aspects offers insights into the alignments and divergences between the knowledge in word associations and pre-trained language models.

1.2 Contributions

This thesis underlines the significance of word associations in the field of NLP as a novel method of acquiring commonsense knowledge and makes three main contributions:

- We present a large-scale resource, WAX, which provides 19K explanations for word associations and relation labels reflecting the high-level structure (Chapter 3). This resource aims to fill a gap in the cognitive psychology community by providing a dataset to study the rationales and structures underlying association. Through modelling and analysing the data, we found that word associations are predominantly driven by semantic relations. Meanwhile, we automate the process of labelling word associations with relevant relations, demonstrating that explanations can disambiguate multiple reasons for word associations, thereby enhancing the performance of relation extractors. Our dataset and initial models pave the way for future work on labelling word associations.

- We highlight the distinction between large-scale word associations (`SWOW`) a and commonsense knowledge graph (`ConceptNet`) through a comprehensive comparison of their structures, knowledge content, and utility for commonsense question-answering tasks (Chapter 4). This is the first study to successfully demonstrate the promise of word associations as a resource for commonsense knowledge.

- We introduce a diagnostic framework to more effectively understand and extract the implicit knowledge encoded in pre-trained language models, and to identify effective strategies for handling challenging scenarios such as abstract and rare concepts. We explore the potential of incorporating word associations, showing that they can further enhance the quality of extracted knowledge. This reflects the unique value of the knowledge encoded in word associations and its complementary role with pre-trained language models.

Collectively, these contributions push the boundaries of interpretable word associations, providing valuable insights into their utility for future research in cognitive psychology and NLP.

1.3 Thesis Structure

Chapter 2 Literature Review provides an overview of the literature on semantic knowledge, examining three dimensions: organisation, acquisition, and application. We start by reviewing the organisation of semantic knowledge in human minds, delineating different theories of organising semantic memory. Building on these theories, we introduce the construction of different large-scale knowledge graphs used in this thesis, focusing on two key graphs: `ConceptNet` and `SWOW`. Next, we provide an overview of the computational models for representing semantic knowledge, including pre-trained language models and knowledge graph representations, in terms of their evolution, architectures, and training objectives. In addition, we discuss various NLP tasks that provide the benchmark applications for this thesis, including semantic link prediction and commonsense question answering, by reviewing the existing datasets and models.

Chapter 3 Explanations and Relations in Word Associations focuses to better understand the knowledge encoded in word associations. To reveal the underlying reasons of why two words are associated, we propose a framework to recover the context of linking two associated words from humans. More specifically, we collect large-scale word association explanations from crowd-sourcing and label them with associative relations. This novel data

set enables us to analyse the distribution of associative types and verify our hypothesis that context in the explanations can alleviate disambiguation between two words and thus improve machine learning models' ability to predict their relations. We design experiments to compare models' ability to predict word association relation types when explanations are presented or not. By assessing a series of state-of-the-art supervised relation classification models, we observed large improvements when explanations are incorporated, compared against only two words used as inputs, demonstrating the utility of word association explanations. This study demonstrates that comprehensive and broad relation information is encoded in word associations, motivating us to investigate the utility of large-scale word associations in NLP in Chapters 4 and 5.

Chapter 4 Commonsense Knowledge in Word Associations systematically compares two commonsense knowledge graphs (KGs): `ConceptNet` and `SWOW` in terms of their graph structure, knowledge capacity, and their benefits to downstream tasks. Although our analysis demonstrates substantial differences between the two KGs, surprisingly, similar improvements are observed when incorporating them into commonsense question answering models. Importantly, our empirical results highlight that the commonsense reasoning ability of pre-trained language models is improved by incorporating knowledge from `SWOW`. Furthermore, our analyses suggest that `SWOW` serves as an alternative commonsense knowledge graph of `ConceptNet`. This motivates us to study the utility of `SWOW` as a tool for accessing semantic knowledge from pre-trained language models, which we explore in detail in Chapter 5.

Chapter 5 Robust Hypernym Extraction from BERT with Anchors and Word Associations In Chapter 5, we present a diagnostic framework to extract and analyse the knowledge with pre-trained language models. This framework, which integrates corpus mining methods with pre-trained language models, features a new prompt optimisation technique. This technique uses 'anchors, automatically extracted sibling concepts, to enhance and direct prompt formulation, improving model prompting effectiveness. As a case study, we use this framework to examine hypernym knowledge encoded in BERT through a set of diagnostic probing tasks, in which we assess model capacity under several conditions, such as abstrac-

tion and frequency of concepts, and consistency across paraphrased contexts. Additionally, we explore the use of word associations as an external tool for acquiring anchors and compare it to anchors automatically mined from pre-trained language models. Our empirical results show that word associations can improve the effectiveness and robustness of extracting hypernyms from pre-trained language models, showing that word associations improve knowledge accessibility through prompting and provide knowledge beyond pre-trained language models.
Chapter 6 Conclusions provides a summary of our thesis, including our key findings and their implications. In addition, we discuss potential future directions for our research.

Chapter 2

Literature Review

This chapter provides a comprehensive overview of the literature on the representation and acquisition of semantic knowledge, the source of commonsense knowledge. We introduce the primitives of semantic knowledge and its organization theories from a cognitive psychology perspective in Section 2.1.1, with a focus on semantic networks in Section 2.1.2. Next, we provide an overview of computational representation techniques for semantic knowledge in Section 2.3, focusing on two main areas: pre-trained language models (Section 2.3.1) that learn semantic knowledge from large text-corpora, and knowledge graph representations (Section 2.3.2) that focus on representing structured semantic knowledge. We then review relevant frameworks for gathering commonsense knowledge from humans (Section 2.2), including techniques of eliciting and organizing knowledge graphs, and compare their pros and cons, with a focus on two large-scale knowledge graphs: `ConceptNet` and `SWOW`. Then, we introduce the tasks relevant to this thesis (Section 2.4), including semantic relation learning in Section 2.4.1 and commonsense question answering in Section 2.4.2. By presenting this comprehensive overview, we aim to provide a deeper understanding of the organization, acquisition and representation of semantic knowledge across different sources and offer motivations for our study.

2.1 Semantic Memory

> "You pile up associations the way you pile up bricks. Memory itself is a form of architecture." — *Louise Bourgeois*

As an integral part of human cognition, memory equips humans with the ability to remember and learn from the past, link the past with the present by association or reasoning, and predict the likely events in the future. Different types of memories are involved in the process of human understanding (Szymanski and Duch, 2007). The most relevant one for this thesis is the *semantic memory*, which stores the general knowledge that people have experienced and learned in the long term (Quillian, 1966, Tulving, 1972).

Semantic memory can be considered a summary of a person's whole living experience and knowledge, allowing humans to understand the basic facts and general knowledge of the world and use them to link with new information (Pitel et al., 2014, Roediger et al., 2017). Human understanding of the world would be impossible without semantic memory (Anderson, 1995).

Given its fundamental importance, researchers have devoted substantial effort to understanding how humans organise, retrieve, and represent knowledge in semantic memory. The key underlying elements are the concepts and their connections (Quillian, 1967, McRae and Jones, 2013). Next, we will introduce the two elements and review two theories in organising semantic memory in human cognition.

2.1.1 Concepts and their Links

Semantic memory encompasses two vital elements: concepts and their links (McRae and Jones, 2013). A concept is an abstract mental representation of objects, events, or relations based on our accumulated knowledge and experiences of the world (Markman and Rein, 2013).

The structuring of concepts in human memory is systematic, not random. Studies have delved into its organization and efficient retrieval. Two key theories, **semantic networks**

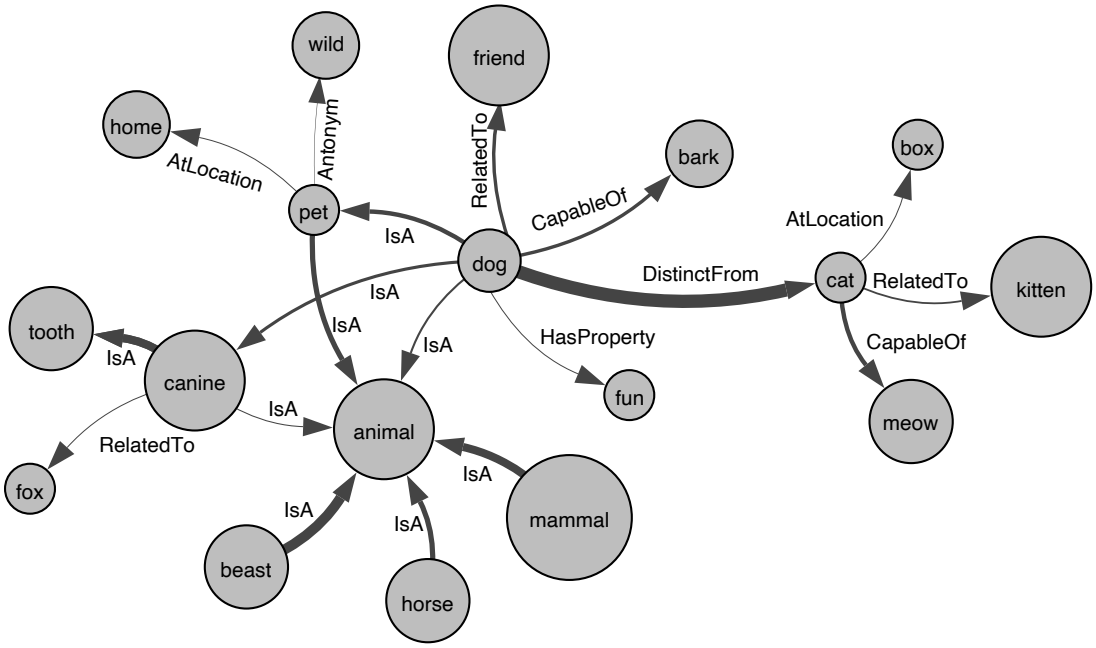


Figure 2.1 A semantic network diagram where nodes symbolise concepts, and edges indicate (a) direction via arrows, (b) semantic relationships via textual labels like ATLOCATION, and (c) strength through edge width, with wider edges signifying stronger connections. Note: this figure is generated from a combined source of `ConceptNet` and `SWOW`; the nodes and directional indicators are shared elements between `ConceptNet` and `SWOW`. The relation labels come from `ConceptNet`, while the edge strengths are derived from `SWOW`.

and **feature models**, have been proposed to simulate the organization of human semantic memory using computers.

A semantic network (Collins and Quillian, 1969, Collins and Loftus, 1975) is a graphical representation of concepts and their connections, with concepts as nodes and their connections as edges. The meaning of a concept can be captured by its neighbours, which are connected via various types of link. These links can be classified into three types (Collins and Loftus, 1975): (a) direction, i.e., the original versus target of the link, which points from one concept to another; (b) relation types, which reveal the specific semantic relationships between two concepts; and (c) strengths, which measure the closeness of the connection between two concepts. An illustration of this can be seen in Figure 2.1. The semantic relatedness between concepts can be determined either by the distance between them in the graph or by the

Concept	Feature	Production Frequency	Feature	Production Frequency
dog	has fur_hair	18	does wag its tail	6
	does bark	17	is a companion	6
	has a tail	16	has paws	5
	is a pet	15	does need exercise	4
	is an animal	14	does play	4
	is man's best friend	14	is friendly	4
	has four legs	10	is loyal	4
	has teeth	9	is trained	3
	has legs	8	does run	3
	is a mammal	8	is cute	2

Figure 2.2 An example of feature representation for the concept *dog* from 30 participants (Devereux et al., 2014). The Production Frequency is the number of participants who listed the feature.

strength of the connecting edges, which can be quantified by the number of shared properties between the two concepts.

In contrast, feature models (Smith et al., 1974, McRae et al., 1997) assume that concepts are represented by a collection of discrete semantic features. These features are commonly elicited using the property generation task (i.e., feature listing) by asking multiple participants to generate a list of features for given concepts (McRae et al., 1997). An example of such a feature representation is illustrated in Figure 2.2. Within this type of model, the typicality of a feature is quantified by its production frequency, which is the number of participants who list the same feature for a given concept. Additionally, the degree of relatedness between different concepts can be quantified based on the overlap of these features.

Feature models are easy to interpret and suitable for representing basic properties relationships; however, semantic networks have more flexibility in representing complex relationships between concepts that allow them to grow rapidly (Steyversa and Tenenbaum, 2005). Additionally, property knowledge in feature models can be implemented in semantic networks (Hollan, 1975). Therefore, semantic networks have been widely used across many fields, such as artificial intelligence (Lehmann, 1992), social science (Segev, 2021), and biomedical research (Bales and Johnson, 2006). In this thesis, we will focus on the semantic network and introduce its development and categories below.

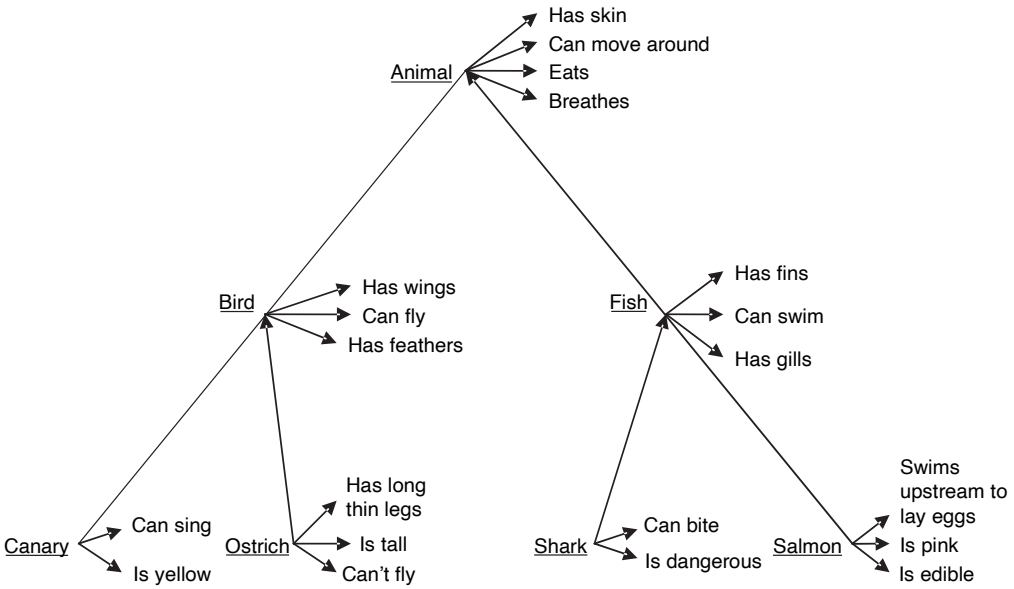


Figure 2.3 Hierarchical semantic network, taken from (Collins and Quillian, 1969). Reprinted by (Balota and Coane, 2008).

2.1.2 Semantic Networks

A semantic network represents a graph $\mathcal{G}$ with concepts $\mathcal{C}$ as nodes and links between concepts as edges $\mathcal{E}$ (see Table 2.1 for a full list of notations). These concepts can be symbolised by single words (e.g., *animal*) or phrases (e.g., *sense danger*) to convey their meanings effectively. The edges $\mathcal{E}$ in $\mathcal{G}$ represent links between these concepts. The meaning of a concept within $\mathcal{G}$ is characterized by its connected concepts and the nature of the links connecting them. An edge e can have link labels (e_ℓ) representing: (a) a relation type r, such as ISA or USEDFOR; (b) a direction d between concepts; or (c) a weight w denoting edge strength. This is illustrated in Figure 2.1.

The development of semantic networks has a long history, beginning with the work of Collins and Quillian (1969). They introduced a hierarchical network that represents the organisation of concepts in semantic memory. As illustrated in Figure 2.3, general concepts are at the top, descending to specific ones at the bottom, interconnected through the hierarchical relation ISA. Each concept (e.g., *bird*) possesses attributes (e.g., *has skin*, *can fly*, or *is*). These attributes are stored at the most relevant hierarchical level. For example, the property *has wings* is anchored at the bird node, eliminating the need to repeat it for specific

Notation	Description
c_1	A head concept or a cue word in word associations
c_2	A tail concept or an associated response word in word associations
r	A relation type
w	A weight value
d	A pointer or direction from one concept to another
$r \in \mathcal{R}$	Relation set
$c \in \mathcal{C}$	Concept set
(c_1, r, c_2)	A triple of head, relation and tail
$(\mathbf{c}_1, \mathbf{r}, \mathbf{c}_2)$	Embedding of head, relation and tail
$\xi \in \mathcal{E}$	Edges in edge set
$\xi_\ell = \{r, w, d\}$	Edge labels with possible types r, w, or d
$\mathcal{G}$	A knowledge graph
θ	Model parameters
$d(\cdot)$	Distance metric in specific space
$\mathbb{R}^d$	d dimensional real-valued space
$\mathbf{x}$	A sequence of input tokens $\mathbf{x} = [x_1, x_2, \cdots, x_T]$
$\mathbf{X}$	A sequence of input embedding $\mathbf{X} = [X_1, X_2, \cdots, X_T]$
X, Y, Z	Variables or placeholders in a text template
$\mathcal{L}$	Loss function

Table 2.1 Notations with descriptions.

bird types, such as canaries or ostriches. This work marked the first attempt to represent semantic knowledge as a network. The authors found a positive correlation between the concept hierarchy and the time individuals take to verify a statement's truth. For example, they found that individuals took longer to determine that "A canary is an animal" than "A canary is a bird". This delay aligns with the hierarchy of the memory structure, where "animal" is positioned above "bird", and therefore requires more inference time when validating the facts. Although this hierarchical model offers efficient knowledge storage, it requires more cognitive effort during retrieval. Furthermore, the model does not fully explain certain behavioural phenomena, such as typicality effects (e.g. why individuals respond faster to "robin is a bird" than "ostrich is a bird") and the latencies of false negative sentences (e.g., why individuals are slower to reject "butterfly is a bird" than "dolphin is a bird").

To tackle those issues, Collins and Loftus (1975) extended the hierarchical model with the spreading activation theory, which organises the concepts according to semantic similarity

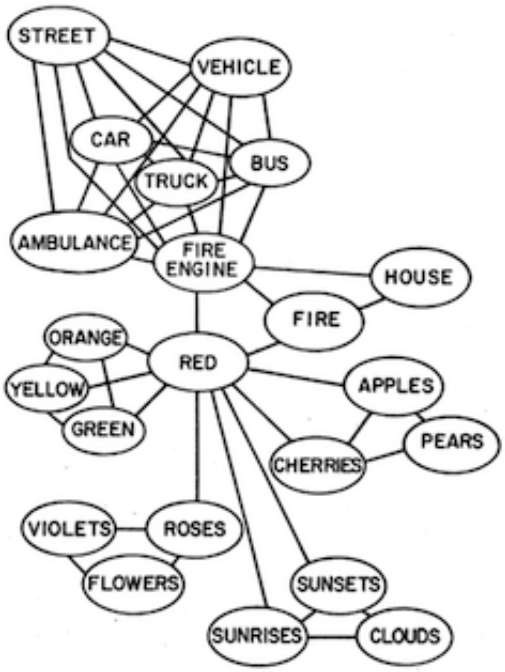


Figure 2.4 Illustration of spreading activation theory from Collins and Loftus (1975), which organises concepts based on semantic similarity. A shorter link denotes a higher similarity.

instead of a hierarchical structure, and no longer distinguishes concepts from features. Semantic similarity refers to an aggregation of all paths that connect two concepts in a network. The core idea is that the more common properties two concepts share, the closer they are in terms of semantic similarity. For example, in Figure 2.4, the different kinds of vehicles (e.g., *car* and *truck*) are highly interconnected and closer than other nodes (e.g., *car* and *red*) in semantic meanings. In this theory, the retrieval of semantic memory is triggered through activation. This means that when we think about a concept, it activates the related concepts in our memory, which in turn activates more related concepts, forming an activated network of interconnected nodes. In this process, the closer two concepts are semantically, the easier and faster they are to be retrieved. Therefore, using the semantic similarity can successfully explain the typicality effects and the latencies of false negative sentences; that is strongly connected concepts are more typical (e.g., *robin* and *bird*) and thus are associated faster than other weakly linked nodes (e.g., *ostrich* and *bird*).

The above two works pioneered later work in two directions. The first involves characterising concepts with explicit relation-linked nodes. This idea has been used to construct

relational semantic networks, such as the WordNet (Miller, 1995) and `ConceptNet` (Liu and Singh, 2004), with enriched relation types and concepts to enhance the flexibility of representing complex relationships among concepts. This research line has been extensively developed in the domain of natural language processing, and the created semantic networks have been applied to improve NLP downstream tasks (Moussallem et al., 2018, Chen et al., 2018, Lin et al., 2019a, Zhou et al., 2020a). Another direction is the **associative network**, which characterizes a concept through its (weighted) connections to other concepts in the network. This coincides with the idea of "free association", an effective task that is used in psychology and cognitive science to study the associations between concepts that people have in their mental lexicon. In contrast to relational semantic networks, which can be considered explicit knowledge stores, associative networks serve as maps of the human mental lexicon and may be used as implicit knowledge stores. However, the extent to which these differences in theoretical network structures reflect practical differences in curated large-scale semantic networks, and their benefits to NLP downstream tasks remain unknown. This serves as the high-level motivation for our thesis.

To provide a better understanding of how various semantic networks are structured, we next introduce representative networks for each of them and their properties.

2.1.2.1 Relational Semantic Networks

A relational semantic network ($\mathcal{G}$) is a graph composed of concepts (nodes) and relations (labelled edges). Edges $\mathcal{E}$ comprise a set of relational triples (c_1, r, c_2). Different semantic networks have been constructed, focusing on different sets of relation types ($\mathcal{R}$) and network structures. We introduce two representative semantic networks that are used in our thesis.

Princeton WordNet (Miller, 1995, Fellbaum, 1998, 2010) is a large lexical semantic network that interrelates concepts that are similar in meaning through a set of lexical and semantic relationships. WordNet groups concepts into sets of synonyms called synsets, which provide short definitions and usage examples. These synsets constitute the inventory of nodes of a graph, where edges are labelled with various types of relationships, including hypernym/hyponym (broader or narrower "IsA" relations), antonym, and other relations.

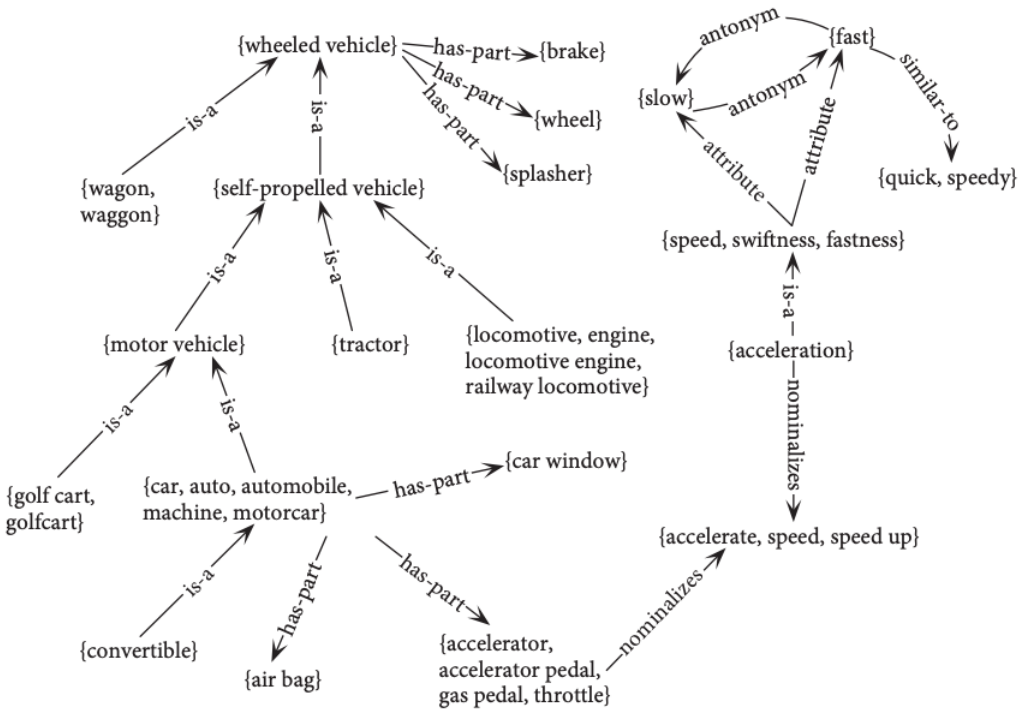


Figure 2.5 An excerpt of WordNet semantic network. Figure from Navigli (2022).

The most recent WordNet (3.0) (Fellbaum, 2010) contains over 117,000 synsets, comprising over 81,000 noun synsets, 13,600 verb synsets, 19,000 adjective synsets, and 3,600 adverb synsets. WordNet is a hierarchical semantic network with general concepts at the top and more specific concepts at the bottom. This is inspired by Collins and Quillian (1969)'s hierarchical model but without the strict tree structure.

The example in Figure 2.5 illustrates the hierarchical structure of WordNet. The hierarchical structure of WordNet enables similarity between synsets to be measured in several ways, based on connections (Wu and Palmer, 1994, Lin, 1998). WordNet has been extended to multiple languages (Ordan and Wintner, 2007, Fellbaum et al., 2006, Bond and Paik, 2012) and its English version has been continually updated (McCrae et al., 2019, 2020). It has been extensively used in various applications, such as textual entailment (Silva et al., 2018), natural language generation (Juraska et al., 2018) and sentiment analysis (Wang et al., 2018b). WordNet was created by experts and thus contains high-quality data. In Chapter 5, we use it as an external knowledge base to examine the quality of knowledge extracted from language models (PLMs), which will be introduced in Section 2.3.1.

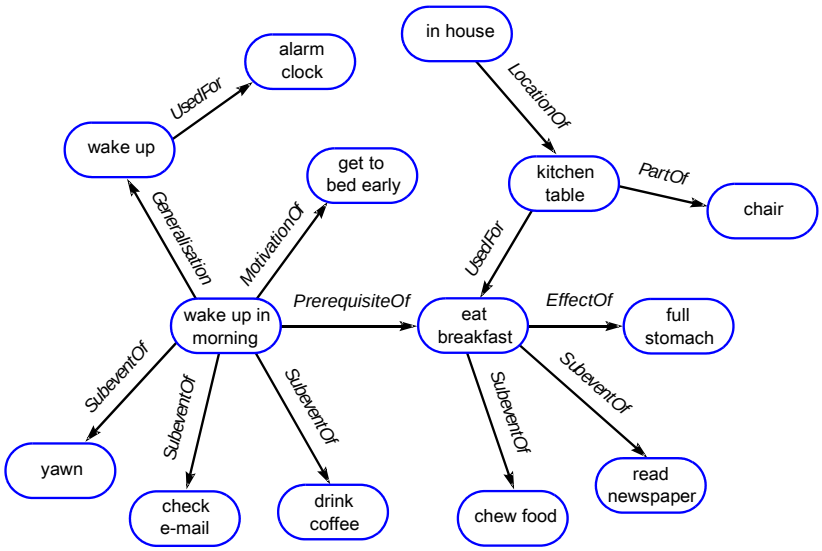


Figure 2.6 Excerpt from ConceptNet's semantic network (Liu and Singh, 2004).

ConceptNet (Liu and Singh, 2004) is a large semantic network that focuses on general or commonsense knowledge, referring to the basic level of practical knowledge concerning everyday situations and events that are commonly shared among most people. Instead of using synsets as nodes, ConceptNet uses natural language words or phrases to represent nodes and contains more semantic relations[1] than WordNet. Unlike WordNet, which uses a hierarchical structure to organise the graph, ConceptNet relaxes the restriction, enabling more flexibility between concepts to be connected. Figure 2.6 illustrates the structure of ConceptNet. Its construction involves seven major resources, such as crowd-sourced facts from Open-Mind Commonsense Sense (Singh et al., 2002), information extracted from Wiktionary (Table 2.2 in Section 2.2.1.1 provides more details). The most recent version ConceptNet 5.5 (Speer et al., 2017) is a multilingual graph, containing over 21 million edges and 8 million nodes, covering 83 languages. Due to its broad knowledge coverage, it has been used for a wide range of downstream tasks, such as commonsense question answering (Lin et al., 2019a), natural language inference (Wang et al., 2019), and machine translation (Caseli et al., 2010). ConceptNet is used in Chapter 4 as a representative

[1] Section 2.2.1.1 will provide a detailed presentation of these ConceptNet relations.

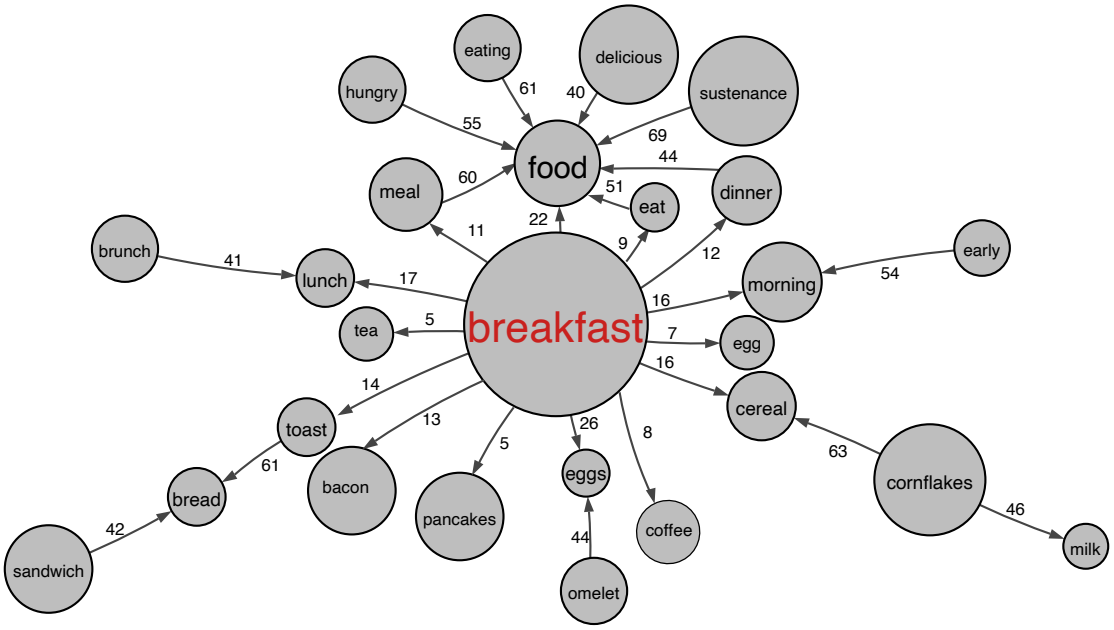


Figure 2.7 Illustration of associations centering on the cue word 'breakfast' from SWOW (De Deyne et al., 2019). The edge arrow indicates the direction of association (from cue to association), and the numbers on the edges indicate the frequency of associations.

engineered resource of commonsense knowledge to study the differences between relational semantic networks and associative networks. We will provide a detailed description of ConceptNet in Section 2.2.

In summary, relational semantic networks explicitly specify the types of relations that connect concepts, which provides good interpretability and enables explainable relation reasoning across concepts. In Chapters 4 and 5, we use relational knowledge from ConceptNet and WordNet, respectively, to assist various tasks.

2.1.2.2 Associative Networks

In an associative network, concepts are connected by spontaneous free association. Unlike the typical semantic network specifying explicit relation labels on the edges, concepts in associative networks are linked with directions and weights, as shown in Figure 2.7. We refer an associative network as a weighted graph. Associative networks encode association strengths and measure the similarity between two concepts with the degree of association overlap.

An associative network can be constructed via the free association tasks, in which participants are presented with a stimulus cue word (e.g., *dog*) and asked to respond with the first words that come to their minds (e.g., *bark, fur, park*). These associations emerge from a wealth of human experiences (Deese, 1966), arising from the consistent exposure of words and concepts over time (Moss et al., 1995, Bower, 2000), thereby mirroring human understanding of these cue words (De Deyne et al., 2013a).

The first word association network was collected by Kent and Rosanoff (1910) on the base of 100 cue words and 1000 subjects to understand the association differences between patients with mental health issues and healthy subjects. Subsequent work (Kiss et al., 1973, Nelson et al., 2004, De Deyne et al., 2019) focused on constructing large-scale word association data bases and extracting meaningful information to understand the structure of human mental representations (De Deyne and Storms, 2008a). The idea is that with the increase of size, an associative network can cover all commonly used words in the language and represent the mental association structure of individuals speaking that language.

The "free" association mechanism and open-ended nature of word associations enable a holistic representation for a concept to be captured. However, the *reasons why* certain associations were elicited are often not reflected in word associations, making understanding their underlying structure difficult.

Association Types One important question about word associations is what drives the associations. Researchers have attempted to assign categories to the (cue, association) pairs via empirical study to answer this question. The conventional approach is to classify associations into three broad categories (Namei, 2004), including paradigmatic, syntagmatic, and clang associations, to represent the "meaning", "use" and "form"-based associations. Following work revise this classification method by providing more fine-grained associative relation types (Fitzpatrick, 2006, Khazaeenezhad and Alibabaee, 2013) and extend it to cover more association types such as phonological (Nissen and Henriksen, 2006, Zareva, 2007) and 'other' to include any other associations (Yu et al., 2011). Among them, Fitzpatrick (2006) propose the most comprehensive inventory of relations, including 4 broad categories with

17 sub-relations. Due to the different foci of cognitive psychologists and NLP researchers, the development of association types deviates from the study of semantic relations in typical semantic networks, although both share many relations. Our Chapter 3 aims to bridge this gap.

In summary, we introduce the fundamental elements of semantic memory and theories of its organisation, focusing on the evolution of semantic networks. Over time, the development of semantic networks has led to two divergent approaches, namely relational semantic network and associative network. The former, which contains interpretable relational knowledge, is well-developed, while the latter is less so. In our thesis, we examine how the relational networks can advance the understanding of the associative networks in Chapters 3 and 4. We also explore to what extent their practical implementations encode similar or different knowledge in Chapter 4.

2.2 Constructing Semantic Knowledge Graphs

Having introduced the theory of using semantic networks to organise semantic memory (Section 2.1.2), we now turn our attention to the practical aspect: the construction of semantic knowledge graphs. Specifically, we will examine two exemplary practical networks, one from the relational network and another from the free associations.

The task of constructing large-scale semantic knowledge graphs is called knowledge acquisition (Ji et al., 2022). The goal is to acquire semantic knowledge by extracting and representing it in a structured form. Typically, such knowledge is either acquired from existing text-corpora resources, such as Web pages (Tandon et al., 2014, 2017) or directly from humans (Lenat et al., 1985, Miller, 1995, Liu and Singh, 2004, von Ahn et al., 2006, Sap et al., 2019a).

Gathering knowledge from text sources offers speed and scalability, but it is subject to reporting bias (Gordon and Durme, 2013). This means that text corpora often highlight significant events while ignoring the obvious daily events, limiting the coverage of commonsense knowledge that can be extracted. On the other hand, obtaining commonsense knowledge

directly from humans offers a way to access knowledge from human minds. However, this approach often demands significant time and effort, sometimes spanning decades. WordNet, for instance, was initiated in 1985 (Miller, 1985), and stabilised to its current version by 2010 (Fellbaum, 2010). Therefore, acquiring semantic knowledge from humans is a complex and challenging task, and it requires careful consideration of several factors.

A key challenge, as mentioned in Chapter 1, is the acquisition bottleneck (Feigenbaum, 1984), referring to the difficulty of obtaining enough commonsense knowledge to approximate what a typical person knows. The difficulty mainly comes from the implicit nature of commonsense knowledge, since humans acquire this knowledge through experiences from the real world, and this knowledge is so obvious or basic that it is rarely accessed consciously, or expressed verbally (Cambria et al., 2011). Thus, how to extract implicit knowledge from text or design suitable priming tasks so that humans can express the knowledge explicitly is one important consideration.

Another challenge is the scale: acquiring the vast amount of commonsense knowledge that humans possess requires significant effort and resources, making it challenging to create a comprehensive knowledge inventory.

To overcome these challenges, many knowledge acquisition systems have been developed. These systems are designed with different priming tasks and purposes, resulting in different knowledge to be captured from different sources. In this section, we emphasise the construction of two types knowledge graphs: commonsense knowledge graphs and word association networks, with a particular focus on the graphs (`ConceptNet` and `SWOW`) that are used in our thesis. We introduce their approaches of graph construction, knowledge content, and discuss their corresponding properties in Sections 2.2.1 and 2.2.2.

2.2.1 Commonsense Knowledge Graphs

Commonsense knowledge graphs aim to capture and structure human knowledge about everyday concepts and events, focusing on the common knowledge that is often taken for granted in everyday conversation and communication. The purpose of commonsense

Figure 2.8 Distribution of sources in English ConceptNet 5.5.

knowledge graphs is to equip computers with the similar knowledge that possessed by humans, enabling them to generate outputs that are more human-like.

Early commonsense knowledge graphs such as Cyc (Lenat et al., 1985) represent commonsense knowledge with high-order logic, which is difficult to integrate into modern NLP models (Gunning, 2018). Recent commonsense knowledge graphs use natural language to represent concepts and their relationships, including ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019a), which are the two large-scale crowd-sourced commonsense knowledge graph. ConceptNet encodes general commonsense knowledge with a diverse range of relationships, while ATOMIC focuses on events and their relationships. While ATOMIC is constructed recently and addresses a specific gap in ConceptNet's coverage of events and their relationships, ConceptNet is a more mature and widely used knowledge graph in the research community regarding general commonsense. Since this thesis is concerned with general commonsense knowledge, we use ConceptNet in Chapter 4 and introduce more details about it below.

2.2.1.1 ConceptNet

ConceptNet (Liu and Singh, 2004) is a large-scale commonsense knowledge graph. Its nodes encode concepts, and edges between the nodes, labelled with the type of relation, indicate how two concepts are related. Nodes in this graph are natural language words and phrases, like *breakfast* or *wake up in the morning*, and edges are from a fixed number of pre-defined relation types, e.g., PREREQUISITEOF, USEDFOR. See Figure 2.6 for a more complete illustration.

Constructing ConceptNet The construction of ConceptNet is a complex and iterative process. The early ConceptNet (Liu and Singh, 2004), which originated from the OMCS Project (Singh et al., 2002), parses templated sentences that express the relationships between concepts and events used in daily activities and converts them into structure triplets. For example, the sentence "The prerequisite of eating breakfast is to wake up in the morning.", is converted into (*wake up in the morning*, PREREQUISITEOF, *eat breakfast*). Since then, ConceptNet has continued to grow by incorporating other languages, and importing other available resources (Speer and Havasi, 2013, Speer et al., 2017). We use the latest version ConceptNet 5.5 (Speer et al., 2017) in this thesis. Now we introduce it in more detail.

Sources To increase the scale, ConceptNet v5.5 integrates multiple existing resources. They can be categorized into: (a) crowd-sourced data from OMCS (Singh et al., 2002) and "Games with a Purpose" (von Ahn et al., 2006); (b) existing expert curated data such as WordNet (Fellbaum, 1998, Bond and Foster, 2013), JMDict dictionary (Breen, 2004) and OpenCyc (Lenat et al., 1985); (c) online web data from Wiktionary and DBPedia (Auer et al., 2007). Noticeably, the largest contributor is Wiktionary, accounting for 86.1% of the total edges. Overall, ConceptNet v5.5 encompasses 8 million nodes and 21 million edges, spanning 85 languages. In this thesis, we focus on the English in ConceptNet, including over 1.5 million concepts. The relative (%) distribution from different sources is depicted in Figure 2.8.

Relation Ontology The relation ontology in the earliest version of ConceptNet (Liu and Singh, 2004) contained twenty-one relations. These included basic semantic relations in WordNet (e.g., SYNONYM, ISA, and *PartOf*), and were extended to include more relations (e.g., CAPABLEOF, DESIRES) based on data collected from OMCS. As its version 5.5, ConceptNet has expanded with a diverse and rich set of semantic relations, comprising 34 core relations, which offer comprehensive coverage of conceptual connections (see Table 2.2). These relations can be either symmetric (e.g., SYNONYM) or asymmetric (e.g., ISA) relations.

While this relation set includes a spectrum that ranges from general relations like RELATEDTO, to specific ones such as USEDFOR, ATLOCATION, and CAUSES. Its own relation

Dimension	Relations
Lexical	FORMOF, DERIVEDFROM, ETYMOLOGICALLYDERIVEDFROM
Similarity	SYNONYM, SIMILARTO, DEFINEDAS
Distinctness	ANTONYM, DISTINCTFROM
Taxonomic	ISA, INSTANCEOF, MANNEROF
Part-Whole	PARTOF, HASA, MADEOF
Spatial	ATLOCATION, LOCATEDNEAR
Utility	USEDFOR, CAPABLEOF, RECEIVESACTION
Creation	CRATEDBY
Desire/Goal	CAUSESDESIRE, MOTIVATEDBYGOAL, DESIRES, OBSTRUCTEDBY
Quality	HASPROPERTY, SYMBOLOF
Temporal	HASSUBEVENT, HASFIRSTSUBEVENT, HASLASTSUBEVENT, HASPREREQUISITE, CAUSES
Other	RELATEDTO

Table 2.2 Relations in `ConceptNet` 5.5. The dimension is taken from Ilievski et al. (2021).

ontology does not contain a predefined hierarchical structure. In a recent study, Ilievski et al. (2021) proposed an ontology that organises the 34 relations into twelve broader dimensions (cf., Table 2.2). This comprehensive ontology serves as a base for designing a specialized relation ontology for word association in our Chapter 3.

Auxiliary Edges `ConceptNet` v5.5 introduces a rich set of labels for edge information. In addition to the aforementioned sources and relations, it contains other fields, namely: (a) weights that signify the strength of an edge; and (b) context, a text sentence describing the relation in text format, along with other information.[2]

However, it's worth noting that these fields are often noisy or missing, limiting their utilities. The weights, for instance, sourced from the original databases, have varied ranges due to differences in data collection. This can result in weight values ranging from decimals (e.g., 0.30) to integers (e.g., 9), causing these weights to lose their intended meanings and often be overlooked. In terms of context, it often takes on a highly templated structure, such as the sentence *A cat is an animal*. This template structure can simplify the parsing process. However, it does not contain additional semantic information, especially when used as training data for models that extract relations from natural language sentences.

[2]For further information, refer to the official website: https://github.com/commonsense/conceptnet5/wiki/

Properties of ConceptNet The first property is that the complex construction process, requiring a significant amount of engineering across multiple iterations, including the developing the complicated data collection interface, highly-engineering process of converting free-text sentences into structured triples and the integration of other resources. Although `ConceptNet` 5.5 is multilingual, most non-English knowledge is automatically extracted from Wiktionary or other multilingual dictionaries. The high complexity construction process of `ConceptNet` makes it difficult to construct `ConceptNet` using native speakers for many different languages.

The second property is the constrained elicitation used in the crowd-sourcing source of `ConceptNet`. To improve the efficiency of collecting knowledge, OMCS-2 (Liu and Singh, 2004) mainly uses templates[3] to elicit commonsense knowledge from participants. This means that participants are biased by the given relation types in the template. This prevents participants from coming up with other types of commonsense knowledge that is related to the concept. For example, when *'Camping is a ___'* is given, participants will respond with *activity*, but ignore other commonly associated concepts such as *nature, trail, and hike*.

The third property is that `ConceptNet` is sparse, compared to other knowledge graphs such as Freebase (Bollacker et al., 2008) whose nodes are a set of fixed entities. This is related to many reasons. One reason is the implicit and nuanced nature of commonsense knowledge, meaning that it is massive and difficult to elicit. Another reason is the nature of using phrases to represent concepts, because natural language is highly compositional, rich in variations of expressions, and ambiguous. This causes concepts that are conceptually related but not equal are often expressed as distinct nodes (Malaviya et al., 2020), e.g., *go to bed early* and *go to sleep early*. Many works focus on completing `ConceptNet` to improve its coverage, including using external resources to mine knowledge (Zhang et al., 2020a), learning neural networks to represent knowledge inside to allow better inference (Li et al., 2016, Malaviya et al., 2020), or using crowd-sourcing to collect knowledge that `ConceptNet` is lacking (Sap et al., 2019a). However, the sparsity issue remains. This

[3] Although Liu and Singh (2004) state that the free-text elicitation is used, but it is unknown how much it is used by participants and the `ConceptNet` 5.5 mainly contain the templated knowledge.

motivates us to explore alternative approaches of eliciting commonsense knowledge in this thesis.

2.2.2 Word Association Networks (WANs)

A word association network is a type of associative network, building from large collectives of word association data (cf., Section 2.1.2.2). In contrast to `ConceptNet`, which is constructed by integrating multiple resources with a high complexity of engineer work, Word Association Networks (WANs) are created by directly collecting spontaneous associated words from human participants. Concepts in these networks are empirically connected by humans, rather than extracted from existing text-corpora or dictionaries.

A WAN is built by eliciting associations from humans using a word association task. According to the task definition (cf., Section 2.1.2.2 and Figure 2.9), participants are presented with a stimulus word (called *cue*) and respond with the words that spontaneously come to mind (called *associations*). Note that in this graph, nodes are predominantly single words according to the task definition. Figure 2.10 presents an example subgraph.

These associations, formed from humans' continuous experience, reflect mental links between concepts (Moss et al., 1995, Bower, 2000, Nelson et al., 2004). They have been used to study the organization of semantic memory (Goñi et al., 2011), to understand the human mental lexicon (Gósy and Kovács, 2002, De Deyne et al., 2016a, Fitzpatrick and Thwaites, 2020), and to investigate the cultural differences (Guerrero et al., 2010, Korshuk, 2005). Though associations can differ among individuals and specific groups due to distinctive experiences (Nelson and Schreiber, 1992, Morais et al., 2013, Zortea et al., 2014, Dubossarsky et al., 2017), aggregating associations across large populations allows us to construct word association norms, capturing shared conceptual views among native speakers (Kiss et al., 1973, Moss and Older, 1996, Nelson et al., 2004).

To improve the validity and applications of word associations, collecting large-scale word associations that span a broad lexicon. To achieve this efficiently, it is essential to focus on the scale and variety across three dimensions: cue words, associations, and participants.

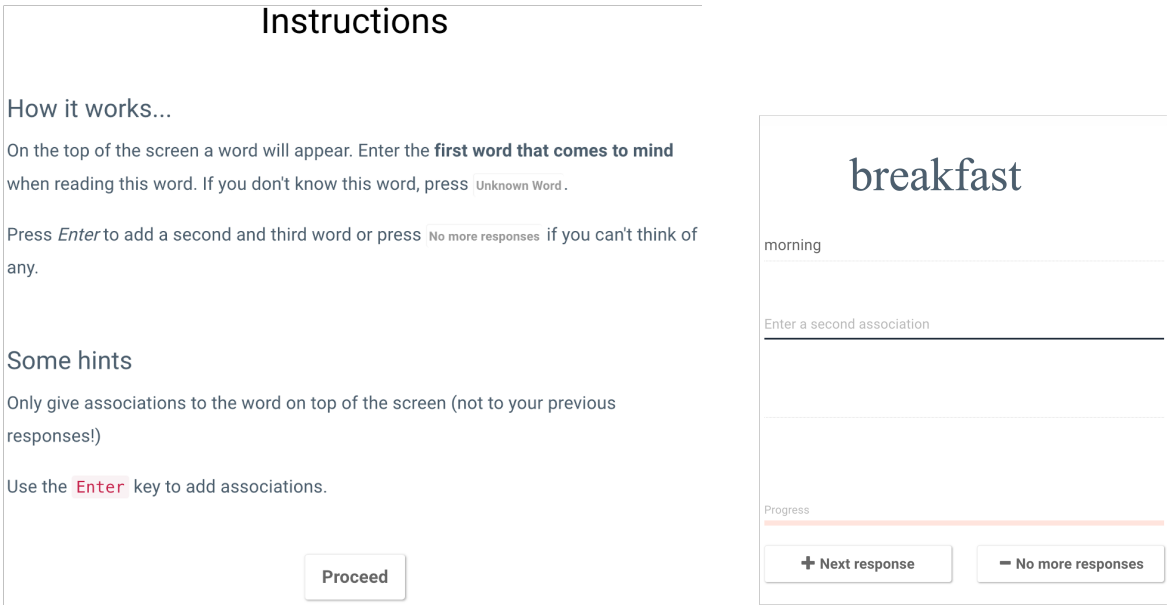


Figure 2.9 The guideline (left) and an example task (right) used in SWOW, taken from its website https://smallworldofwords.org/en.

Firstly, this requires collecting associations for thousands of concepts from a large number of participants, which can be achieved through crowd-sourcing. A few large-scale WAN have been created using this approach, for example, the University of South Florida norms (USF; Nelson et al. (2004)), collected a set of 5,019 cues from over 6,000 participants from college students. The Edinburgh Associative Thesaurus (EAT; Kiss et al. (1973)) collected responses for 8,400 cues. Meanwhile, the Small World of Words (SWOW; De Deyne et al. (2019)), containing over 12,000 cue words, was collected from over 90,000 participants from the general public.

In addition to increasing the number of participants, the diversity of participants can also impact the diversity of associations. Both USF (Nelson et al., 2004) and EAT (Kiss et al., 1973) were collected from college students, while SWOW was collected from the general public. The latter approach allows participants with diverse backgrounds to contribute their associations, thereby increasing the diversity of associations as well as the overall scale.

To better capture a wide range of associations, it's important to collect not just the first response but also the second and third ones that come to mind, as these add valuable diversity (De Deyne and Storms, 2008b). For example, given the word *breakfast*, a participant may

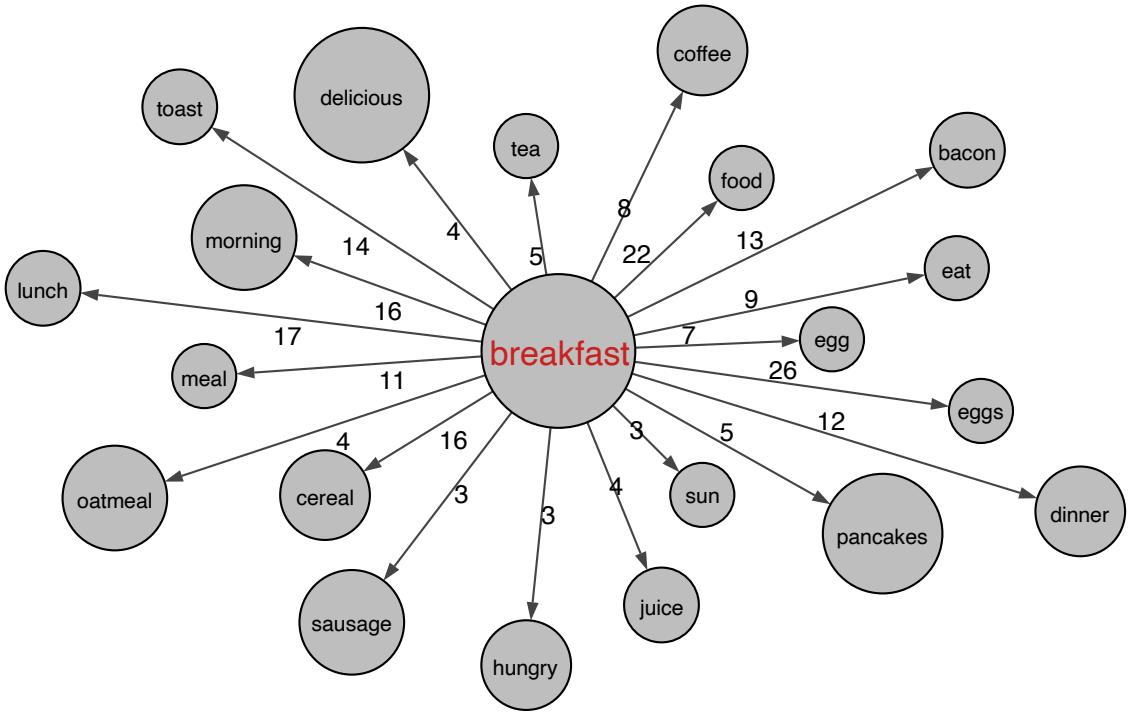


Figure 2.10 Associations center around the cue word 'breakfast' from SWOW (De Deyne et al., 2019) (repeated from a sub-graph of Figure 2.7).

respond with the top three words that come to mind, such as *morning*, *food*, and *sun*. This method produces a more diverse and denser WAN (De Deyne et al., 2013b) and has become a standard practice (De Deyne et al., 2013b, De Deyne and Storms, 2008a, De Deyne et al., 2019, Cabana et al., 2023). We adhere to this method for gathering word associations in Chapter 3. In Chapters 4 and 5, we use the SWOW, which was compiled in a similar manner. Given the high relevance of SWOW to this thesis, will provide a more detailed introduction in the following section.

2.2.2.1 SWOW

The Small World of Words is an ongoing project collecting word associations in 18 languages, including Dutch (De Deyne and Storms, 2008b), English (De Deyne et al., 2019), Spanish (Cabana et al., 2023), Mandarin, Vietnamese, Arabic, etc.[4] We use the English version dataset (De Deyne et al., 2019) in Chapters 4 and 5. To clarify, in this thesis, the term

[4]https://smallworldofwords.org/en/project/stats

SWOW is used to specifically refer to its English version, SWOW.en, unless otherwise explicitly stated.

Constructing SWOW SWOW is collected through its own website.[5] Figure 2.9 presents the guideline from the website and an example task. First, a participant read the instructions to understand the process before doing a task. In the task, a cue word (e.g, *breakfast*) is displayed at the top of the screen, and participants are asked to generate the first three responses that come to mind. During data collection, a task includes a batch of 14-18 cues that are presented to each participant.

Several strategies are introduced to control the quality during data collection,[6] including the instructions (cf., Figure 2.9 left) and the post-processing after the data collection. The instructions include: (a) a clear guideline for participants to generate single-word responses instead of sentences; (b) allowing a participant to skip a word if it is unknown to them; (c) allowing the participant to move to the next cue word if they are unable to produce three different associations for a given cue. The collected data is further cleaned by removing responses from unreliable participants (e.g., those producing n-gram responses or non-unique responses) and is normalized through spelling checks.

The set of cue words in SWOW is based on prior work on semantic priming (Hutchison et al., 2013), word association norms (Nelson et al., 2004), and property generation norms (McRae et al., 2005). During data collection, the cue words presented to the participants are dynamically sampled to ensure even coverage of both frequent and less-frequent words.

As a result, the version of SWOW released in De Deyne et al. (2019) is currently the largest English word association dataset. Collected between 2011 and 2018, it contains 12,217 cue words and 3,665,100 responses from over 90,000 participants with diverse backgrounds. Importantly, for each cue word, responses from 100 unique participants are included. This large-scale provides rich information about the diversity of associations.

[5] https://smallworldofwords.org/en/project/home

[6] This strategy is used in our Chapter 3 when collecting our own dataset.

Edges As shown in Figure 2.10, edges in SWOW include both the directions of associations and their weights (numbers on the edges). Typically, the direction is FORWARD, pointing from a cue word to an association word (e.g., *breakfast* → *morning*), meaning that the cue word elicits the association. However, sometimes cues and associations can be mutually associated if they are both in the cue pool and elicit each other (e.g., *breakfast* ↔ *morning*).

In SWOW, edges are weighted by the fraction of participants who made that association. This weighting provides important insights into the strength of the connection between two words (Deese, 1966, Cramer, 1968), and can aid in estimating their semantic similarity. De Deyne et al. (2019) propose several methods to leverage word associations in measuring word similarity and compare these methods to human similarity judgments. They discovered that a random walk measure is the most effective. This measure calculates similarity between word pairs by examining the distributional overlap of the direct and indirect paths they share in the network. We adopt this approach as a measurement for semantic similarity in Chapter 5.

Properties of SWOW The first property of SWOW is the simplicity and effectiveness of the collection framework. The graph in SWOW is formed and grows by collecting associations from non-expert volunteers over time. The platform is consistently active and allows any participant to add new associations at any time. The word association task itself is simple and straightforward, allowing any native speaker can participate. This makes it practical for the SWOW project to expand to multiple languages in a sustainable way. To initiate the process for a new language, we can begin by sampling an initial set of high-frequency cue words from a corpus. As responses accumulate, we then sample frequent words from these responses in an iterative manner to expand our set, as exemplified by the Rioplatense Spanish study (Cabana et al., 2023). Although this thesis focuses primarily on English, it is important to recognise that expanding the scope to include a diverse range of languages can significantly enhance the potential applications of this work. Not only would this benefit speakers of various languages worldwide, but it would also offer valuable insights from a basic research point of view. Current theories in cognitive science are disproportionately based on studies

conducted in English, leaving us unsure about the universal applicability of these insights. By incorporating diverse linguistic and cultural perspectives could contribute to a more complete and universally applicable understanding of cognitive processes.

The second property is the unconstrained nature of word associations, compared to the OMCS template sentences used by ConceptNet Free associations do not add any relational constraints between the links of cues and associations, thus eliciting broader associations that may not be elicited from given templates. This also results in more direct connections among words, which might require multiple hops to connect in ConceptNet (Yao et al., 2022).

The third property of SWOW is that its graph structure is different from that of ConceptNet. Compared to the sparse connections among concepts in ConceptNet, cue words in SWOW are exposed enough to collect a wide range of associations, resulting in much denser connections and more complete representations for each cue word. Another aspect is the edge information; although specific relational edge labels are missing in SWOW, the association strength and elicitation direction can be used to distinguish the connections among words. This motivates our approach in Chapter 4, where we provide a deeper and systematic comparison between SWOW and ConceptNet.

In summary, in this section, we introduced two types of semantic networks. The differences between them are predominant, ranging from task priming to diversity and graph structures. ConceptNet has been widely used in NLP as an external KG to assist with commonsense reasoning tasks, while the knowledge in SWOW is still under explored. This motivates our study in Chapters 4 and 5.

2.3 Computational Representations of Semantic Knowledge

A critical question in NLP is how machines can represent meanings of concepts and understand them, just as humans do. This question has led to the development of various algorithms and models to learn the meanings of concepts. We provide a review on distributional representations based on text (Section 2.3.1) as well as structured knowledge (Section 2.3.2).

Understanding these computational representations is important for grasping the foundational principles and motivations underlying the thesis presented.

Text corpora have been considered as a proxy for the knowledge humans have experienced and accumulated throughout their lives (Kumar, 2021) and learning the meanings of concepts directly from text has been widely explored. Similar to how word associations are built up in human minds through repeated co-occurrence of word forms (Moss et al., 1995), the statistics of word co-occurrence in corpora are used to learn the meanings of concepts. This is called the distributional hypothesis (Harris, 1954) which claims that "you shall know a word by the company it keeps" (Firth, 1957), suggesting that the meaning of a word can be inferred by its usage, i.e., its distribution in text.

Distributional semantic models (DSMs) represent a significant approach to learn word meanings from natural language in text corpora, grounded in the distributional hypothesis. In DSMs, words are represented as vectors or embeddings. DSMs fall into two categories: "count-based" (Lund and Burgess, 1996, Landauer and Dumais, 1997, Jones and Mewhort, 2007), which directly derive meanings from word co-occurrences and were predominant before neural networks, and "prediction-based", which use neural models to learn embeddings. Notable models include Word2Vec (Mikolov et al., 2013), which learns to predict target words or their context. Subsequent models like GloVe (Pennington et al., 2014) and FastText (Joulin et al., 2017) expanded on Word2Vec, integrating features like global corpus statistics and character-level details. Word embeddings successfully encode word meanings into low-dimensional dense vectors and have proved useful in a wide range of applications. Yet, they face limitations: the out-of-vocabulary (OOV) issue (Pinter et al., 2017), where untrained words can't be represented, and the polysemy problem, as each word has a static embedding unchanged by context (Peters et al., 2018).

To solve the OOV issue, subword tokenization, such as WordPiece (Devlin et al., 2019) and byte-pair encoding (Gage, 1994), has been proposed to break down words into smaller units, allowing the model to represent new or unseen words as combinations of more frequent subwords. By using subword tokenization, the model can better handle OOV words and improve its ability to accurately represent the meaning of these words. For example, the word

"happiness" could be split into three tokens: *"hap"*, *"##pi"* and *"##ness"* in WordPiece. This allows the model to represent rare and unseen words, improving its ability to handle OOV words and the coverage of vocabulary.

To solve the polysemy issue, dynamic embeddings were recently proposed. Different from static embeddings that are represented as a single, fixed embeddings after training, dynamic embeddings are computed dynamically based on how a word occurs in a context (Melamud et al., 2016, McCann et al., 2017, Peters et al., 2018). For example, the embedding for *bank* should depend on its context, leading to two different embeddings in the following two sentences: *She cashed a cheque at the bank*, and *He sat on the bank of the river and watched the currents.* The most recent and effective approach to obtain contextual embeddings is through pre-trained language models (e.g., BERT (Devlin et al., 2019)). In short, a PLM is a neural language model trained on large-scale corpora that is used to dynamically compute the contextual embeddings. PLMs have achieved striking success in NLP and will be the fundamental models to represent semantic knowledge in Chapters 3, 4 and 5. We now introduce them with more depth.

2.3.1 Pre-trained Language Models

The concept of pre-training comes from transfer learning, referring to a model that learns from one domain first and then transfers it to other domain or tasks (Erhan et al., 2010, Howard and Ruder, 2018). In general, a language model assigns probabilities to a sequence of words, indicating how likely they would occur in natural language. A pre-trained language model refers to a neural network that is pre-trained with language model objectives to learn general representations of words from massive unlabelled text data (Jurafsky and Martin, 2023). We use such models from Chapters 3 through 5.

Two main training objectives have been used to pre-train language models. One training objective is to predict the next token given the past tokens. For example, given the tokens *A canary is a kind of* ___ , the model is trained to predict the next likely token (e.g., *bird*). Language models pre-trained with this objective are called autoregressive language models

(ALM) (Dai and Le, 2015, Howard and Ruder, 2018, Radford et al., 2018).[7] Another objective, the "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953), has been proposed to predict a token that based on the bi-directional context directly. For example, in a sequence *A* `[MASK]` *is a kind of bird.*, where the `[MASK]` is a token that is randomly masked, the training objective is to predict the token in the masked position utilizing the full sentence as context. Chapter 3, 4 and 5 use the models trained with MLM training objective. One pre-trained language model in Chapter 3 uses a combination of the two training objectives. We will provide more details for each of these models in Sections 2.3.1.3 to 2.3.1.4.

Conventional model architectures such as the LSTM (Hochreiter and Schmidhuber, 1997) have been used for language models. However, LSTM only allows the information to flow in one direction sequentially (either left-to-right or right-to-left), limiting the capacity of capturing longer context dependency and making it difficult for parallel computation. Recently, a novel architecture Transformer (Vaswani et al., 2017), illustrated in Figure 2.11, has been proposed to enable the computation of longer sequences in parallel, improve the ability to represent long-context information and increase computation speed. Since then, the Transformer has become the dominate architecture for pre-training language models. We use PLMs based on Transformer through Chapters 3 to 5. The details of the Transformer will be described in Section 2.3.1.1.

PLMs acquire general word representations through training on extensive general domain corpora. However, these general representations differ from those obtained through traditional supervised approaches directly targeting specific downstream tasks due to the domain discrepancy. Consequently, applying knowledge in PLMs on downstream tasks necessitates adapting the general representations to the specific domains of those tasks. This adaptation can be achieved through two typical strategies: **feature-based** and **fine-tuning**. The **feature-based** strategy (Dai and Le, 2015, Akbik et al., 2018, Peters et al., 2018) uses pre-trained language models to extract contextual embeddings to initialise embeddings when training models for downstream tasks. This is simple and convenient to obtain initialisa-

[7] Note that models, such as ELMo (Peters et al., 2018), use both forward (left-to-right) and backward (right-to-left) autoregressive objectives. However, ELMo was not designed for use as a generation model.

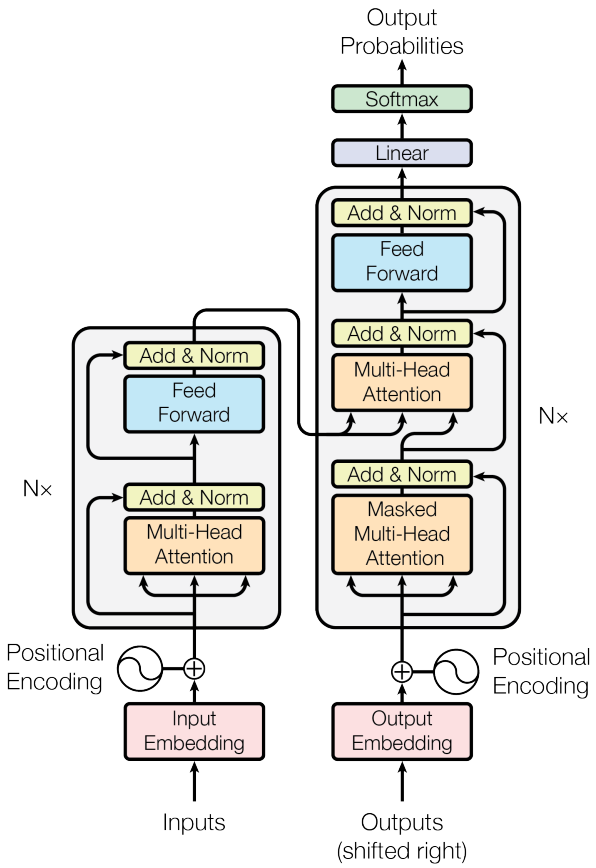


Figure 2.11 The Transformer - model architecture from Devlin et al. (2019).

tion for general representation and it is suitable when the training objectives are different between PLMs and target tasks. We employ the feature-based approach in Chapter 4 to acquire concept embeddings as initial concept representations, which are further updated with the target task. In contrast, the fine-tuning approach (Howard and Ruder, 2018, Devlin et al., 2019, Radford et al., 2018) directly updates pre-trained language model parameters using task-specific data for downstream tasks. During fine-tuning, the task-specific layers replace the top layer of the language model and are trained to adapt to the target tasks, with all parameters in the other layers being updated accordingly. The latter approach allows universal semantic representations obtained during pre-training to be better transferred into downstream tasks, and meanwhile adjusts the parameters based on the tasks. We adopt the fine-tuning strategy in Chapter 3 and Chapter 4 on our tasks for relation classification and question answering.

2.3.1.1 Transformer Architecture

Transformer architecture has become the essential backbone of many PLMs, including masked language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020) and BART (Lewis et al., 2020). We use these models in this thesis from Chapter 3 through 5. Therefore, we introduce the architecture of Transformer below.

A Transformer (Vaswani et al., 2017) is an encoder-decoder neural network architecture that was originally developed for machine translation. Figure 2.11 illustrates the framework. Both the encoder and encoder are composed of a stack of N transformer blocks. The output of one block is the input to the next block. Each transformer block consists of three key components: feed-forward networks, self-attention layer and normalisation layer (Ba et al., 2016).

Self-Attention The key to the Transformer lies in the self-attention layer, which allows the model to understand longer contexts. For example, given the sentence "The trophy would not fit in the brown suitcase because it was too big.", when processing the mention *it*, the self-attention layer allows the model to look at other positions directly and identify the most relevant words *brown suitcase* that are relevant to *it*.

Formally, the self-attention layer maps a sequence of input embedding $\mathbf{X} = [X_1, X_2, \cdots, X_T]$ to a sequence of output embedding $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_T]$. Each Z_i is computed by a weighted sum of other inputs in the sequence. Importantly, the calculation for each Z_i is independent of all other input and output, allowing the parallel computation across the whole sequence and therefore improving the efficiency. The scope of other inputs varies from all proceeding context (including itself) $[X_1, X_2, \cdots, X_i]$ to a entire sequence $\mathbf{X}$, depending on how the attention is being used in encoder or decoder. We illustrate the idea of self-attention being used in Transformer encoder.

First, the input embedding $\mathbf{X}$ is the concatenation of sub-token embeddings and position embeddings (explained in detail below). Self-attention is based on a dot-product attention

among three linear projections of $\mathbf{X}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \tag{2.1}$$

$$\mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \tag{2.2}$$

$$\mathbf{V} = \mathbf{X}\mathbf{W}^{V}, \tag{2.3}$$

where $\mathbf{W}^{Q} \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}^{K} \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}^{V} \in \mathbb{R}^{d \times d_v}$ are specific trainable matrices responsible for projecting $\mathbf{X}$ into three different subspace representations, i.e., $d > d_h$. d denotes the input embedding dimension, d_k, d_k and d_v denote the projected dimension. The matrices $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are termed as query, keys, values respectively. They are used to perform the dot-product attention:

$$\mathbf{Z} = \text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V},$$

where $\sqrt{d_k} = \frac{d}{h}$ is used as a scaling factor to stabilise training by rescaling potentially large values resulting from the dot product of two large vectors. Inside the softmax $(.)$ represents the self-attention matrix, which shows how much each input pair X_i and X_j is relevant to the output Z_i. The softmax function converts these values into a probability distribution, ensuring their total equals 1.

Multi-head Attention Instead of utilizing a singular process of self-attention, Transformer uses a mechanism called multi-head attention, to capture different aspects of relationships among inputs (e.g., syntactic, semantic, or position). The mechanism subdivides the above self-attention process into multiple distinct heads. Each head, indexed by i, is equipped with its individual set of queries $\mathbf{Q}_i$, keys $\mathbf{K}_i$, and values $\mathbf{V}_i$. These components are projected from their corresponding parameters: $\mathbf{W}_i^{Q} \in \mathbb{R}^{d \times d_h}$, $\mathbf{W}_i^{K} \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_i^{V} \in \mathbb{R}^{d \times d_h}$, where it's important to note that d_h is, by design, smaller than d. Multi-head attention first projects $\mathbf{X}$ into multiple smaller subspaces with dimension d_h and then merges them back. The final

representation is a concatenation of the outputs from all heads:

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})\mathbf{W}^O,$$
$$\text{where head}_\text{i} = \text{Attention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i),$$

where $\mathbf{W}^O \in \mathbb{R}^{hd_h \times d}$ is a trainable matrix that projects the representation back into the original dimension d.

Positional Embeddings Alongside multi-head attention, the Transformer introduces the concept of positional embeddings, which are used to provide the model with information about the position of each token in the sequence. This is necessary because the self-attention mechanism used in the Transformer model does not consider the order or position of the words in the sequence by default. The positional embedding is a vector with the same dimension as the input embeddings and is added to the input embeddings to form the final representation of each token. The position information can be either an absolute or relative position of a token in a sequence. Originally, Vaswani et al. (2017) use the trigonometric functions such as sine and cosine to encode the position of each token along each dimension of the embedding vector:

$$PE_{(i,2j)} = \sin\left(\frac{i}{10000^{2j/d}}\right)$$
$$PE_{(i,2j+1)} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

where i is the absolute position and j is the dimension. The idea is to encode the position as a combination of sine and cosine waves with different frequencies and phases. This is particularly important for tasks that require the model to understand the sequential order of words, such as language modelling or machine translation.

Subsequent work of PLMs built on top of Transformers either only use its decoder (Section 2.3.1.2), encoder (Section 2.3.1.3), or retrains the original encoder-decoder (Section 2.3.1.4). Figure 2.12 illustrates their differences.

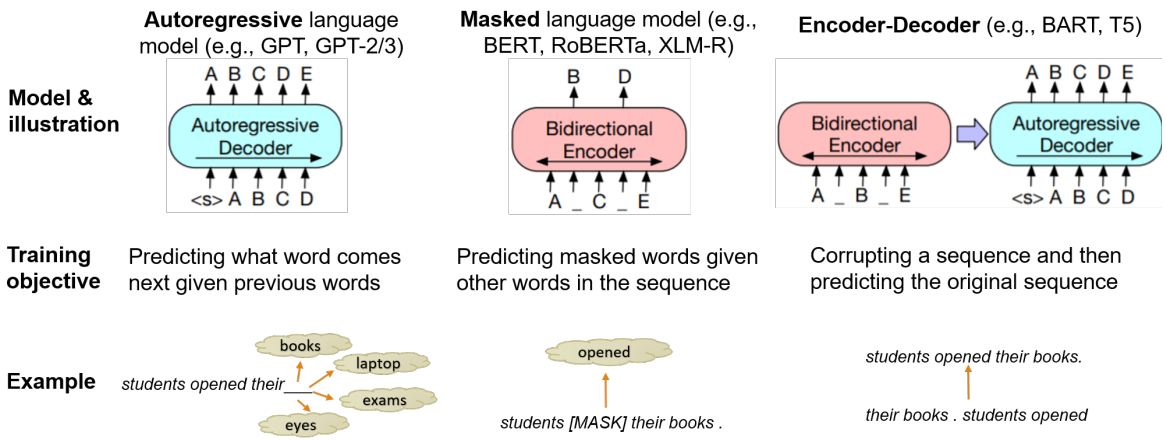


Figure 2.12 Three types of PLMs, adapted from Min et al. (2023) and Lewis et al. (2020).

2.3.1.2 Autoregressive Language Models

Autoregressive language models, also known as casual or unidirectional language models, predict the next token from past or left tokens, as illustrated in Figure 2.12 (left). Given a sequence of tokens $\mathbf{x} = [x_1, x_2, \cdots, x_T]$, the model is trained with $\mathcal{L}_{\text{LM}}$ loss function:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log p(x_t \mid \mathbf{x}_{<t}; \theta), \tag{2.4}$$

where $\mathbf{x}_{<t} = x_1, x_2, \cdots, x_{t-1}$, and θ are the language model parameters.

OpenAI GPT (Radford et al., 2018) is the first deep autoregressive language model trained on a transformer decoder architecture with a large amount of data from BooksCorpus (Zhu et al., 2015) and Word Benchmark (Chelba et al., 2014). This model shows that pre-training on unlabelled data and fine-tuning on labelled data in the target domain improved downstream tasks such as commonsense reasoning and sentiment analysis.

GPT-2 (Radford et al., 2019) has more than 10 times the parameter size of the original OpenAI GPT and is trained on a new large-scale dataset called WebText, which contains millions of pages. This increase in data and parameters significantly enhances the model's capabilities. The model also unifies multitask learning by introducing task conditioning into the formation of pre-training, represented as $P(\text{output} \mid \text{input}, \text{task})$. This allows the model to produce different outputs for the same input, depending on the task at hand. Building on all

these capabilities, GPT-2 exhibits remarkable performance on zero-shot setting, suggesting that, given sufficient capacity, language models can learn universal representations that are directly transferable to downstream tasks. We use GPT-2 to encode graph knowledge in semantic networks such as `ConceptNet` and associative networks in Chapter 4.

GPT-3 (Brown et al., 2020) is 100 times larger than its predecessor, GPT-2, and is trained on a mixture of datasets of approximately 500 billion tokens from sources like Common Crawl and additional web and book datasets. This trend of continually increasing model size has lead to new advancements, as evidenced by the emergent abilities observed in Wei et al. (2022). By the time this thesis was being written, more large language models such as ChatGPT and GPT-4 (OpenAI, 2023), trained on dialogue datasets with reinforcement learning, exhibit remarkable improvements in natural language generation and understanding. These models, trained with dialogue datasets and reinforcement learning from human feedback, enable complex human-like conversations on topics such as coding and mathematics. While the GPT family currently dominates the development of causal language models, several other models contribute to this field. See other survey papers for a more complete review (Qiu et al., 2020, Amatriain, 2023, Zhao et al., 2023).

2.3.1.3 Masked-Language Models

Masked language models are a family of language models that are trained using the Masked Language Model (MLM) objective, i.e., predict missing words in a sentence given its surrounding context. This is illustrated in Figure 2.12 (middle). The bidirectional nature of MLMs allows them to capture both left and right context, which helps to capture longer-range dependencies in the sentence. This is particularly important for natural language understanding tasks, such as question answering and text classification. We use PLMs trained on the MLM objective in Chapters 3 and 4 as encoders for semantic knowledge understanding tasks.

The first developed masked language model is BERT (Devlin et al., 2019), which is pre-trained on the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia

(2,500M words). It comes in two model sizes: $\text{BERT}_{\text{base}}$ with 117M parameters, comparable to GPT-1, and $\text{BERT}_{\text{large}}$ with 340M parameters. We use $\text{BERT}_{\text{large}}$ in Chapter 5.

BERT employs the MLM training objective, which randomly masks 15% of the tokens from the input and trains the language model to predict the original word of the masked token based on the context on its left and right. The loss function is:

$$\mathcal{L}_{\text{MLM}} = -\sum_{\hat{x}\in m(\mathbf{x})} \log p(\hat{x} \mid \mathbf{x}_{\setminus m(\mathbf{x})}; \theta), \tag{2.5}$$

where $m(\mathbf{x})$ and $\mathbf{x}_{\setminus m(x)}$ denote the masked words from $\mathbf{x}$ and the remaining words respectively. θ denotes the parameters in the masked-language model.

Except for $\mathcal{L}_{\text{MLM}}$, BERT also includes the next sentence prediction (NSP) task to capture sentence relationships. The task aims to predict whether a sentence is the actual next sentence that follows the previous one, which can be formatted as a binary classification task:

$$\mathcal{L}_{\text{NSP}} = -\log p(t \mid \mathbf{x}, \mathbf{x}'), \tag{2.6}$$

where $t = 1$ when $\mathbf{x}$ and $\mathbf{x}'$ are consecutive sentences, otherwise 0. The overall loss function in BERT is the sum of $\mathcal{L}_{\text{MLM}}$ and $\mathcal{L}_{\text{NSP}}$.

After pre-training, BERT can be used for downstream tasks, such as text classification and question answering, however, those tasks are different from the tasks in pre-training. To handle various downstream tasks, BERT unifies the inputs formats for pre-training and fine-tuning for better adaptation. Firstly, a special token `[CLS]` is inserted into the first token of every sequence. The final hidden state of `[CLS]` token aggregates all of the sequence information, which can be used as the input for downstream classification tasks with a task-specific layer on top (e.g., softmax for classification, or any transformation for regression). Secondly, BERT uses a special token `[SEP]` to separate input sequences when multiple sequences are concatenated, so that BERT can distinguish different segments of the input, such as questions vs answer options in the question answering task or two text inputs in textual similarity tasks. Furthermore, every token is attached with a learned segment embedding, which has the same dimension as token embeddings, to distinguish the segment

they belong to. The input embedding of every sequence is the sum of token, position[8] and segment embedding.

BERT has demonstrated the effectiveness of masked-language models by advancing state-of-the-art in eleven natural language understanding tasks, such as SQuAD v1.1 (Rajpurkar et al., 2016), GLUE (Wang et al., 2018a), and MNLI (Williams et al., 2018). Another key impact of BERT is that it also empirically proves that general representations can be learned via pre-training on large text corpora and fine-tuning on a smaller task-specific dataset. This approach has since become a standard practice in NLP.

Subsequent work based on BERT has focused on improving its architecture and training methodology. We introduce two variants that are relevant to our thesis.

RoBERTa RoBERTa (Robustly Optimized BERT Pre-training Approach; Liu et al. (2019)), a modified version of BERT, is different from BERT in several aspects. Firstly, the next sentence prediction training objective is removed as Liu et al. (2019) empirically show that without NSP yield comparable or better performance compared to with NSP. Secondly, while BERT uses static masking to generate masks during data prepossessing, RoBERTa uses dynamic masking to randomly mask a sequence every time it's feed into the model during pre-training. The authors found that dynamic masking is comparable or slightly better than static masking. Thirdly, RoBERTa uses advanced pre-training strategy, including a larger and more diverse corpora with longer training time, longer sequences and larger batch sizes.

RoBERTa outperforms $\text{BERT}_{\text{large}}$ across a wide range of tasks, showing that the power of masked-language models can be further stimulated with better training strategies and more data. Therefore, RoBERTa is used in Chapter 3 as a text encoder.

ALBERT Another relevant variation is ALBERT (A Lite BERT; Lan et al. (2020)), which focuses on designing smaller models to reduce the computational cost and memory requirements. ALBERT achieves this by implementing two key strategies: factorized embedding parameterization and cross-layer parameter sharing.

[8] In BERT, the position embedding is learned during training, whereas in the Transformer, the position encoding is a fixed function of the position index and a set of learned parameters.

Factorized embedding parameterization reduces the number of model parameters by factorizing the embedding matrix into two smaller matrices, where one first projects the tokens into a lower dimension and another one projects them into the hidden space. The factorization prevents model parameters from growing with the increase of hidden size. While cross-layer sharing of parameters—including feed-forward network and attention parameters—across all layers of the model serves to further reduce the model size, it also prevents the number of model parameters from growing with the increase in network layers.

As a result, the two strategies allow ALBERT to scale up to larger size of hidden dimensions while maintain smaller parameters, e.g., $\text{ALBERT}_{\text{xxlarge}}$, the largest version of ALBERT, increases the hidden dimension size to 4096 with a total of 235M parameters, which only around 70% of $\text{BERT}_{\text{large}}$'s parameters.

In addition to the masked language model training objective in Equation (2.5), ALBERT uses a sentence-order prediction (SOP) loss, which is similar to Equation (2.6), but employs a different negative sampling strategy. In contrast to BERT, which samples negative examples from different documents, ALBERT swaps the order of positive samples to obtain negative samples.

Lan et al. (2020) found that this strategy consistently improve downstream task performance. $\text{ALBERT}_{\text{xxlarge}}$, consistently outperform $\text{BERT}_{\text{large}}$ and $\text{RoBeRTa}_{\text{xxlarge}}$ on the GLUE benchmark (Wang et al., 2018a), demonstrating that designing a more efficient parameter sharing and scalable training can further improve the capacity of masked-language models. Therefore, we use $\text{ALBERT}_{\text{xxlarge}}$ in Chapter 4.

Despite the success of masked-language models in various NLP tasks, one of their major limitations is their inability to perform generation tasks, such as text summarisation and machine translation, as they require the model to generate new text rather than predict missing words in a given sequence. To address this limitation, some models have been proposed that use both the encoder and decoder, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), to combine natural language understanding and generation in a unified training approach, which we will describe in more details below.

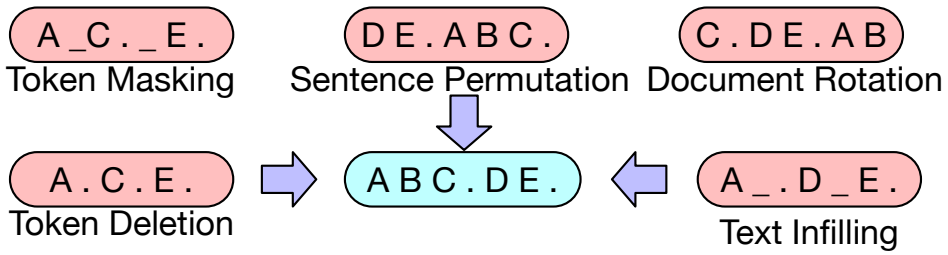


Figure 2.13 Illustration of six strategies used by BART to corrupt an original sequence **ABC.DE.**. Adapted from Lewis et al. (2020).

2.3.1.4 Encoder-Decoder Language Models

Encoder-Decoder Language Models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) use both the encoder and decoder Transformer architecture, as shown in Figure 2.12 (right). In this architecture, the training objective is denoising autoencoders (DAE; Devlin et al. (2019), Wilson (1953)), which are trained to reconstruct the original input from corrupted input. Except for the random masking used in BERT, the corruption can be done with arbitrary transformations, such as random masking, deleting, inserting or replacing some tokens in the original inputs. See Figure 2.13 for an illustration of these corruptions. During pre-training, the encoder learns the representation for the corrupted inputs and the decoder learns to recover the corrupted tokens with the representation from the encoder. Given a sequence of tokens $\mathbf{x} = [x_1, x_2, \cdots, x_T]$, and its corrupted version $\hat{\mathbf{x}}$, the loss function is:

$$\mathcal{L}_{\text{DAE}} = -\sum_{t=1}^{T} \log p(x_t \mid \hat{\mathbf{x}}, \mathbf{x}_{<t}; \theta), \tag{2.7}$$

By pre-training on the DAE objective, PLMs are expected to learn more robust and meaningful representations of the input data, even when it's noisy and incomplete. This fits into our scenarios that relationships are missing in the associative network described in Section 2.1.2.2 and we use this kind of model in Chapter 3. We now introduce the details of BART (Lewis et al., 2020), the model we used in Chapter 3.

The architecture of BART can be seen as a generalisation of BERT and GPT, with a bi-directional encoder and a unidirectional decoder (see Figure 2.12, right), which is trained on the same corpora as RoBERTa with 160G of news, books, stories and web text. BART aims

to train a strong and robust sequence-to-sequence model that can reconstruct the corrupted sequences. That is, the input of BART is a corrupted sequence, and the output is the full sequence of original inputs.

BART uses a series of strategies for different granularity of corruption. On the token level, BART uses token masking (same as BERT) to randomly replace original tokens with `[MASK]` token, and token deletion that randomly delete tokens, and therefore requires the model to predict the position of deleted tokens. On the span level, text infilling (Joshi et al., 2020) is used to replace a text span in the original text with a mask token and requires the decoder to recover the masked tokens. In order to learn document structure, BART uses sentence permutation that shuffles sentences in documents and requires the decoder to recover the order, as well as document rotation that requires the decoder to identify a the first token of the document that is replaced with a random token. Those corruption operations improve the robustness of BART, particularly text-infilling, which has been shown to be the most effective form of corruption.

Similar to other PLMs, after pre-training, BART can be used for downstream tasks by fine-tuning. For fine-tuning, original (uncorrupted) sequences are feed into both encoder and decoder, the final hidden states from the decoder will be used for classification. BART performs comparable to RoBERTa on natural language understanding tasks, but achieves consistently boost on generation tasks such as summarisation. We use BART in Chapter 3 to recover missing semantic relations between two concepts.

2.3.2 Knowledge Graph Representation

Beyond free-text knowledge in large corpora, often utilised as a reflection of human knowledge accumulation (Section 2.3.1), human knowledge is also amendable to representation in structured formats, such as semantic networks (Section 2.1.2), also known as knowledge graphs (KGs). Knowledge graph representations have proven beneficial for various NLP downstream tasks, a subject that we will explore in depth in Chapter 4. However, a challenge arises in how to transfer the symbolic knowledge in knowledge graphs into a vector space that captures their semantic properties. In this section, we will introduce models that are specifi-

cally designed for this transformation, including embedding-based models (Section 2.3.2.1) and Transformer-based models (Section 2.3.2.2)

Conventional embedding-based approaches focus on learning a low-dimensional embedding for each entity and relation in a knowledge graph based purely on the structure of KGs (Wang et al., 2017, Ji et al., 2022). This kind of approach is simple and effective to learn, however, it heavily relies on fixed embeddings and is difficult to generalise to unseen knowledge. This is crucial when the knowledge graph is incomplete or sparse, i.e., a massive inventory of concepts which are loosely connected.

Inspired by the success of the Transformer in transfer learning (Devlin et al., 2019, Brown et al., 2020), Transformer based KG models have been proposed to capture the compositional semantics of concepts and utilise the transferable knowledge in PLMs (Bosselut et al., 2021, Clouatre et al., 2021, Daza et al., 2021). Importantly, PLMs can generate the new knowledge based on the learnt knowledge even when the relation or concepts are unseen. We investigate the effectiveness of the two types of representations in Chapter 4.

2.3.2.1 Embedding-based Models

Embedding-based models offer an approach to capturing the intricacies of structures like knowledge graphs. In this paradigm, every concept or relation in a knowledge graph translates into a low-dimensional dense vector, thereby generating an embedding matrix. These embeddings can be used to perform a wide range of tasks, including knowledge graph completion, and can also be integrated into downstream applications, such as question answering, the focus of Chapter 4.

Of the myriad methods developed for learning these embeddings, TransE (Bordes et al., 2013) was first proposed and emerges as a popular choice. It has demonstrated its prowess in representing knowledge graphs, including WordNet and Freebase (Bordes et al., 2013). While there are modifications of TransE tailored to harness diverse relations and interactions between concepts (Wang et al., 2014, Lin et al., 2015), other models prioritize learning from relation paths (Neelakantan et al., 2015) or sub-graphs (Velickovic et al., 2018, Schlichtkrull et al., 2018) rather than triplets. Despite this, TransE stands out for its simplicity and efficacy

in training and application. Accordingly, we employ TransE in Chapter 4 for encoding semantic networks.

The core of TransE is its translation-based scoring function, which interprets relations as translations between two related concepts. For instance, in the triple (*poodle*, ISA, *dog*), the relation ISA serves as the translation operation that connects *poodle* and *dog*. Building on this idea, when a relation triple (c_1, r, c_2) holds true, the embedding for c_2 tends to align closely with the sum of the embeddings for c_1 and r, expressed as $(\mathbf{c}_2 \approx \mathbf{c}_1 + \mathbf{r})$. The training of TransE incorporates a margin-based ranking function, which scores the distances between positive and negative triples:

$$\mathcal{L} = \sum_{(\mathbf{c}_1,\mathbf{r},\mathbf{c}_2)\in S} \sum_{(\mathbf{c}'_1,\mathbf{r},\mathbf{c}'_2)\in S'} [\gamma + d(\mathbf{c}_1 + \mathbf{r}, \mathbf{c}_2) - d(\mathbf{c}'_1 + \mathbf{r}, \mathbf{c}'_2)], \tag{2.8}$$

Here, S signifies a batch of positive samples, while S' represents a batch of negative samples derived from S by substituting c_1 or c_2 with a randomly chosen concept from the graph. The function $d(\mathbf{c}_1 + \mathbf{r}, \mathbf{c}_2) = |\mathbf{c}_1 + \mathbf{r} - \mathbf{c}_2|_N$, where N denotes the L1 or L2 norm, measures the distance between two vectors under a relation $\mathbf{r}$. γ is the margin between positive and negative samples.

While embedding-based models are simple and efficient, they also have inherent limitations. Notably, when a knowledge graph is sparse or incomplete, the ability to represent knowledge is constrained, and the learned embeddings are limited to concepts within the graph. This poses a challenge in scenarios where new knowledge cannot be retrieved in the existing graph.

2.3.2.2 Transformer-based KG Representation Models

In contrast to embedding-based models that rely on static embedding matrices to store learnt knowledge, Transformer-based models transfer knowledge in knowledge graphs into model parameters. This involves incorporating knowledge graph elements, such as triples, paths, or sub-graphs into language models to effectively capture the semantics and structures of knowledge graphs. The rich structure knowledge in knowledge graphs can in turn enhance

the language understanding capabilities of pre-trained models, which can then be used for downstream tasks.

There are two common approaches for combining knowledge graphs with PLMs: joint-training and fine-tuning. Joint-training (Yasunaga et al., 2022, Wang et al., 2021b) involves training language model on textual knowledge using language model objectives described in Section 2.3.1, as well as graph knowledge tasks like link prediction. This allows for better unification of knowledge from the two sources, however, it requires training from scratch, which can be costly and time-consuming. Fine-tuning, on the other hand, involves converting knowledge graph elements, such as triples, paths, or sub-graphs into natural language descriptions, which are then used to fine-tune language models. This involves converting relation or concept ontologies into their natural language (e.g., HASFIRSTSUBEVENT $\rightarrow$ *has first subevent*) or using a more sophisticated model to retrieve natural language sentences from external corpora (Bansal et al., 2022). This approach is popular and effective as it is less computationally expensive and can be generalised across different knowledge graphs. For those reasons, we use the fine-tuning approach in Chapter 4.

One notable fine-tuned model is COMET (Bosselut et al., 2019), which fine-tunes OpenAI GPT (Radford et al., 2018) with individual triplets, aiming to predict c_2 given c_1 and r from a knowledge graph. This is suited for tasks like automatic knowledge graph construction. In contrast to COMET using individual triplets, Path Generator (Wang et al., 2020) fine-tune GPT-2 with paths formed from consecutive triplets sampled from a knowledge graph. The task is to predict the paths (intermediate nodes and relations) given two concepts. For instance, given *(predator, animal)*, the model is trained to predict the path *(predator, DistinctFrom, prey, IsA, animal)*. The loss function can be defined as:

$$\mathcal{L} = -\sum_{t=|c_1|+|c_2|+1}^{t} \log p(x_t \mid \mathbf{x}_{<t}) \tag{2.9}$$

where $\mathbf{x}_{<t}$ are the prefix which is the concatenation of concept pairs (c_1, c_2) and the `[SEP]` token. The Path Generator learns to maximize the probability of the observed paths given the concept pairs. During inference, it can generate the most likely path of connecting two

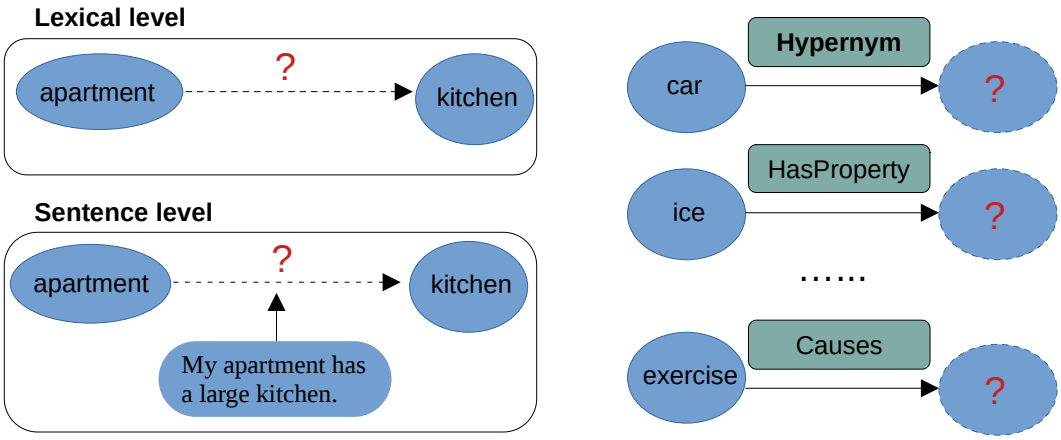
concepts without explicitly specify the relationships. This is suited for scenarios, where concept pairs are known while their underlying connections are missing. We use the idea of Path Generator in Chapter 4 to transfer knowledge in associative networks and study their utility for question answering tasks.

2.4 Tasks: Relation Learning and Commonsense Reasoning

Commonsense is the basic level of practical knowledge and judgment that we all need to help us live in a reasonable and sound way.[9] Equipping machines with commonsense reasoning ability requires understanding the meanings of concepts as well as their links, especially for the relations existing among them (Singh et al., 2002, Davis and Marcus, 2015). These relations form the backbone of knowledge graphs like WordNet (Miller, 1995) and `ConceptNet` (Liu and Singh, 2004), benefiting various NLP tasks, such as machine translation (Zhao et al., 2021) and question answering (Bauer et al., 2018, Mihaylov and Frank, 2018). On the other hand, large-scale word association networks (WANs) encode concepts and their connections with associated concepts, but their relation labels are missing (cf., Sections 2.1.2.2 and 2.2.2). This leads us to ask (a) what are the relations encoded in word associations, and (b) to what extent WANs can benefit these applications in NLP? Therefore, developing models that can predict these relations is important to enhance our understanding of word associations and their application in real-world scenarios.

In our thesis, we aim to gain a deeper understanding of the relationships between associated words and explore the utility of WANs in downstream tasks. To achieve this goal, we need to predict the missing relationships among associated words, which is the task of relation classification in semantic link prediction. Therefore, we review and provide the relevant datasets, task framing, and methods in Section 2.4.1. Along with this, we introduce a variant of task in semantic link prediction, named concept prediction. This task aims to predict a missing concept given one concept and a relation, and it is the subject of interest

[9] https://dictionary.cambridge.org/dictionary/english/commonsense



(a) relation prediction (b) concept prediction

Figure 2.14 Illustration of semantic link prediction tasks: (a) relation prediction and (b) concept prediction. The question mark indicates the missing element to be predicted in each task. In task (a), the missing element is the relation PART-WHOLE. In task (b), the missing elements are the *vehicle* for car, *cold* for ice and *sweet* for exercise. In this thesis, we mainly focus on the relation hypernym.

in Section 5. In Section 2.4.2, we introduce the downstream task commonsense question answering regarding the datasets and models.

2.4.1 Semantic Link Prediction

The task of semantic link prediction aims to predict the missing link between concepts. This task can be categorized into two sub-tasks: (a) relation classification that aims to predict a missing relation r between a concept pair (c_1, c_2); and (b) concept prediction that predicts a missing concept c_2 for a given a known concept c_1 and relation r. Figure 2.14 presents examples for each task.

Typically, relation classification between word pairs can be categorized into two types lexical level and sentence level (cf., Figure 2.14 a). The former aims to classify the relation between a given pair of words, while the latter focuses on identifying the relation between a word pair based on a sentence expressing the relation. Both tasks are relevant to the identification of word relationships, as explored in Chapter 3.

For the concept prediction task, approaches are classified based on the external resources they use, such as text corpora, pre-trained embeddings, or language models (cf., Figure 2.14 b), to capture the context. In Section 2.4.1.3, we will provide more details of these different approaches.

Note that in Section 2.3.2, we introduced models for knowledge graph representation, which are primarily used for knowledge graph completion (KGC). These models aim to fill missing links, either relations $(c_1, ?, c_2)$ or concepts $(c_1, r, ?)$ within a graph. The semantic link prediction tasks discussed in this section, while bearing similarities to KGC, diverge in crucial ways. Firstly, while KGC models rely on the inherent structure of incomplete graphs for training, the models in our current discussion utilise varied input types, ranging from the lexical level to the sentence or corpus level. Secondly, KGC's main objective is to fill missing relations or concepts in existing graphs, whereas our focus here is on predicting relations or concepts without a pre-existing graph containing relation labels. This distinction is vital as we aim to identify relations in contexts where current word association networks entirely lack them and the relation ontology is yet to be fully defined.

Next, we provide an overview of these datasets in Section 2.4.1.1 and introduce the corresponding methods for each task (Sections 2.4.1.2 and 2.4.1.3). Moreover, in Section 2.4.1.3, we delve into a specific type of relation: hypernym, focusing on the task of hypernym extraction. This task involves identifying the fundamental hierarchical relationship between concepts and serves as the primary focus of our Chapter 5.

2.4.1.1 Datasets

The construction of datasets lies in the central of advancing the development of models. High quality datasets are important for both training and evaluation. Here, we review the existing datasets for predicting semantic links between general concepts, including the datasets for relation prediction (lexical and sentence relation level) and concept prediction.

Datasets for lexical relation classification The task of lexical relation classification aims to predict the semantic relation r for a given concept pair (c_1, c_2). Initial studies provided datasets that often focused on a single relation like synonymy (Landauer and

Dumais, 1997), similarity (Finkelstein et al., 2002), or hypernymy (Baroni et al., 2010). However, these datasets were found to be either too narrow or too broad to capture the complex relations among concepts, thus limiting the potential of models to identify a wide range of relationships.

In response to this need, datasets containing more diverse relations have been constructed. For instance, BLESS (Baroni and Lenci, 2011) emerged as a popular dataset that broadened the scope of lexical relation classification. As opposed previous datasets that mainly consider nouns, it includes a wider range of part-of-speech categories (nouns, verbs, or adjectives) and five common relations: *synonym*, *meronym*, *co-hyponym*, *attribute*, and *event*. This dataset has been instrumental in expanding the scope of lexical relation classification.

Furthermore, the DiffVec dataset by Vylomova et al. (2016) encompasses 15 relations by integrating resources from prior work. Similarly, the BATS dataset (Gladkova et al., 2016) introduced a hierarchical relation ontology, distinguishing between morphological and semantic relations with four broad relationships (namely: inflections, derivation, lexicography, encyclopedia) and 40 fine-grained relations. Despite these advancements, the semantic relations mainly cover the taxonomic relations such as those in WordNet (e.g., HYPERNYMS, MERONYMS), more diverse situational or event relations aren't adequately represented (e.g., USEDFOR, HASPREREQUISITEOF). This means that only a limited number of relations related to commonsense knowledge are covered, indicating the need for further improvements in dataset construction.

Classifying lexical relations based solely on two words raises a unique challenge, particularly in the face of word ambiguity. For instance, the words *reading* and *education* could be associated with either the *UsedFor* or the *Causes* relation without context. However, if a context like "Reading regularly enhances the mind and leads to a well-rounded education" is given, the ambiguity can be resolved, and the relationship between *reading* and *education* can be clearly identified as the *Causes* relation. This highlights the necessity of sentence level relation classification, which can provide more context signals.

Datasets for sentence level relation classification The direction of sentence level relation classification for general concepts has been largely shaped by the availability and

characteristics of datasets. Currently, these datasets tend to focus on simple conditions, such as a single type of concepts of a single part of speech (Noun-Noun).

SemEval Tasks (Girju et al., 2007, Hendrickx et al., 2010) have driven this research line, providing datasets that focus on fine-grained relationships between noun pairs. The dataset for SemEval-2007 Task 4 (Girju et al., 2007) was constructed using a combination of text mining and human annotations, with sentences extracted from the web using heuristic patterns to identify specific word pair relations. Those sentences are further verified by human annotators. This task was initially formulated as a binary classification, aiming to determine whether a given relation is expressed in a sentence.

SemEval-2010 Task 8 (Hendrickx et al., 2010) expanded this work by reformulating the task as a multi-way classification, where each example is labelled with multiple relations from a set of ten. This task also provided a larger dataset, with 10,717 annotated examples compared to 1,529 in SemEval-2007 Task 4.

Recently, Dognin et al. (2020) constructed a dataset in commonsense domain by aligning triplets in `ConceptNet` and OMCS sentences, following the distant-supervision approach of Mintz et al. (2009) for aligning the sentences in Wikipedia articles with triplets in Freebase. However, this dataset is not publicly available and, as Han et al. (2020) point out, datasets created by distant supervision can be noisy, which can mislead both training and evaluation.

Despite these advancements, the creation of such datasets remains challenging due to the effort required for human annotations and the complexity of labelling a wide range of semantic relations for general concepts. Consequently, the two SemEval tasks remain the primary publicly available datasets for general concepts, but they are limited to noun-noun pairs.

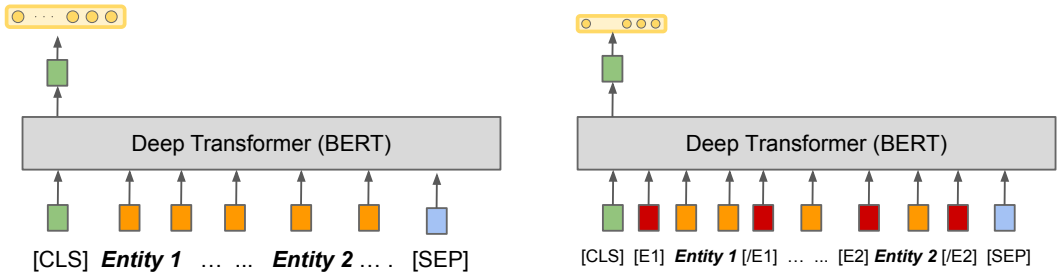
The scarcity of annotated datasets for general concepts across a broad range of semantic relations underscores the critical role of datasets in this field and the challenges faced in predicting relations between general concept pairs. This lack of resources also motivates our study of Chapter 3, where we introduce a sentence-relation parallel data set to facilitate further research.

Datasets for Concept Prediction with a Focus on Hypernym Extraction Concept prediction is the task of predicting a target concept given a concept and a relation type. This task has been widely used for ontology building, e.g., constructing the ontology of animal involves extracting all hyponyms of *animal*. Prior work has explored various relations, including Hypernym (Hearst, 1998), Part-Whole (Girju et al., 2006), Causal Relations (Girju, 2002, 2003), and Property (Baroni et al., 2010, Kelly, 2014).

Among them, hypernymy, a major semantic relation, is noteworthy for its role in organizing human memory and concept categorization (Miller and Fellbaum, 1991, Murphy, 2004). This relation is fundamental for tasks such as taxonomy construction (Snow et al., 2006, Navigli et al., 2010). In Chapter 5, we specifically focus on the hypernym knowledge extraction task.

Datasets for hypernym extraction can be categorized into three types: hand-crafted taxonomies, hypernym detection datasets, and hypernym discovery datasets with large corpora. Traditionally, hypernym extraction from text corpora has been evaluated as part of broader taxonomy evaluation tasks (Bordea et al., 2015, 2016, Hovy et al., 2009), with knowledge bases such as WordNet, Wikidata, or DBPedia employed for evaluation purposes. However, as Camacho-Collados (2017) pointed out, the evaluation of taxonomy often relies on the human judgments of a set of randomly sampled (c_1, r, c_2) triples. This process is costly and hardly reproducible from one taxonomy to another, as they vary in domains and structures.

In response, later studies aimed to simplify and streamline the evaluation process by constructing hypernym detection datasets. In these datasets, a set of hyponym-hypernym pairs is given, and models are evaluated based on whether they can successfully identify the hypernym relation. These datasets are often constructed based on a single or a combination of existing resources such as lexicon and commonsense knowledge graphs like WordNet and `ConceptNet`, as seen in BLESS (Baroni and Lenci, 2011), EVAL (Santus et al., 2015), and LEDS (Baroni et al., 2012), or Encyclopedic knowledge bases such as Wikidata (Vrandečić, 2012) or DBPedia (Auer et al., 2007), as in SHWARTZ (Shwartz et al., 2016).



(a) STANDARD – [CLS] (b) ENTITY MARKERS – [CLS]

Figure 2.15 Architectures for extracting relation representations from deep Transformers network (Baldini Soares et al., 2019). (a) depicts a model with STANDARD input and [CLS] output and (b) use ENTITY MARKERS input while using [CLS].

As a complementary dataset for hypernym detection, the dataset provided in SemEval-2018 Task 9 (Camacho-Collados et al., 2018) is the only dataset specifically designed for retrieving hypernyms from a target corpus for a given input concept. This dataset includes both concepts and entities, spanning three languages (English, Spanish, and Italian) and two specific domains of knowledge in English (Medical and Music).

In Chapter 5, we look at the task of hypernym extraction and use several of the introduced data sets as benchmarks, including BLESS, EVAL, LEDS, and SHWARTZ.

2.4.1.2 Relation Prediction

The task of relation prediction aims to predict a missing relation r that might hold between two concepts (c_1, c_2). One typical approach formulates relation prediction as a classification task, where models are trained to predict the relations between words. This requires labelled datasets as we mentioned in Section 2.4.1.1 to train relation classifiers to classify relationships. For the lexical level, classifiers often rely on training semantic distributional representations such as word embeddings to capture the meanings of words and learn to distinguish different relationships (Weeds et al., 2014, Necşulescu et al., 2015). However, this approach has been criticised for classifiers not truly understanding the relationship but learning individual characteristics of words. For instance, Levy et al. (2015) found that trained hypernym classifiers are prone to hypernym frequency, where common hypernyms (e.g., *people*, *animal*) are often classified as hypernyms regardless of the other word.

For the sentence level classifiers, neural networks such as LSTM (Xiao and Liu, 2016) and CNN (Zeng et al., 2014, Wang et al., 2021a) have been used to obtain sentence representations. Recently, Transformer-based architectures have gained popularity and are being used as sentence encoders to represent sentences. The paradigm of pre-training on large corpora and fine-tuning on downstream tasks have shown to be effective for relation classification tasks. One simple strategy is to use the representation from `[CLS]`[10] as the sentence representation for relation prediction, as shown in Figure 2.15 (a). However, the standard `[CLS]` representation is general for the whole sentence, and not aware of the specific relation between two concepts. Given the limited resources in general concepts domain, models in relation extraction mostly focus on relationships between named entities. Those models attempt to capture both the contextual and entity information by using more entity markers or type information to assist the relation prediction. Highlighting the entities in a sentence has been shown effective in relation classification (Peters et al., 2019, Baldini Soares et al., 2019, Zhang et al., 2019).

Baldini Soares et al. (2019) proposed a general-purpose extractor called "Matching the Blanks", which trains a relational encoder to directly extract relational information from text using BERT as the base model. They explored various strategies for extracting relational information, such as using the representation of the `[CLS]` token as the relation representation, inserting entity markers into the input sentence to augment entity information, as illustrated in Figure 2.15 (b). The model is further trained on BERT to learn similar representations of relations for sentences mentioning the same pair of entities. Although this model was proposed for entities, the results demonstrated that models can successfully transfer to other datasets with further tuning on labelled datasets that contain general concepts, as shown by improvements on SemEval 2010 Task 8 (Hendrickx et al., 2009) dataset. Inspired by the benefits of adding entity markers, our Chapter 3 employs the entity aware version from this framework as a baseline model.

Encoder-decoder models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) have been explored for structure prediction tasks such as relation classification. The idea

[10] `[CLS]` in BERT aggregates all sequence information and is often used for fine-tuning classification tasks, see Section 2.3.1.3.

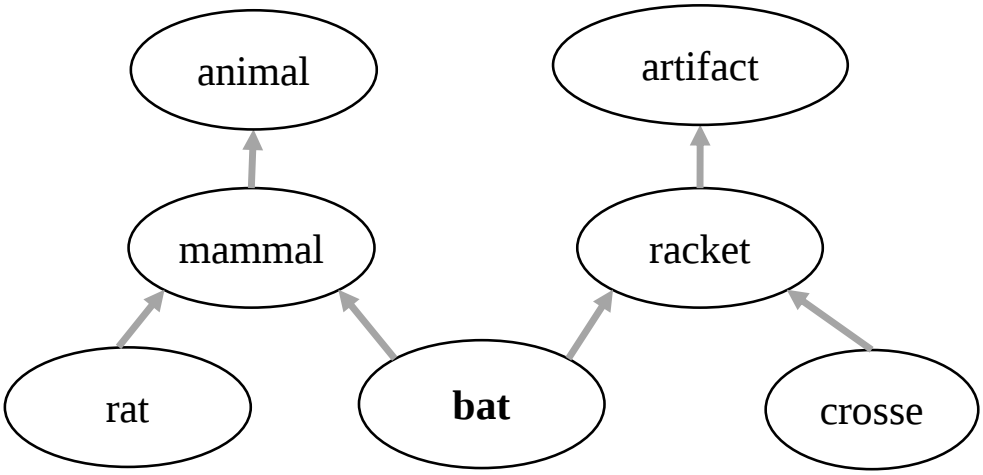


Figure 2.16 An example of identifying different hypernyms for *bat* by using different anchors or siblings (e.g., *rat* or *crosse*).

is to translate raw sentences containing the entities or implicit relational information into a structured relational triplet. For example, given the sentence *my apartment has a large kitchen* fed into the encoder, the decoder is trained to generate the triple (*apartment*, Part-Whole, *kitchen*). This process usually requires the linearization of the relational triplet, i.e., converting the structure format into natural language expression.

Inspired by this, Huguet Cabot and Navigli (2021) proposed the REBEL model, which converts a raw input sentence into a triplet that is linearized into a sequence of text with entity markers to indicate their roles. For example, the triple (*apartment*, Part-Whole, *kitchen*) will be represented as <*triplet*> *apartment* <*subj*> *kitchen* <*obj*> *Part-Whole*. REBEL is fine-tuned based on the BART model and has shown its effectiveness on a widely used relation extraction dataset. We use this linearization approach in our Chapter 3. The advantage of using seq-to-seq models for relation classification lies in their flexibility for formatting the input and output, enabling the generation of multiple triplets simultaneously.

2.4.1.3 Concept Prediction

In the task of concept prediction, given a concept c_1 and a relation r, the aim is to determine which concepts are compatible under this relation. In Chapter 5, we focus on the relation of hypernym and the task of extracting hypernyms.

Hypernym extraction is the task of identifying broader categories for specific words, emphasizing the hierarchical *IsA* relationship between concepts. In this scenario, when provided a hyponym *X* (e.g., *bat*), the objective is to identify its potential hypernyms *Y* (e.g., *mammal*, *animal*).

In the task of hypernym extraction, models identify hypernyms using only a hyponym as input. The complexity of hypernym extraction arises from the limited information available. Approaches in the line of hypernym extraction fall into three categories: (1) corpus-mining with patterns; (2) leveraging word embeddings; and (3) prompting PLMs.

Patterns-based Pattern-based approaches use patterns that express hyponym-hypernym relations to extract hypernyms from large corpora. This line of research was pioneered by Hearst's patterns (Hearst, 1992), which utilise a small set of hand-crafted lexico-syntactic patterns to extract hyponym-hypernyms. One notable example is the pattern *Y such as X and Z*, where *Z* is the sibling of *X* that share the same hypernym *Y*. By using this pattern on the sentence *Mammals such as rats and bats* can extract the hyponym-hypernym pairs (*rats*, *mammals*) and (*bats*, *mammals*). Subsequent work aimed to increase coverage by automatically mining patterns or generalizing patterns with part-of-speech tags or dependency parsing (Snow et al., 2004). Follow-up works show that those patterns are effective in constructing ontologies. For instance, Hovy et al. (2009) reused the pattern (*Y such as X and Z*) to mine hypernyms by bootstrapping, where *Z* is an open slot that shares a hypernym with *X* and is filled with automatically mined words from the Web. By iteratively filling *Z* with different concepts, different *Y* can be mined. This *Z* is named 'anchor', which provides grounding contextual information. Figure 2.16 illustrates this, where providing *bat* with anchors such as *rat* and *crosse* enables the identification of different hypernyms (*mammal* or *racket*). The anchors provide additional information to facilitate the prediction of the correct hypernym. This method of 'anchoring' has been shown to improve the coverage and quality of automatically extracted hypernym knowledge. A similar idea of using anchors to improve hypernym classifiers is used in Snow et al. (2004) and Bernier-Colborne and Barrière (2018). These patterns are of high quality, however, such high-quality patterns are rare and difficult

to identify across all relationships and parts of speech. Additionally, they suffer from low recall because they heavily rely on the lexical co-occurrence between X, Y, and Z, making it difficult to generalise to concepts that have never occurred together in corpora.

Embedding-based The embedding-based approach identifies hypernyms in vector space, taking advantage of extracting word pairs from similar contexts even if they have never been co-observed in text. These embeddings are acquired either through Distributional Semantic Models (DSMs) constructed from large corpora or using pre-trained word embeddings like GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). Given a hyponym, its hypernym can be identified using relation-specific distance measures in the embedding space. To capture the asymmetric characteristic inherent in the hypernym relation, various measurements have been proposed with different assumptions. For instance, the distributional inclusion hypothesis (Weeds and Weir, 2003, Kotlerman et al., 2009) assumes that the prominent contexts of a hyponym should be subsumed within those of its hypernyms. Another example is the distributional informativeness hypothesis (Santus et al., 2014), which posits that a hyponym is more context-specific and therefore more informative than its hypernym. However, these approaches perform inconsistently across different relationships and datasets, making it difficult to select the best measurement (Shwartz et al., 2017).

Prompting-based Prompting-based approach emerged with the recent advancements in PLMs. Prior work (Petroni et al., 2019) has demonstrated that using pre-trained language models for zero-shot prompting is an effective and direct method for retrieving semantic relational knowledge. In a zero-shot setup, the LM is probed using prompts without task-specific labelled examples, relying solely on its general knowledge from pre-training. This method assesses the model's ability of represent relational knowledge and generalise to new tasks.

This approach enables hypernym extraction without dependence on domain corpora. This feature equips them with the capability of generating a list of potential hypernyms for a given hyponym by prompting. For example, given a hyponym *robin* and if we want to know its hypernyms, we can prompt PLMs with an input like *A robin is a type of* `[MASK]` and ask a

language model to predict the most likely words to fill the `[MASK]` token (Ettinger, 2020). We can then evaluate whether a given hypernym, e.g., *animal*, is within the predicted list and calculate its rank within the list. This approach provides an effective way to extract hypernym knowledge.

However, to what extent PLMs understand hypernym knowledge and what is the most effective approach to extract them remains to be explored. Furthermore, the design of prompts greatly affects the outputs of hypernym extraction, as PLMs are sensitive to the surface forms of prompts (Webson and Pavlick, 2022).

Several studies have explored prompting hypernyms from language models, each with different foci and approaches. Consistency across paraphrased prompts has been examined to determine if language models conduct superficial associations based on partial understanding or possess a complete understanding of a concept. Ravichander et al. (2020) focused on the consistency across different morphologies (e.g., singular vs. plural hyponyms) and found that PLMs fail to retrieve consistent knowledge. Hanna and Mareček (2021) investigated the effects of single hypernym patterns (e.g., *X is a Y*, *A Y such as X*) on prompting PLMs in the financial domain and showed that performance varies with patterns. They also proposed a prompt *X is a Y. So is Z*, which mirrors the anchored pattern (*Y such as X and Z*), where *Z* is the nearest neighbour of *X* in FastText word embeddings (Bojanowski et al., 2017). Another focus is testing the properties of hypernym knowledge, where Lin and Ng (2022) targeted the transitive property of hypernym knowledge in BERT and found that models do not always adhere to the transitive rule.

In summary, while the literature shows positive results for hypernym extraction, particularly with the method of prompting PLMs, it also highlights a number of unresolved challenges. This gap motivates our study in Chapter 5, where we delve into marrying the longstanding pattern-based approach from the literature with the recent prompting approach, aiming to tackle these persistent challenges.

Data set	Question and answer options	Knowledge Type
CSQA	*Q1: Where can I stand on a river to see water falling without getting wet?* (A) waterfall (B) **bridge** (C) valley (D) stream (E) bottom	SPATIAL
	Q2: What is the hopeful result of going to see a play? **(A) being entertained** (B) meet (C) sit (D) rush (E) happy	CAUSES
	Q3: How does a person begin to attract another person for reproducing? **(A) kiss** (B) genetic mutation (C) have sex (D) birth of new person (E) marry	HASSUBEVENT
OBQA	*Q1: As a car approaches you in the night, the headlights* (A) remain at a constant (B) turn off **(C) become more intense** (D) recede into the dark	DEFINITION
	Q2: The moon's surface (A) is smooth on the entire surface (B) contains an internal core of cheese (C) is filled with lakes **(D) contains large cavities cause by explosions**	CASUAL
	Q3: What is the most likely to be an effect of acid rain on an aquatic environment? (A) increase in plant growth (B) increase in fish population **(C) decrease in plant life** (D) cleaner and clearer water	PROPERTY
MCscript2.0	T: I am at work. I have a guest sit at the bar. The ordered themselves a beer. I check that he is of age, and that his license is valid. I then go to the beer cooler, and grab a nice cold mug, and fill it up with beer. I place a napkin down and set the beer on top in front of the bar guest. I present him the check and tell him no rush, whenever he is ready. He then places his cash with the receipt. I go to cash him out, offer to be right back with his change, and he responds with, "Keep the change". I like nights like this. *Q1: Why did they receive a nice tip?* **(A) the customer was happy with the service** (B) the customer was in a rush	-

Table 2.3 Examples of multiple choice question-answering tasks. The correct answers are in **bold**. The last column shows the required commonsense knowledge type provided by from the original dataset papers. The symbol - denotes that MCscript2.0 does not provide the relation.

2.4.2 Commonsense Question Answering

Commonsense Question Answering (CQA) refers to a task where models draw upon their "commonsense knowledge" to answer questions presented in natural language. For instance, to answer a question like, "What should you do before going to bed?" models would need to apply commonsense knowledge to answer likely actions such as *brushing your teeth* rather than *eating breakfast*. Performance on CQA tasks is a measure of a model's ability to reason about and understand the world in a manner similar to human cognition. Thus, this task is one effective, and widely used way, to assess the extent of commonsense reasoning in models, contributing to other more complex tasks in NLP such as text summarisation and dialogue

systems. In Chapter 4, we use this task to examine the commonsense reasoning capabilities of recent PLMs, investigating whether word associations can enhance their reasoning ability.

The task of CQA is typically a multiple-choice task, where a model is given a question and a list of possible answers. A model is expected to select the most likely answer option based on commonsense knowledge. Crucially, most answer options are superficially related to the question, and commonsense reasoning is necessary to identify the right one. Table 2.3 shows examples from a variety of CQA datasets.

Addressing CQA questions effectively requires models to go beyond the overt information given in the question and answer and draw upon underlying connections or rationales (Talmor et al., 2019). This implies that models need to comprehend not just the literal semantics of the inputs, but also infer associated implications or assumptions. One substantial challenge in this domain is the implicit nature of commonsense knowledge (Davis and Marcus, 2015), which humans acquire through day-to-day experiences but is often not directly stated in texts, making it difficult for models to learn. Other complexities include reasoning with multiple information fragments, managing ambiguity inherent in natural language, and navigating diverse forms of commonsense knowledge. These challenges necessitate a deep understanding of the connections and associations between concepts in questions and answers, as well as the reasoning behind them.

2.4.2.1 Datasets

To test models' commonsense reasoning ability, a number of benchmark datasets have been proposed (Levesque et al., 2011, Gordon et al., 2012, Zellers et al., 2018, Sap et al., 2019b, Talmor et al., 2019). In Chapter 4, we choose three datasets targeting general commonsense for experiments. See Table 2.3 for examples from these datasets. CSQA (Talmor et al., 2019) is a challenging dataset targeting general commonsense knowledge. The creators asked crowd workers to generate valid questions based on sub-graphs sampled from `ConceptNet`. To correctly answer a question, broad types of commonsense knowledge between concepts are needed, such as spatial, causality, and sub-event relationships (cf., Tab 2.3 for examples). OpenBookQA (Mihaylov et al., 2018) focuses on assessing models multi-hop reasoning

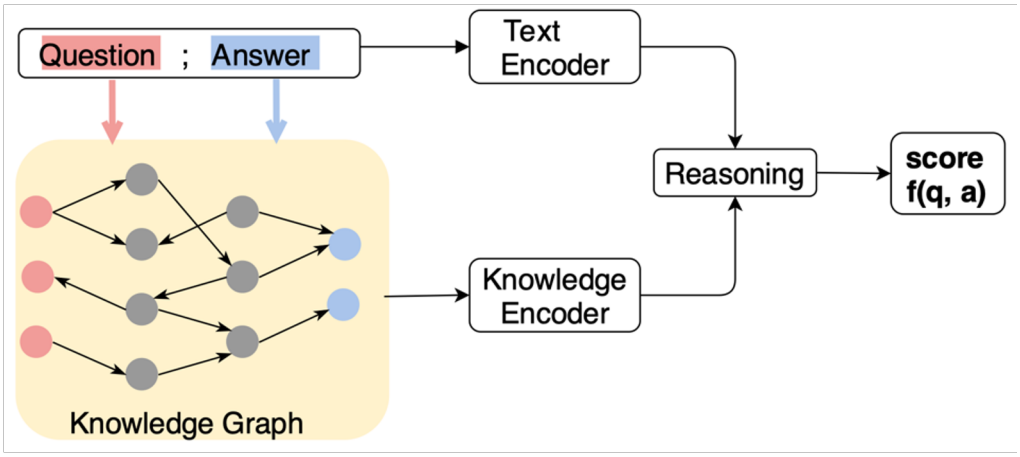


Figure 2.17 A general KG-augmented framework from multiple-choice commonsense question-answering, adapted from Wang et al. (2020).

skills in elementary science questions. Answering these questions requires models to draw on scientific facts from an open book with broad commonsense knowledge. Different from the above two datasets which provide only the question as context for the answer options, MCScript 2.0 (Ostermann et al., 2019) provides a narrative context describing daily events such as *doing the laundry* and *baking a cake*. The questions are designed to require "script" knowledge about everyday situations. Previous work (Lin et al., 2019a, Mihaylov and Frank, 2018) has shown that using knowledge from external knowledge graphs such as `ConceptNet` can improve performance. However, to what extent CQA can benefit from associative knowledge in large-scale word associations remains unknown, motivating our study in Chapter 4.

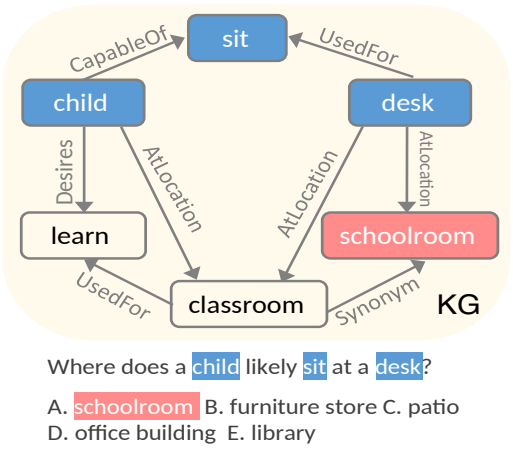
2.4.2.2 Modelling Approaches

On the methodological end, a model typically encodes a question along with each of its answer option separately into an embedding space and generates a score for each answer. The answer option with the highest score is chosen as the correct answer. The method for generating these embeddings has evolved over time. Initially, models that were widely used in natural inference (Chen et al., 2017) and reading comprehension (Seo et al., 2017) were adopted for this purpose (Talmor et al., 2019). These models focused on encoding the interactions between a question and an answer, using bilinear transformation or heuristic operations

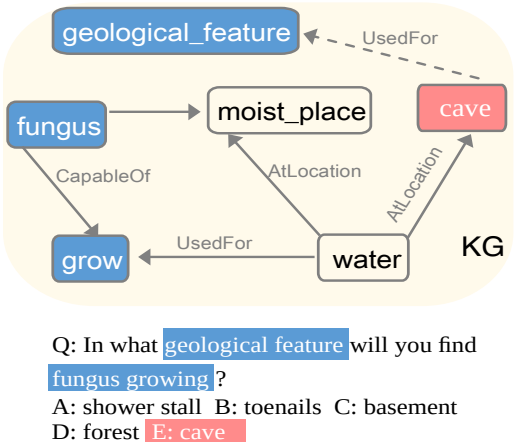
such as element-wise subtractions and multiplication to capture nuanced differences and similarities. Later on, fine-tuned PLMs with an appended classifier were deployed (Talmor et al., 2019). However, a significant gap still exists compared to human performance. To address the deficit of commonsense knowledge, external knowledge from knowledge graphs was incorporated into models (Mihaylov et al., 2018, Lin et al., 2019a). Recently, there has been a line of work that incorporates KGs with PLMs.

A general framework for augmenting QA models with external knowledge graphs is illustrated in Figure 2.17. The framework entails three components: (a) a text-encoder encodes the representations of a question and its answer option; (b) a knowledge encoder that represents the relevant knowledge, retrieved from a KG for a question-answer pair, in vector space; (c) a reasoning module which integrates the representations from (a) and (b) to derive an overall score. The answer with the highest score is selected as the correct answer. Different models vary in the types of knowledge retrieved from a KG and the knowledge encoder used for encoding graph knowledge. They can be broadly classified as static or dynamic models based on whether only existing knowledge can be retrieved or knowledge that doesn't exist in a KG can be dynamically generated. Figure 2.18 illustrates two examples for the two types of models.

In static KG-augmented models, different levels of knowledge can be retrieved, varying from basic concepts (Wang et al., 2019) to relational triplets (Mihaylov and Frank, 2018) or paths (Santoro et al., 2017), to sub-graphs (Lin et al., 2019a). For example, Lin et al. (2019a) employed GconAttn (Graph Concepts Attention, Wang et al. (2019)), a model that applies graph matching for question-answer pairs. This model transforms the text forms of both the question and its answer into two subgraphs by extracting relevant knowledge from a KG. It then aligns the two subgraphs with a concept-level attention and pooling strategy, producing a fixed knowledge embedding. Although GconAttn allows interactions among concepts, it ignores relational knowledge, i.e., the edge labels and the paths that connect the concepts in the KG. To better utilise the relational knowledge in a KG, Wang et al. (2020) employed the Relation Network (RN; Santoro et al. (2017)) to encode knowledge representation by performing path-level attention over paths retrieved from a KG for each pair



(a) Relevant knowledge exists in a KG (Feng et al., 2020)



(b) Relevant knowledge (dashed line) is missing in a KG (Wang et al., 2020).

Figure 2.18 Illustration of two types of situations regarding knowledge: (a) the existing knowledge in a KG is sufficient to answer a question, and (b) the knowledge required to answer the question is missing from the KG, necessitating models to dynamically complete the relevant knowledge. This corresponds to two types of KG-augmented models: (a) a static model that retrieves evidence from `ConceptNet` to answer questions, and (b) a dynamic model that generates missing relations between question and answer concepts as evidence. Blued nodes are retrieved question concepts and red nodes are answer concepts. A solid line denotes an existing edge in a KG, while a dashed line indicates a model-generated dynamic edge absent in the static KG.

of question-answer concepts. While static-KG augmented models can use reliable knowledge in existing KGs, they are reliant on known knowledge and cannot provide additional necessary knowledge on demand.

Dynamic-KG augmented models transfer knowledge from static KGs to a model that can dynamically infer necessary knowledge on demand. This line of models often relies on the power of generative models for learning and inference. Wang et al. (2020) proposed PG-Global to generate dynamic path knowledge embeddings with a pre-trained Path Generator, which is a fine-tuned GPT-2 (Radford et al., 2019) with multi-hop paths sampled from a KG using random walks. The details of training the Path Generator are described in Section 2.3.2.2. This model transfers knowledge from a KG into a language model, which is used as a dynamic KG to generate path representations on-demand. Additionally, they proposed PG-Full (Wang et al., 2020) to unify the static knowledge embedding from RN and the dynamic knowledge embedding from PG-Global.

While these models present diverse strategies for addressing CQA, it remains unknown to what extent they can leverage commonsense knowledge encoded in word associations to improve CQA. We explore this in Chapter 4.

2.5 Discussion and Chapter Summary

2.5.1 Discussion

We have provided a comprehensive review of the relevant materials for our thesis. We now present discussions on several key points and themes of this research, and summarise this chapter.

Variations in Commonsense Knowledge At the individual level, we acknowledge that people with different beliefs and backgrounds possess different basic knowledge. This variation is seen in everyday practices and preferences. By using normalised data collected from a wide range of participants like SWOW (De Deyne et al., 2019), we expect to capture a more uniform and comprehensive representation of concepts within a specific language.

At the cultural level, we acknowledge that commonsense knowledge is cultural specific (Anacleto et al., 2006). For example, consider the use of chopsticks as the primary eating utensils in many Asian cultures, in contrast to the predominant use of spoons and forks in Western countries. This thesis primarily focuses on Western, English-speaking cultures, and therefore, the cultural nuances discussed are largely reflective of these societies.

The Rapid Evolution in NLP At the time of our research, the models we employed were considered state-of-the-art. However, it is important to note that the field of NLP is rapidly evolving. Subsequent to our study, the field has seen the emergence of new, advanced models that have significantly improved performance and capabilities. The pace of this progression is highlighted by the recent strides in large language models, including ChatGPT (Bian et al., 2023), GPT-4 (OpenAI, 2023), and Llama2 (Touvron et al., 2023). These models, trained with human feedback, are designed to follow human instructions and produce outputs

closely aligned with human expectations (Touvron et al., 2023). Consequently, there has been a marked improvement in commonsense question answering tasks, even under zero-shot prompting conditions (Bian et al., 2023, Bang et al., 2023). Nevertheless, these models continue to face challenges, such as in situations that require temporal, event, and social commonsense understanding (Laskar et al., 2023). Additionally, their alignment with human commonsense reasoning, particularly in pinpointing essential commonsense knowledge and rationales for answering questions, is not always consistent. This might cause trust issues for applying these models to real scenarios (Ji et al., 2023). Thus, exploring ways to enhance the understanding of commonsense knowledge within these models and to distil reliable commonsense knowledge remain promising avenues for future research.

Semantic Knowledge in KGs and PLMs Large-scale KGs, as introduced in Section 2.2, and representational models of PLMs, which are reviewed in Section 2.3, represent two pivotal aspects of our research. On the one hand, KGs, being highly interpretable and accessible, offer significant advantages and opportunities for integration with PLMs. Exploring how symbolic knowledge from KGs differs and how it can complement and enhance the capabilities of PLMs form the core of our investigation in Chapter 4. On the other hand, questions such as what knowledge is contained in the hidden parameters of PLMs, and how to effectively extract structured knowledge from PLMs, remain to be explored. This motivates our Chapter 5. Our thesis focuses on comparing KGs and PLMs in semantic knowledge representation and examines their utilities for commonsense reasoning tasks. However, the exploration of how recent advancements in larger language models might reshape this relationship is a critical area for future studies (Pan et al., 2023).

2.5.2 Chapter Summary

In this chapter, we presented the necessary theoretical and methodological background for the main thesis chapters. Specifically, we reviewed the theories and practical implementations of semantic memory organization, with a focus on the semantic network (Section 2.1). Particularly, we reviewed the construction of two large semantic networks, `ConceptNet`

and SWOW, discussing their differences (Section 2.2). In addition, we reviewed the models for learning semantic knowledge from free text using pre-trained language models (Section 2.3.1) and from knowledge graphs (Section 2.3.2). Finally, in Section 2.4, we reviewed two tasks: relation learning and commonsense question answering, which are representative of a suite of benchmarks for machine commonsense reasoning and will be used throughout this thesis.

Next, we address the three research questions outlined in Section 1.1, focusing on understanding the rationales and relations inside word associations (Chapter 3), utilizing large-scale word associations for commonsense question answering tasks (Chapter 4) and leveraging word associations to improve the knowledge extraction from PLMs (Chapter 5).

Chapter 3

Explanations and Relations in Word Associations

In this chapter, we address the first research question: why are certain words associated and what kinds of relational knowledge do word associations contain? Word associations are among the most common paradigms for studying the human mental lexicon (Kent and Rosanoff, 1910, Deese, 1966). While the words that are elicited (Nelson et al., 2004, Kiss et al., 1973, De Deyne et al., 2019) and their general categories – such as paradigmatic, syntagmatic and phonological (Kent and Rosanoff, 1910, Osgood et al., 1954, Namei, 2004, Fitzpatrick, 2006, Fitzpatrick and Thwaites, 2020) – have been well-studied, surprisingly little attention has been given to the question of *why* participants produce the observed associations. Answering this question would not only advance understanding of human cognition, but could also aid machines in learning and representing basic commonsense knowledge.

Motivated by this, we propose a framework to directly elicit reasons for associations from human participants. Building on top of this, we collect a large, crowd-sourced dataset of English word associations with explanations, labelled with high-level relation types, which we refer to as WAX. We present a series of analyses of the provided explanations and design several tasks to probe to what extent current pre-trained language models capture the underlying relations and explanations. Our experiments show that models struggle to capture

the diversity of human associations, suggesting WAX is a rich benchmark for commonsense modelling and generation.[1]

This chapter builds on the paper:

> Chunhua Liu, Trevor Cohn, Simon De Deyne and Lea Frermann. 2022. WAX: A New Dataset for Word Association eXplanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 106–120, Online only.

3.1 Introduction

Word associations (Kent and Rosanoff, 1910, Deese, 1966, Kiss et al., 1973) serve as a prevalent paradigm in cognitive science for probing the human mental lexicon (Nelson et al., 2004, Fitzpatrick, 2006). As introduced in Section 2.1.2.2, they reflect spontaneous human associations between concepts. In a typical study, a participant is presented with a cue word (e.g., *bagpipe*) and asked to spontaneously produce the words that come to mind in response (*music*, ...). Large-scale crowd-sourcing studies have resulted in the creation of numerous word association norms, with a detailed overview available in Section 2.2.2.1. The largest English word association graph, SWOW (De Deyne et al., 2019), covers over 12K cues and 3M responses, drawing from thousands of participants. It serves as a vital resource for understanding basic human conceptual knowledge. This repository of shared associations is used to predict concept similarities (De Deyne et al., 2016c), to probe the gender bias in human minds (Du et al., 2019) and to provide external knowledge for commonsense question answering (Chapter 4 and Liu et al. (2021)).

However, existing word association data sets like SWOW only provide cue-association pairs, but do not further distinguish between different types of associations, leaving the reasons and relations behind associations unknown. To fill this gap, we constructed a novel data set to recover the underlying reasons by collecting associations together with free-text

[1] Data and code are available at https://github.com/ChunhuaLiu596/WAX

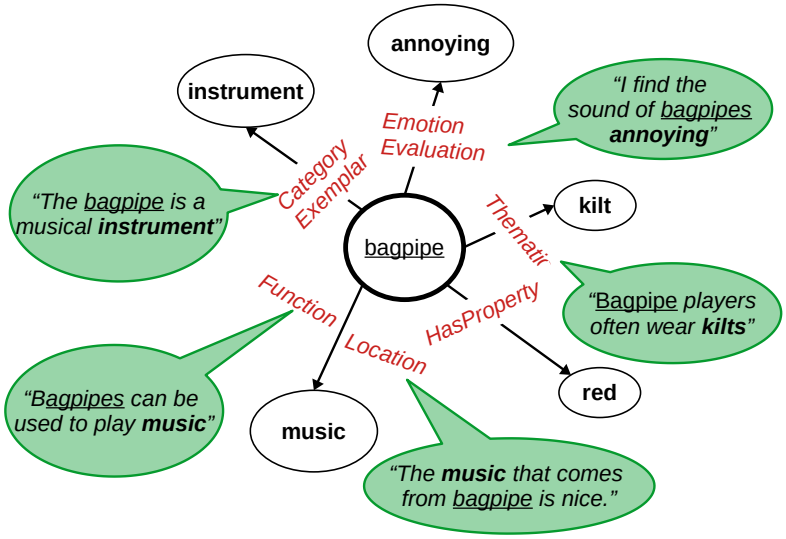


Figure 3.1 Excerpt of WAX, which consists of associations between cue words (bagpipe) and associations (kilt, red, ...) together with association explanations (speech bubbles) and discrete relation type labels (edge labels). Some associations are supported by distinct relation types and explanations (e.g., bagpipe→music).

explanations from human participants, and distill high-level relation types from them. Our data set can enhance our understanding of the *reasons* and *types* for conceptual associations in humans, and can serve as an explicit knowledge resource for reasoning models.

Our data set WAX (Word Association eXplanations) encodes English word associations with diverse explanations and high-level relation types, and is illustrated in Figure 3.1. In a large crowd-sourcing study, we (a) collected human word associations by presenting participants with a cue word (*bagpipe*) and collecting the association words that spontaneously came to mind (*music*, *kilt*, ...) (Figure 3.1, circles); (b) asked the same participants to provide *explanations* in short sentences that describe how the cue is linked with their corresponding associations (Figure 3.1, speech bubbles); and (c) labelled explanations with a relation type adapted from a predefined set (Fitzpatrick and Thwaites, 2020, Wu and Barsalou, 2009, McRae et al., 2012, Speer et al., 2017) (e.g., FUNCTION, edge labels in Figure 3.1). We ensure data quality through several layers of careful annotator training and data filtering.

Compared to existing work on categorising word associations (Piermattéo et al., 2018, Fitzpatrick, 2006), WAX is larger in size, grounds associations in explanations, and has been released to the research community, supporting future research on understanding and mod-

elling conceptual knowledge. WAX complements existing commonsense knowledge graphs, which either involved decades of manual and engineering work (ConceptNet; Speer et al. (2017)), limiting their ability to reflect the natural diversity in human associations (ATOMIC; Sap et al. (2019a)); or only indirectly link concepts via a shared scene (CommonGen; Lin et al. (2020)). WAX results from a new, scalable method of collecting general commonsense knowledge, while maintaining both quality and diversity of associations and explanations, and can be extended to other languages in a straightforward manner.

We annotated a subset of WAX with high-level, discrete relation labels, enabling us to quantify the diversity of human mental relations, and to evaluate machine learning models in their ability to (a) distinguish different relations; and (b) generate plausible association explanations. Our experiments using pre-trained language models demonstrate the value of WAX as a rich and challenging data set for a variety of commonsense modelling and generation tasks. In sum, this chapter's main contributions are:

- A large data set of word associations with free-text explanations, providing the justification for the relation, and relation labels, which can support scalable studies of the human mental lexicon, and the development of models of relation extraction, commonsense knowledge, and explanation generation.
- Extensive experiments demonstrating the utility of WAX for commonsense relation classification and explanation generation.
- Insights into the relative ease of predictability of different relation types, giving rise to future development of targeted models, as well as relation ontologies that are tailored to 'empirical' relations emerging from the data.

With our WAX dataset and experiments, we addressed the first research question of *what relational knowledge is encoded in word associations?* Our findings indicate that the rationales behind associations fit into a high-level relation ontology and are mainly driven by semantic relations.

3.2 Background

Our work relates to several research lines, including word associations, commonsense knowledge graphs, and explainability.

3.2.1 Explaining Word Associations

Word associations, as reflections of human mental lexica, have been studied extensively in psychology (Kent and Rosanoff, 1910, Deese, 1966, Cramer, 1968). Recognizing their importance, researchers have turned to crowd-sourcing large-scale word association datasets across various languages, e.g., English (Kiss et al., 1973, Nelson et al., 2004, De Deyne et al., 2019), Dutch (De Deyne and Storms, 2008b), and Japanese (Joyce, 2005). Among them, SWOW (De Deyne and Storms, 2008b, De Deyne et al., 2019) is the largest multilingual word association graph, covering 18 languages.[2] Section 2.2.2 provides a detailed description of its construction process and approach. These graphs have been widely used as tools to study characteristics and structures of semantic memory (De Deyne et al., 2013a, Steyversa and Tenenbaum, 2005). However, the graphs only include directed associations between word pairs, rendering the underlying reasons for association unknown.

Types of mental associations were previously studied in cognitive psychology (Read, 1993, Sinopalnikova, 2004, Fitzpatrick, 2006, Santos et al., 2011, Yokokawa et al., 2002). While various relation schemes have been developed,[3] the procedure of labelling predominately falls into two categories: (1) cue-association classification and (2) reason-based classification.

The majority of studies (Osgood et al., 1954, Orita, 2002, Yokokawa et al., 2002, Namei, 2004) classify word association types based solely on cue-association pairs, due to the challenges of obtaining reasons behind word associations on a large scale. Typically, researchers gather word associations from participants and then manually annotate cue-association pairs with relations. Such a method necessitates that researchers infer relations from the cue-associations, thereby introducing potential misunderstandings and inaccuracies.

[2] https://smallworldofwords.org/en/project/home
[3] See Section 3.2.3 for further details on different relation ontologies.

Conversely, a few studies (Fitzpatrick, 2006, Piermattéo et al., 2018) adopt the reason-based classification, albeit on a small data scale. In this approach, relations are labelled based on explicit rationales provided by participants who generated the word associations. This research line shows that relations of word associations can be recovered by (1) asking subjects to *explain* (in spoken words or in writing) the reasons for the produced association, then (2) inferring a relation based on the explanations. By directly sourcing explanations from participants, this method minimizes the subjective interpretations of researchers. This not only promises improved labelling accuracy but also ensures a more genuine representation of participants' cognitive processes, making the findings more robust and generalizable. However, the data sets from these studies often were small (e.g., 100 cues) and were not made available to the research community.

We follow the methodology from the reason-based works, both to recover the association reasons (see our method description in Section 3.3) and to label a subset of our word associations with relation types. In contrast with previous work, we provide a large-scale data set by gathering explicit explanations and relation types, to encourage future work on automatic association inference and relation labelling.

3.2.2 Perspectives from Commonsense Repositories

In word association graphs, cue words are typically surrounded by a rich set of associations (60 on average in SWOW) provided by multiple participants responding to the same cue. Naturally, those associations could be considered as shared, basic knowledge, or a source of commonsense knowledge. Equipping machines with such resources has attracted substantial attention (Davis and Marcus, 2015), for instance by incorporating existing resources like ConceptNet (Speer and Havasi, 2012) into models to solve downstream tasks like question answering, as discussed in Section 2.4.2.

However, acquiring such commonsense knowledge is a challenge because it is vastly diverse and not often explicit in language, leading to data scarcity. Commonsense knowledge is typically collected either in free-text format (OMCS: Singh et al. (2002)) or structured databases (e.g., ConceptNet: Speer et al. (2017); ATOMIC: Sap et al. (2019a)). Despite

efforts to collect it in various formats, the implicit nature of commonsense knowledge means these databases often remain less dense than desired, as discussed in Section 2.2.1.1. Therefore, exploring new approaches for acquiring commonsense knowledge is an ongoing research question. Understanding relations in word associations could increase its potential as a source of acquiring commonsense knowledge.

In Section 3.3.2, we explore the underlying associative relations in word associations, leveraging the relation inventory adapted from well-developed relational commonsense knowledge graphs, aiming to better understand the associative network, as motivated in Section 2.1.2.2.

Additionally, as introduced in Section 2.3, pre-trained language models (PLMs) were tested as commonsense repositories (Petroni et al., 2019, Shwartz and Choi, 2020, Bhargava and Ng, 2022) by probing the extent of commonsense knowledge encoded in PLMs or using PLMs to construct (or complete) commonsense knowledge graphs (Malaviya et al., 2020, Zhou et al., 2020b, Hao et al., 2023). Integrating existing knowledge (free-text or structured) with PLMs has been shown effective for improved machine reasoning (Wiegreffe et al., 2022, Moghimifar et al., 2021), and having machines explain why a certain association exists could bridge between structured and text representations. We explore association explanation in Section 3.5.

3.2.3 Relation Inventory

Concept meanings are shaped by their relational links with other concepts (Aitchison, 1994). By understanding these relations, we gain insights into concept meanings and the connections behind them (Kumar, 2021). Extensive research has been conducted to ascertain which relations are most effective in capturing these interconnections among concepts, resulting in various relation ontologies, each with a unique focus. We will introduce the development of these relation schemes, point out their connections, and show their contribution to the relation ontology for our study in Section 3.3. This ties back to topics introduced in Section 2.1.2, including word association categorisation, property generation, and commonsense relations.

	Paper	Relations
Word Associations	Osgood et al. (1954)	**3 main**: Paradigmatic, Syntagmatic, Clang
	Fitzpatrick (2006)	• **4 main, 17 sub-categories:** • Meaning-based: defining synonym, specific synonym, hierarchical/lexical set, quality, context, conceptual • Position-based: consecutive xy, consecutive yx, phrasal xy, phrasal yx, different word class collocation • Form-based: derivational, inflectional, similar in form only, similar form association • Erratic: false cognate, no link
	Cremer et al. (2011)	• **4 main, 17 sub-categories:** • Direct meaning-related: coordinate, subordinate, superordinate, antonym, partonym (part whole), partonym (whole part), context-independent, goal/target, synonym • Indirect meaning-related: subjective association, composite word, context-dependent • Form-based association: change of affix, similar form • Other: non-classifiable, repetition, no response
Property Generation	Garrard et al. (2001)	• **4 main:** Sensory, Functional, Encyclopaedic, and Categorization
	Cree and McRae (2003)	• **4 main, 9 sub-categories:** • Visual: color, parts and surface properties, motion • Sensory-processing channels: smell, sound, tactile, taste • Function • Encyclopedic
	Wu and Barsalou (2009)	• **4 main, 26 sub-categories:** • Entity: associated abstract entity, behavior, external component, surface property, internal component, larger whole, made-of, quantity, systemic feature • Situation: action/manner, associated entity, function, location, origin, participant, time • Taxonomic: coordinate, individual, subordinate, superordinate, synonym • Introspective: affect emotion, cognitive operation, contingency, evaluation, negation

Table 3.1 The overview of relevant relation ontology in various studies.

Word Association Relations Research in cognitive psychology has emphasised the importance of categorising association types. A summary of these categorisations is provided in

Table 3.1. Osgood et al. (1954) introduced three primary relations: paradigmatic (meaning-based), syntagmatic (position-based), and clang (form-based). Specifically, paradigmatic associations connect words like *table* and *desk* by meaning, syntagmatic ones like *reading* and *glasses* by syntactic roles, and clang associations link phonologically similar words like *boat* and *coat*. While paradigmatic words are interchangeable in context, syntagmatic ones are not. This three-category system, though foundational (Wolter, 2001, Orita, 2002, Namei, 2004), has faced criticism for its simplicity and therefore fails to capture the diverse and granular nature of word associations. Motivated by this, Fitzpatrick (2006) divided each broad category into several sub-categories, resulting in 17 types (cf., Table 3.1 Row 2). Additionally, Cremer et al. (2011) split the paradigmatic category into two: direct meaning-related for context-independent inherent taxonomic associations (e.g., *dog–cat* edge is CO-HYPONYMS and *teeth-mouth* edge is PARTWHOLE relation) and indirect meaning-related for subjective or context-dependent associations (e.g., *strong–muscles*). This relation scheme has a high overlap with the relation scheme proposed by Wu and Barsalou (2009) for labelling features in the property generation task, which will be described below.

In our work, we adhere to the four broad categories from Fitzpatrick (2006), namely meaning-based, position-based, form-based, and 'other'. Due to data scarcity, form-based and clang-based associations were excluded during the pilot study stage, while the 'other' category was utilised during data collection but omitted in the final dataset. Refer to Table 3.4 (Linguistic group) for details on the final ontology. In particular, we place emphasis on developing meaning-based relations by integrating relation ontologies from property generation and commonsense knowledge graphs.

Property Generation Relations Property generation is a key task for acquiring semantic features from semantic memory (see Section 2.1.1). For a given concept (e.g., *knife*), the task of property generation is to generate a list of properties (e.g., *is sharp*, *used for cutting*, $\cdots$). To better understand and categorise these features, researchers have devised various relation schemes (cf., Table 3.1 bottom).

Garrard et al. (2001) proposed a four-way categorisation encompassing sensory, functional, encyclopaedic, and categorising information. In a different approach, Cree and McRae (2003) suggested categorising them into nine types based on brain regions. A more detailed hierarchical scheme was introduced by Wu and Barsalou (2009) for concrete living concepts, such as a dog. This scheme comprises 26 sub-categories within four main categories: entity, situation, taxonomic, and introspective. For clarity, the taxonomic category mostly aligns with a subset of relations in WordNet, including CO-HYPONYMS, HYPERNYMS, and HYPONYMS. The entity category captures inherent features of the concept, like PARTOF and MADEOF. Situational relations, on the other hand, focus features co-occurring in specific contexts, such as LOCATION for *a cupboard in a kitchen*. Lastly, introspective relations capture subjective features, such as emotions and evaluations. See Table 3.1 (last row) for the full list of fine-grained relations. This comprehensive scheme has gained widespread acceptance in subsequent research (McRae et al., 2005, Kremer and Baroni, 2011, Recchia and Jones, 2012, Bolognesi et al., 2017).

Our interest in this line of relation schemes stems from the observation that property generation often mirrors word associations. Prior study (Santos et al., 2011) has noted this overlap, suggesting that early properties in the generation process often arise from a word association process. Furthermore, properties generated later typically describe objects and situations. This overlap between word associations and property generation is evident in the relation coding scheme proposed by Santos et al. (2011), which includes four categories: taxonomic, linguistic, property, and situational, along with an additional 'other' category. The linguistic class corresponds to a mixture of meaning-based (e.g., SYNONYM, ANTONYM), syntagmatic-based (e.g., forward and backward compounds like *reading-glasses* and *bee-honey*), and form-based association relations (e.g., root or sound similarity). This scheme indicates the intertwined nature of word associations and property generations. Therefore, in our study, we adopted three broad categories from Wu and Barsalou (2009), namely situational, taxonomic, and concept-properties (cf., Table 3.4 Column 1). The fine-grained relations for each category are developed by combining multiple relevant studies (Wu and Barsalou, 2009, Speer et al., 2017, Cremer et al., 2011, Fellbaum, 2010).

Relation	Definition	Examples
Hypernym	From concepts/events to superordinates	breakfast → meal, fly → travel
Hyponym	From concepts/events to subtypes	meal → lunch, walk → stroll
Instance Hypernym	From instances to their concepts	Austen → author
Instance Hyponym	From concepts to their instances	composer → Bach
Part Meronym	From wholes to parts	table → leg
Part Holonym	From parts to wholes	course → meal
Derivation	Lemmas w/same morphological root	destruction ↔ destroy
Entails	From events to the events they entail	snore → sleep
Antonym	Semantic opposition between lemmas	boy ↔ girl, buy ↔ sale

Table 3.2 Core WordNet Relations, Definitions, and Examples, taken from Jurafsky and Martin (2023). The direction of the arrows indicates whether the relation is asymmetric (→) or symmetric (↔).

Relations in Relational Semantic Networks In relational semantic networks, the relation schemes primarily revolve around the two expansive networks discussed in Section 2.2. Comprehensive details of these relation schemes are elaborated in Table 3.2 and Section 2.2.1.1 (Table 2.2). The evolution of the `ConceptNet` relation schemes not only encompasses relations from WordNet but also extends them leveraging sentences sourced from OMCS. Therefore, we used relations in `ConceptNet` as a reference when developing our relation ontology.

Our interest in this line of relation scheme lies in its well-developed and its overlap with relation schemes used in word associations and property generation, as shown in several studies. For instance, Vinson and Vigliocco (2008) developed a relation scheme to categorise the semantic features for both objects and events (or nouns and verbs), and found that many features reflect the semantic relationships used in WordNet, e.g., HYPERNYM and PART HOLONYM. McRae et al. (2012) provided an overview of the semantic relations and associative relations, showing that the taxonomy relations (SYNONYM, ANTONYM, HYPERNYM, CO-HYPONYMS), exist in property generation and word associations also exist in WordNet (Miller, 1985). The authors further compare the coding scheme in Santos et al. (2011) against an integrated semantic relation scheme and found a high overlap between them. Consequently, a unified relation scheme for associative relations in word associations and semantic relations has been advocated.

3.2.4 Explainable Commonsense

Previous work used generated explanations to improve downstream task performance, e.g., on question answering (Shwartz and Choi, 2020) and natural language inference (Rajani et al., 2019). Less research has attempted to generate explanations to construct structured commonsense resources. Dognin et al. (2020) align ConceptNet with OMCS using heuristic rules and propose dual learning to transfer between a knowledge graph and free text. However, their language data is templated, and their dataset is not public. Other work has retrieved representative contexts from large corpora (Hendrickx et al., 2009), or used templates to construct sentences from triples (Petroni et al., 2019). In Section 3.5 we use WAX to generate explanations that reflect the naturalness and diversity of human explanations.

Another related data set, CommonGen (Lin et al., 2020), consists of crowd-sourced, short sentences describing a scene that includes a given set of concepts (common objects and actions). CommonGen is designed to test machines' compositional ability, but relations between concepts are implicit in the description. Compared to their work, WAX is more explicit, eliciting concept associations from workers directly; more specific as each explanation focuses on a relation between an association pair; and more general (incl. adjectives, adverbs, and abstract concepts). WAX could hence be used to augment knowledge graphs like SWOW (De Deyne et al., 2019) with relation labels, or free-text explanations.

3.3 The WAX Corpus

We present our two-stage framework for collecting word association relations between pairs of concepts (words) by crowd-sourcing explicit explanations of the relations (Figure 3.2). In Phase 1, we collect associations and free-text explanations to elicit the underlying reasoning. In Phase 2, we label a subset of (cue, association, explanation)-triples (c, a, e)[4] with relation types r to characterise the inventory of common relation types. We collect the WAX dataset by crowdsourcing via Amazon Mechanical Turk (MTurk).[5] Participants were informed what

[4]Throughout the chapter, we use c, a, e, r to denote cue, association, explanation and relation respectively.

[5]Our study received ethics approval with the application reference number of 2021-22495-22206-5 from The University of Melbourne ethics review board.

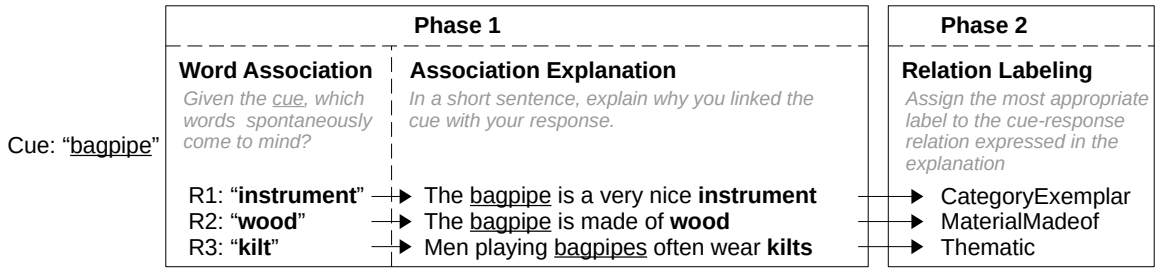


Figure 3.2 Data collection framework for WAX: In Phase 1, participants give associations in response to a given cue word and subsequently provide explanations. In Phase 2, a separate group of participants labels the relations. Table 3.4 lists the full relation ontology and detailed definitions.

data will be collected, how the data will be processed and used, and asked for their explicit consent. To avoid potential confronting content, we removed profane words[6] before sampling cue seeds in Phase 1 (Section 3.3.1).

3.3.1 Phase 1: Eliciting Explanations

We adopt the reason-based classification approach, as introduced in Section 3.2.1, by following prior work (Fitzpatrick, 2006, Piermattéo et al., 2018). In phase 1, we collect both (a) word associations and (b) their corresponding explanations from the same annotator. Here, an *explanation* is defined as a one-sentence justification illustrating the rationales behind the provided word association. By sourcing both the association and its explanation from the same individual, we ensure a direct and genuine insight into the thought process behind the association.[7]

In our data collection, we follow the procedure described by De Deyne et al. (2019) to collect the top three associations. Given a cue word, a worker first generates up to three spontaneous associations (Figure 3.2, left), and immediately after provides the aforementioned explanation of *why* they linked the cue and each association (Figure 3.2, center). The

[6] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

[7] While we could have annotated existing word associations with explanations, this would require inference of another person's reasons for the association. To remove this confound, we elicit associations and explanations from the same worker.

resulting explanations will serve as our text corpus of sentences expressing relations between concept pairs. More examples can be seen in Table 3.8.

We used a set of 1,100 single-token cues, sampled from SWOW, ensuring a balanced distribution over the POS tags noun, verb, adjective, and adverb; as well as abstract and concrete concepts.

Task Design and Payment We use MTurk to collect explanations for the sampled cues by presenting each cue to 10 different workers. We present an annotation batch with 5 randomly sampled cues. During the annotation, each worker (1) produces up to three associated words for each cue, and (2) writes an explanation for each association, as shown in Figure 3.2 (left). Workers can skip cues if their meaning is unknown, or provide fewer than three responses if they cannot think of more. Each batch is paid with $0.66 reward with extra bonus up to $1, depending on the number of known cues, associations, and explanations. Each task takes approximately 5 minutes, as estimated in a pilot study. We paid an average of $1.48 per batch, resulting in an hourly wage of $17.76 (all amounts in US dollars), which is higher than both the US ($7.25 USD) and Australian minimum hourly wage ($14.76 USD).

Quality Control Word associations and underlying reasoning are subjective, hence standard quality assessment via annotator agreement does not apply. Instead, we introduced a number of strategies to control quality: clear guidelines, careful selection of workers, and filtering of undesired explanations.

Our guidelines provide a comprehensive task description accompanied by examples. These guidelines can be further explored in the Appendix, where we've included a screenshot of the annotation interface.[8]

For worker selection, we adopted a two-step approach. We first chose fluent English speakers from five countries (United States, United Kingdom, Canada, Australia, and New Zealand) with over 95% HIT approval rates. We then conducted a pilot study where each worker was assigned a task containing five cue words. For each cue word, they were expected to generate up to three associations. From the results, we selected workers who were familiar

[8]See Appendix A for the key instructions. The complete guidelines are also available in our github repository.

with all the cue words and successfully produced a total of at least ten associations. This pilot study ensured that we selected proficient workers well-acquainted with the task.

After collecting the explanations, we retain only valid explanations. A valid explanation must (1) include the cue and association words, or a morphological variant (e.g., plural); (2) be a single sentence of 5 to 20 words. We removed explanations that did not meet the criteria above or follow trivial templates, and batches where more than 3 of the 5 cues were marked *unknown*.

The final data set includes the annotations of 258 workers and comprises 15K unique cue-association pairs along with 19K explanations. For more data set statistics, see Table 3.3 (left).

3.3.2 Phase 2: Relation Labelling

Labelling relations inherent in word associations has long captivated the cognitive psychology community. Such labelling provides a valuable tool, enabling studies of language development and highlighting differences in conceptual understanding across diverse groups. Building on Phase 1, Phase 2 aims to understand the high-level relation types within these explanations. We augment the dataset above with explicit relation labels (Figure 3.2, right). Given a triple of cue, association, and explanation (c, a, e), annotators choose the most appropriate relation type from a fixed relation inventory. This is crucial, serving as a lens into the variety of underlying association types and shedding light on the primary relations that drive human cognition in word associations. As previously discussed in Section 2.4.1.1, the lack of parallel sentence-relation datasets poses challenges for model development concerning conceptual relations. By providing a human-annotated dataset with explanation-relation pairs, we pave the way for automating the labelling of word associations. This human-curated data serves as a test bed, evaluating machines' abilities in extracting or generating word association relations. We will now introduce the relation inventory, and then describe the process of relation labelling.

	Full WAX	Relation Labelled
# unique a	6,128	453
# unique (c,a)	15,337	520
# unique (c,a,e)	19,228	725
Vocab size	10,180	1,656
Avg len(e)	9.71	10.1

Table 3.3 The statistics of the full WAX, and its subset labelled with manually relations. Avg len(e) is the average length of the explanation (in words).

Relation Inventory As introduced in Section 3.2.3, we consider relation ontologies from various sources. We adapt (a) an established semantic relatedness taxonomy of 28 relation types from cognitive studies of the human mental lexicon (Wu and Barsalou, 2009, McRae et al., 2012), and (b) linguistic relation types (PHRASE, SOUND-SIMILARITY and LEXICAL) from word associations (Santos et al., 2011) and (c) two event-related relations (HASPREREQUITISTE and RESULTIN) from ConceptNet (Speer et al., 2017). These relations reflect four broad categories that were defined in previous work (McRae et al., 2012, Wu and Barsalou, 2009), including TAXONOMIC, SITUATIONAL, CONCEPT-PROPERTIES, and LINGUISTIC. See Table 3.2.3 for their detailed subcategories. In multiple pilot annotations, we assessed the confusability and applicability of the relations to our association data. We conflated associations which were (i) similar (e.g., ACTION and BEHAVIOR), (ii) rare (e.g., ORIGIN), (iii) of opposite directionality (e.g., PARTOF and LARGERWHOLE), since this distinction was often not reflected in the explanations. For example, the explanation "even the sweetest cherry has a pit" is applicable to both *cherry* $\rightarrow$ *pit* and *pit* $\rightarrow$ *cherry*. The final relation ontology consists of 16 relation types,[9] as shown in Table 3.4, which includes the definition of each relation.

Relation labelling We sampled 757 (c, a, e) triples from the data from Phase 1, excluding recurring template-like explanations (e.g., "A is a type of B"). This results in a challenging data set that mirrors the diversity and complexity of real-world language. Moreover, it necessitates a deep understanding of the relationships within explanations beyond mere

[9]We introduced a 'None-of-the-Above' relation to account for relations not covered by our predefined set. However, annotators rarely used this option.

Group	Relation	Definition/Example
Concept-Properties	HASPROPERTY	Cue has association as a property; or the reverse. Possible properties include shape, color, pattern, texture, size, touch, smell, and taste; or inborn, native or instinctive properties. *pineapple-yellow* | *pineapple is yellow inside.*
	PARTOF	A part or component of an entity or event. *restroom-toilet* | *every restroom has a toilet.*
	MATERIALMADEOF	The material of something is made of. *paper-trees* | *paper is made from trees.*
	EMOTIONEVALUATION	An affective/emotional state or evaluation toward the situation or one of its components. *afraid - monsters* | *little kids are frequently afraid of monsters.*
Situational	TIME	A time period associated with a situation or with one of its properties. *bless - birthday* | *they bless her on the birthday.*
	LOCATION	A place where an entity can be found, or where people engage in an event or activity. *shark-sea* | *sharks usually live in the sea.*
	FUNCTION	The typical purpose, goal or role for which cue is used for association. Or the reverse way. *blender-mix* | *i use the blender to mix ingredients.*
	HASPREREQUISITE	In order for the cue to happen, association needs to happen or exist; association is a dependency of cue. Or the reverse way. *smoke-burning* | *if you see smoke, something is definitely burning.*
	RESULTIN	The cue causes or produces the association. Or the reverse way. A result (either cue or association) should be involved. *poison-death* | *too much poison can cause death.*
	ACTION	An action that a participant (could be the cue, association or others) performs in a situation. Cue and association must be among the (participant, action, object). *rabbit-hop* | *rabbits hop along everywhere they go.*
Taxonomic	CATEGORYEXEMPLAR	The cue and association are on different levels in a taxonomy. *rabbit-animal* | *a rabbit is a small type of animal.*
	SAMECATEGORY	The cue and association are members of the same category. *summer-spring* | *summer and spring are both seasons.*
	SYNONYM	The cue and association are synonyms. *gently-tenderly* | *gently has a similar meaning as tenderly.*
	ANTONYM	The cue and association are antonyms. *objective-subjective* | *the opposite of objective is subjective.*
Linguistic	COMMONPHRASE	The cue and association is a compound or multi-word expression or form a new concept with two words. *north-face* | *north face is a very popular brand of clothing.*

Table 3.4 The ontology and definition of associative relations used for labelling WAX. The examples are followed after the definition with the format of *cue-association* | *explanation*.

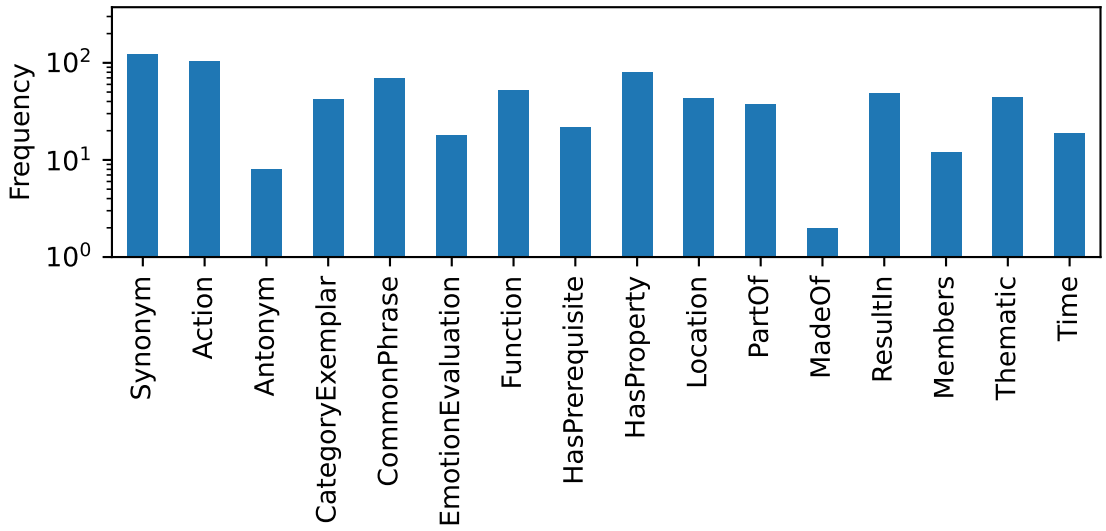


Figure 3.3 Relation distribution of WAX human labelled data.

superficial pattern matching. We included cues with all POS from Section 3.3.1 except for adverbs.[10] Examples are provided in Figure 3.1 and Table 3.8.

Task design and Payment MTurk annotators were given the 16 relation types, their definitions (Table 3.4), and examples. Each batch consisted of 30 (c,a,e) tuples, and a worker selected the most appropriate relation per tuple. This task takes approximately 22 minutes, based on a pilot study. Each batch is paid at a minimum \$1 with extra bonus up to \$8, depending on the quality of the annotation. We paid an average of \$5.92 per batch, resulting in an hourly wage of \$17.36.

Quality Control To ensure the highest quality of annotations, we adopted several strategies: (a) We provided detailed instructions to the annotators. (b) We conducted a training phase by sampling a set of triplets (c,a,e). These were annotated by two authors who ensured that a consensus was reached on the relation labels, which were then treated as ground truth. Using this labelled set, we trained our crowd workers. They were presented with the triples, and their annotations were compared with our ground truth to assess accuracy. (c) We selected 10 reliable crowed workers who achieved accuracy > 0.5 in the training phase. All workers, regardless of their selection status, were compensated for their contributions with

[10]Associations with adverbs have received little attention and are not well-covered by existing relation ontologies.

Type	Cue	Association	Explanation	{Annotation: count}
Retained	honey	sweet	honey is a very sweet substance.	{HASPROPERTY: 5}
	gypsy	europe	gypsies now mainly live in europe.	{LOCATION: 4, THEMATIC: 1}
	baked	fried	baked and fried are two ways to prepare food.	{SAMECATEGORY: 3, THEMATIC: 1, PARTOF: 1}
Discarded	buddy	together	buddies love to spend time together.	{ EMOTIONEVALUATION: 1, RESULTIN: 1, THEMATIC: 1, LOCATION: 1, HASPROPERTY: 1}
	breath	oxygen	when you breath you inhale oxygen.	{RESULTIN: 1, Action: 1, PARTOF: 1, HASPREREQUISITE: 1, FUNCTION: 1}
	faithful	committed	being faithful in a relationship involves being committed to the other person.	{SAMECATEGORY: 1, THEMATIC: 1, PARTOF: 1, SYNONYM: 1, HASPREREQUISITE: 1}
	staff	employed	staff is the people employed by a particular organization.	{PARTOF: 1, THEMATIC: 1, HASPREREQUISITE: 1, SAMECATEGORY: 1, FUNCTION: 1}

Table 3.5 Samples of retained (top) and discarded (bottom) instances in WAX labelled set. The Annotations column indicates the labels assigned to the instance together with assignment count out of 5 annotations.

an hourly wage of $17.36. (d) Throughout the annotation process, we maintained continuous feedback to the annotators. (e) For each (c,a,e), we collected labels from five workers for each (c,a,e). If a label has three or more votes it is selected; otherwise the instance is labelled by two authors of the study, and the voting test is re-applied. After this, 28 instances are still not assigned a label with three votes, and are discarded. See examples of retained and discarded in Table 3.5. We obtained an annotator agreement (pairwise Cohen's κ) of $\kappa = 0.42$, indicating moderate agreement.

The final labelled data set consists of 725 (c,a,e)-triples, covering 520 unique (c,a) pairs, labelled with one of 16 relations. The corresponding relation distribution is shown in Figure 3.3, showing that the relations are present in the data to varying degrees (e.g., the top 4 relations cover 52% of the overall labelled data). Table 3.3 presents the statistics of the full WAX. Our collected data, made public for research purposes, does not contain personal information except for the worker ID, which we have redacted from the dataset.

Questions and Examples
Q1: Does the explanation express a valid reason for associating (c,a) ? Example: raspberries can be made into jam.
Q2: Does the relation label express a valid relation for (c,a) ? Example: (nature, beautiful, HASPROPERTY)
Q3: Does the relation label express the relation for (c,a) that is described in the explanation? Example: (space, stars, PARTOF, space has a lot of stars in it.)

Table 3.6 Questions and examples for WAX dataset quality check.

Criteria	**WAX**	**Random**
Q1: e valid explanation for (c,a)	0.98	-
Q2: r valid relation for (c,a)	0.79	0.26
Q3: r valid relation for (c,a,e)	0.76	0.20

Table 3.7 Manual validation accuracy for assessing explanations and their relation labels, as well as whether they are concordant with the cue and association pair. Also, the judged accuracy of instances with randomly corrupted relation labels is also presented in the Random column.

3.3.3 Corpus Analysis

Quality In a final round of quality control, we examined the overall consistency of WAX. We designed three questions to manually examine its key elements: explanations, relations, and their alignment (see examples in Table 3.7). Q1 asks whether the generated explanation expressed a valid relation for the (c,a) pair. Q2 verifies the relation label quality by asking whether the given relation is valid for the (c,a) pair. Q3 looks into the alignment between explanations and relations by asking whether the explanation e indeed expresses the relation label r.

We presented a random sample of 100 (c,a,e,r)-tuples from WAX to two qualified annotators[11] to answer the three questions. We additionally mixed in 50 (c,a,e) with a randomly assigned relation label r, as a reference point for random performance.[12] Table 3.7 shows the results. We can see that almost all explanations express valid links between cue and association (Q1), demonstrating the validity of the explanations from Phase 1. Close to 80% of the relations are considered valid for (c,a) (Q2) and (c,a,e) (Q3). To put this in

[11] One native speaker of English who was not involved in the project, and one of the authors.
[12] Note that the explanation for (c,a) was not randomised as this would have resulted in a trivial baseline.

perspective, the respective accuracy on the random sample was significantly lower. To the best of our knowledge, WAX is the first large-scale data set with explanations of conceptual associations.

Diversity Conceptual associations may result from factual knowledge, cultural or societal norms, or individual experiences. Here, we analyse the extent to which different annotators produced divergent associations and/or explanations (cf., the *bagpipe* $\rightarrow$ *music* association in Figure 3.1). The presented numbers are a lower bound on diversity, because WAX was collected from a small number of MTurk annotators, which were themselves not screened for diversity and are likely a homogeneous group of (western) English native speakers.[13]

15% (N=2358) of the (c,a) pairs in the full WAX[14] were produced by more than one annotator (3.5 times on average), raising the question whether a single typical relation or multiple distinct ones connect these concepts. We look into this by examining the labelled subset. For 59% (N=51) of these ambiguous (c,a) pairs, all annotators expressed the same underlying relation. Examples include (*grater*, *cheese*, FUNCTION), (*flowing*, *water*, ACTION) and (*reading*, *glasses*, COMMONPHRASE). For the remaining 41% (N=36) annotators expressed between 2 and 5 *different* relations. An example is the pair (*goalie*, *save*) produced by three annotators, with relations FUNCTION ($1\times$) and ACTION ($2\times$). Table 3.8 presents the above examples together with explanations in WAX.

Analysis revealed that in cases where *different* relations emerged for the same (c,a) pair, these relations were predominantly situationally related. This observation suggests that the relations associated with a (c,a) pair are dynamic and influenced by context. To illustrate this, we quantify to what extent two broad categories often cooccur. Figure 3.4 shows that the (situational, situational) is the most prevalent, succeeded by (situational, taxonomic) and (situational, concept-property). Additionally, certain fine-grained relation pairs appear with higher frequency, e.g., (FUNCTION, LOCATION), (ACTION, HASPREREQUISITE), (HASPROPERTY, PARTOF), and (HASPREREQUISITE, SYNONYM). To provide concrete

[13] We removed another layer of potential ambiguity in Phase 2, where we assigned a single label to each association by majority voting, even though some explanations may support several underlying relations.

[14] 16%(N=87) in the labelled proportion, accounting for 43% (N=312) of the labelled (c,a,e,r) tuples.

Concept Pairs	Explanations and Labels
(*grater*, *cheese*)	(1) a grater is great to make shredded cheese. (FUNCTION) (2) he shredded the cheese with the grater. (FUNCTION) (3) i use a grater to grate cheese for my meal. (FUNCTION)
(*flowing*, *water*)	(1) the water is flowing down the gutter. (ACTION) (2) water flows when you turn on the faucet. (ACTION) (3) water is often seen flowing through hills and valleys. (ACTION)
(*reading*, *glasses*)	(1) he needs his reading glasses. (COMMONPHRASE) (2) my father needs reading glasses. (COMMONPHRASE) (3) the old man had to use reading glasses as it was difficult to see up close. (COMMONPHRASE)
(*goalie*, *save*)	(1) another job of the goalie is to save the shots on the goal. (FUNCTION) (2) the goalie reached his glove out and made a big save. (ACTION) (3) the goalie had a great night, making a save on all but one of the shots he faced. (ACTION)
(*igloo*, *cold*)	(1) an igloo is very cold to the touch. (HASPROPERTY) (2) the igloo is a cold place. (HASPROPERTY) (3) when it's cold, you can build an igloo out of snow. (HASPREREQUISITE)
(*heaven*, *god*)	(1) heaven is where god lives. (LOCATION) (2) heaven is the place where one can be with god. (LOCATION); (3) it is said that heaven and hell were created by god. (ACTION)
(*correction*, *fix*)	(1) if you are making a correction, you are fixing something. (SYNONYM) (2) to fix the mistake is a needed correction. (HASPREREQUISITE) (3) in order to make the correction, he had to fix his mistake. (HASPREREQUISITE)

Table 3.8 Example WAX (c,a) pairs produced by >1 annotator, each with three explanations (1)–(3) and corresponding relation labels. The first three examples are *unambiguous* associations, where all explanations describe the same relation, while the last four are *ambiguous*, with explanations covering distinct relations.

examples of this phenomenon, refer to Table 3.8 (bottom), which showcases specific instances where these relations manifest differently based on context.

In Section 3.4 we explore the task of association relation classification, and evaluate our models on the challenging, ambiguous subsets described above to gauge the extent to which associative ambiguity is captured in different transformer-based classifiers.

3.3.3.1 Clustering Explanations

While classifying associative relations into a pre-defined ontology is an important task, both for comparability with prior cognitive work, and for model development and evaluation, it is informative to also group explanations in a purely data-driven way and compare the result

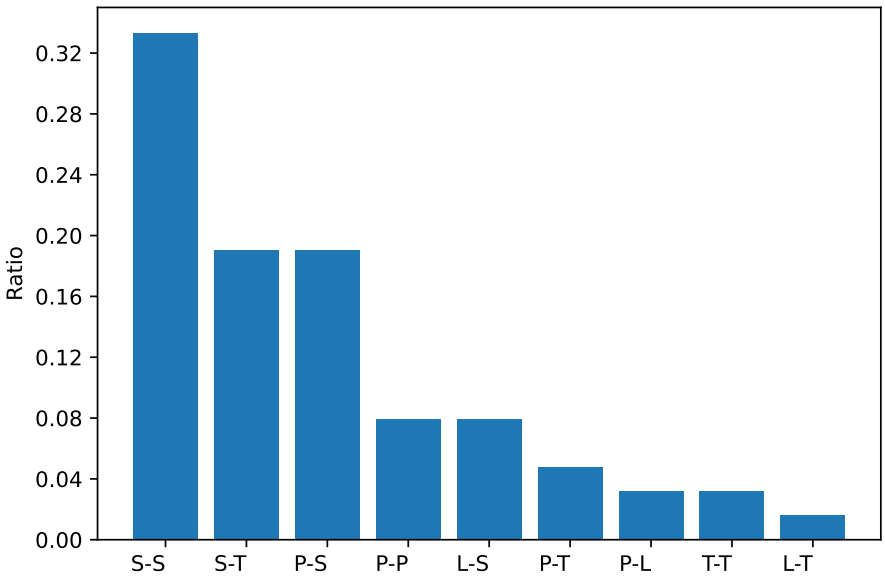


Figure 3.4 Distribution of two coarse relations occurring concurrently for the same (c, a) pairs. S means Situational, T denotes Taxonomic, P means concept-property, and L stands for Linguistic.

with established relation inventories. To this end, we cluster all 19K WAX explanations using K-means into 75 clusters.[15] In order to abstract away from signals specific to cue and association words, and focus on the general 'linking information', we masked cue and association tokens in the explanations and embedded the result with BERT-base (mean pooling over the final layer). We visualised each cluster by its top TFIDF trigrams.

Table 3.9 summarises the clustering results. We compared our clusters with our relation ontology, observing that it is broadly aligned with some categories in our relation ontology in Table 3.4. Some clusters capture the relations in our ontology directly (LOCATION, FUNCTION), although some relations are conflated (SYNONYM, ANTONYM). One general 'similarity'-focused cluster emerged, confirming previous findings on the tendency of English native speakers to associate words based on general meaning similarity (Fitzpatrick, 2006). A second set of clusters captures 'generic associations' (GENERIC 1-2) such as 'If you are c then you a' or 'c is associated with a'. The third (smallest) set is topical, with explanations

[15]We experimented with smaller numbers of cluster but found that this number produced the most nuanced representations, and tried TFIDF instead of BERT embeddings which lead to highly skewed cluster memberships.

Cluster	Representative TF/IDF 3-grams
{SYNONYM, ANTONYM }	'the opposite of' 'opposite of is' 'is the opposite' 'is synonym for' 'another word for'
SIMILAR	'has similar meaning' 'similar meaning as' 'as has similar' 'meaning as has'
LOCATION	'keep my in' 'my in my' 'put my in' 'on my face' 'many in my'
FUNCTION	'be used to' 'can be used' 'when you have' 'there is usually' 'in order to'
ACTION	'in charge of' 'charge of the' 'was in charge' 'the helped the'
TIME	'am about something' 'if am about' 'if something will' 'something will happen'
GENERIC1	'when you are' 'if you are' 'something you are' 'it when you'
GENERIC2	'referred to as' 'associated with being' 'think of as' 'in the past'
TOPICAL1	'in movie called' 'starred in movie' 'was in movie' 'books and movies'
TOPICAL2	'the game the' 'of the game' 'the ball in' 'to catch the' 'the game was' 'to win the'

Table 3.9 Representative sample of explanation clusters, represented by the top TF/IDF 3-grams. Each row corresponds to a cluster, with the cluster names manually labelled. Top: clusters aligning with predefined relations; center: topic-like clusters; bottom: generic clusters.

referring to GAMES (sports) or ENTERTAINMENT (movies and music). Overall, we find that taxonomic and event-related (HASPREREQUISITE, RESULTIN) relations are well-captured, while property relations (PARTOF, HASPROPERTY) are reflected to a lesser extent.[16] This observation aligns with research showing that personal experiences (events and scenarios) inform word associations as well as conceptual representations more broadly (Barsalou, 1983). The presence of GENERIC and TOPIC clusters implies that future relation ontology could benefit from incorporating idiosyncratic relations to better capture more subjective or personal experiences in word associations.

3.4 Relation Classification

Automatic prediction of relation types or generation of explanations can support commonsense knowledge graph completion, enhance our understanding of such knowledge in pre-trained language models, or inform explainability research. In the following sections, we present a series of experiments to demonstrate how WAX can support progress toward some

[16] Our modelling approach masks cues and associations in explanations, potentially diluting property relations, which are often directly reflected in cue-association pairs rather than in the context in explanations.

Relation	Trigger phrase	Relation	Trigger phrase
ANTONYM	opposite	FUNCTION	used
PARTOF	part of	CATEGORYEXEMPLAR	type of, form of
HASPREREQUISITE	require, need to	MATERIALMADEOF	make of/by/with
LOCATION	grow on, grown in, live in, on the, find	SYNONYM	similar, synonym, another word, define

Table 3.10 Templates used to automatically label explanations. The trigger word is the text between cue and association in the explanation.

of these goals. This section addresses relation classification, before we study explanation generation in Section 3.5. We construct a relation classification task using our relation type ontology as the ground truth, as a 16-way classification problem to predict a single relation type r from either only (c,a)-pairs (we call this model -EXP) or the full explanation e, which by construction includes c and a (+EXP).[17] We can thus test whether access to explanations, which lay out *why* two concepts are associated, improves relation prediction over and above the knowledge available to PLMS via large-scale pre-training. For example, predicting a relation (e.g., FUNCTION) for the pair (*bagpipe*, *music*) is arguably simplified (or constrained) with access to an explicit explanation such as "*Bagpipes* are used to play *music*".

3.4.1 Dataset

As the labelled portion of WAX is both small in size and skewed in relation distribution (Figure 3.5 blue), we augment its *training* portion with data from Wu and Barsalou (2009) and ConceptNet (Speer et al., 2017), which include concept pairs and their relation, but no explanations. To create labelled explanations, we find (c,a,r') edges in these external resources that are also in the unlabelled portion of WAX, (c,a,e), and then map the known relation label into our inventory, $r' \rightarrow r$, thus constructing full (c,a,e,r) tuples. In addition, we identified frequent patterns in the WAX explanations, and devised a small set of templates to extract the corresponding relations (e.g., 'a is part of c', indicates a PARTOF relation). All templates are listed in Table 3.10, which includes trigger words and phrases used to

[17] Another natural formulation is multi-class classification given as input a (c,a) pair with *all* produced explanations, which we leave for future work.

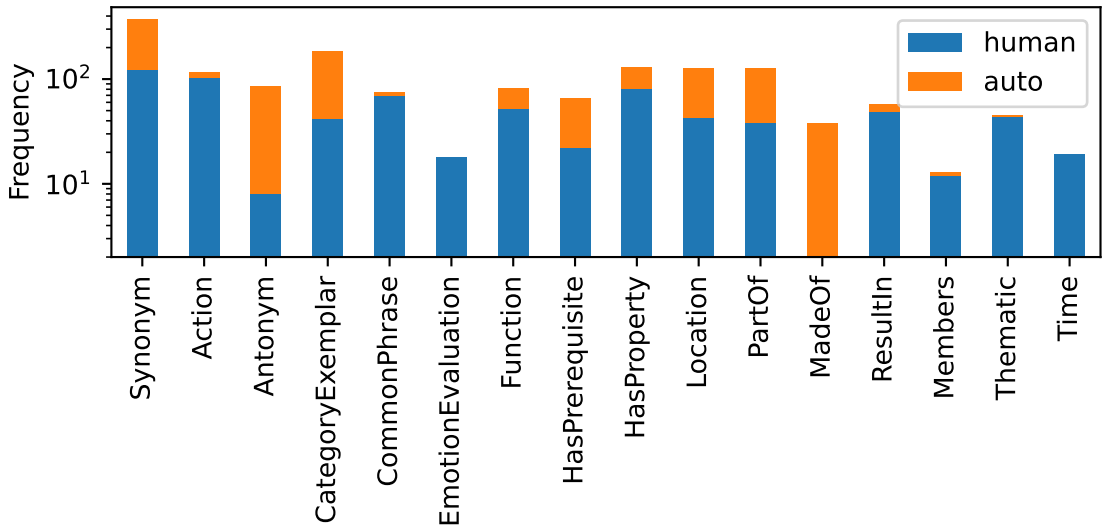


Figure 3.5 Relation distribution of WAX labelled data, including human labelled subset and auto-augmented subset.

automatically map recurring, templated WAX explanations to relations. Those relations were verified independently by two authors of this study, and we retained only instances where both agreed on their validity. An additional 835 labelled explanations were obtained and merged with crowd-sourced labelled data. Figure 3.5 (blue + orange) presents the relation distribution after data augmentation. This data was further split into train, dev, and test sets, resulting in 948, 300, and 312 (c, a, e, r)-tuples, respectively.

3.4.2 Method

We experimented with discriminative and generative seq-to-seq methods for relation prediction. Specifically, we chose BERT (Devlin et al., 2019) and BART (Lewis et al., 2020) as representatives. For details on model architecture and pre-training, readers are referred to Section 2.3.1.4. In the following, we introduce how these models are applied to the relation prediction task.

For the discriminative model, we fine-tuned the BERT-base-cased (Devlin et al., 2019)[18] with our training data and tested its performance. To verify the effectiveness of the ex-

[18]It outperformed other BERT versions, incl. BERT-large.

planations, we fine-tuned the -EXP and +EXP separately with different input formats. For explanation-agnostic models -EXP, we used the simple template "[CLS], c, [SEP], a" as inputs. For explanation-aware models +EXP, we concatenated the full explanation e along with the (c,a) in the format of "[CLS], e, [SEP], c [SEP], a". We used BERT to embed the input and used the hidden representation of the [CLS] token as input to a discriminative classification layer.

In addition, we followed Huguet Cabot and Navigli (2021) and framed relation prediction as a sequence to sequence generation problem by generating (c,a,r) given (c,a,e) for +EXP, or given (c,a) for -EXP, using teacher forcing. Although less direct, the approach is motivated by recent successes in formulating the classical (structured) prediction problem as a seq-to-seq (Bevilacqua et al., 2021, Nayak and Ng, 2020). Including c and a in the output leads to more focused r predictions. We fine-tuned BART-large (Lewis et al., 2020) as the generative model. We represented the encoder input as "e <subj> c POS_c <obj> a POS_a", and the decoder input (with teacher forcing at training time) as "<triplet> c <subj> a <obj> r". <...> are sentinel token, and POS_x the POS tag of argument x.[19]

For baselines, we used the majority class and a logistic regression (LR) classifier with TF-IDF features derived from the training data. All models were trained on the training set, and hyperparameters were selected based on the dev set. The core hyperparameters used were as follows: (a) for BERT, we used the AdamW_hf optimiser, set the maximum steps to 500, used a learning rate of 5×10^{-5}, and a batch size of 8; (b) for BART, the optimiser was AdamW, with a maximum of 1000 steps, a learning rate of 2×10^{-5}, and the same batch size of 8.

3.4.3 Results

Main results Table 3.11 presents the results. The fine-tuned PLMs outperform the baseline models (LR and Majority-Class) by a large margin, and BART performs better than BERT, suggesting the promising direction of modelling word association relations with seq-to-seq frameworks. We further explore this direction in Section 3.5. +EXP models (fine-tuned with

[19] We use the code base from https://github.com/Babelscape/rebel.

	Model	P	**R**	**F1**	**Acc**
	Majority-Class	1.1	6.7	1.9	16.3
-EXP	BERT-base (zero-shot)	1.3 (±1.0)	8.1 (±1.5)	1.6 (±0.7)	5.6 (±3.0)
	LR	5.4 (±0.0)	8.4 (±0.0)	4.5 (±0.0)	18.6 (±0.0)
	BERT-base	24.8 (±1.9)	26.8 (±1.9)	20.7 (±1.2)	32.8 (±1.9)
	BART-large	35.0 (±1.3)	46.5 (±2.7)	36.3 (±1.1)	46.4 (±1.1)
+EXP	BERT-base (zero-shot)	0.4 (±0.4)	5.2 (±1.8)	0.7 (±0.6)	4.9 (±6.4)
	LR	29.9 (±0.0)	17.7 (±0.0)	16.0 (±0.0)	22.1 (±0.0)
	BERT-base	34.2 (±4.3)	40.2 (±5.2)	32.7(±3.7)	45.5 (±2.5)
	BART-large	**47.1** (±1.1)	**53.3** (±0.7)	**45.0** (±0.8)	**53.1** (±1.1)

Table 3.11 Experimental results on relation classification, showing macro-averaged precision, recall and F1, and accuracy for models with access to the full explanation (+EXP) or to cue and association only (-EXP). We report the mean and standard deviation over four runs, using different seeds, on the overall test with 312 instances. All models, except for the Majority-Class and BERT-base (zero-shot), have been fine-tuned (trained) on the training data.

full explanations) performed substantially better than -EXP models (fine-tuned on (c,a) pairs without context), suggesting that explanations provide signals over and above the knowledge already encoded in PLMs. This is confirmed by comparing against a BERT zero-shot model, which yielded results close to a random guess across 16 relations and consistently performed worse than the majority class baseline.

In light of these findings, it is essential to consider human evaluations and their alignment with computational models. Based on our manual validation results presented in Table 3.7, we observed that the consistency in human evaluations for both (c,a,e) and (c,a) criteria lies between 76% and 79%. This led us to estimate that the overall human accuracy in this validation task to be within this range. However, when compared to the performance of models, a noticeable discrepancy emerges. This reveals a substantial gap between model and human performance, indicating ample scope for future enhancement.

Relation diversity We evaluated our models separately on two challenging data subsets to investigate whether models capture the relation diversity discussed in Section 3.3.3: (1) (c,a) pairs with *multiple* explanations that all refer to the *same* relation type (Table 3.12,

	Model	Ambiguous relations (N=131)				Unambiguous relations (N=181)			
		P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
	Majority-Class	0.5	7.1	0.9	6.9	1.9	8.3	3.1	23.2
-EXP	BERT-base (zero-shot)	1.1	4.1	1.4	4.6	1.7	9.1	2.0	6.4
-EXP	LR	2.0	7.7	1.8	7.6	9.6	11.0	7.7	26.5
-EXP	BERT-base	23.9	23.2	18.8	26.2	22.6	25.1	21.0	37.6
-EXP	BART-large	29.7	34.3	28.8	38.7	41.7	**46.1**	41.1	51.9
+EXP	BERT-base (zero-shot)	0.4	3.9	0.7	4.1	0.5	4.9	0.9	5.5
+EXP	LR	23.1	14.5	10.9	16.0	32.3	16.5	16.1	26.5
+EXP	BERT-base	33.2	34.7	29.7	40.7	34.0	35.1	31.7	48.8
+EXP	BART-large	**42.8**	**45.7**	**38.1**	**47.4**	**46.4**	45.4	**41.8**	**57.4**

Table 3.12 Experimental results on relation classification for Ambiguous (left) and Unambiguous (right) relations. The column names are consistent with those used in Table 3.11. Numbers are the mean of averaged four runs, we ignore the std here for readability.

right block); and (2) (c, a) pairs with *multiple* relations that refer to *different* relation types (Table 3.12, left block). Transformer-based models outperform LR, with BART performing best. The difference between BART +EXP vs BART -EXP increases compared to overall results for both F1 and Acc in Table 3.11, confirming the value of explicit explanations for these challenging subsets. Unsurprisingly, the ambiguous relation scenario is the most challenging.

We further analyse how model predictions differ from human labels on both relation-ambiguous and unambiguous (c, a) pairs. We inspect predicted labels from the best-performing model BART. Table 3.13 shows representative examples comparing human and model-predicted relations for unambiguous instances (one true relation, top) and ambiguous ones (multiple true relations, bottom). Although predictions diverge from gold labels, especially for the challenging ambiguous subset, the model labels are often reasonable. Consider (*discuss*, *talk*) with the explanation "to *discuss* something you must *talk* about it" and gold label CATEGORYEXEMPLAR, was predicted by the model as HASPREREQUISITE. It is not uncommon that taxonomic (CATEGORYEXEMPLAR) and associative or situational associations (HASPREREQUISITE, ACTION) relations are both valid for an explanation (Santos et al., 2011), leading to confusions by both our human annotators and model predictions. Our raw relation annotations include at least 5 annotations per (c, a, e) tuple, and hence

(c,a)	Synonym	Prerequisite	Antonym	MadeOf	Location	PartOf	Function	ResultIn	Property	Emo-Eval	Time	Phrase	Action	Thematic	CategoryEx	SameMemb
darkness-light			⊗													
pocket-wallet					⊗											
skunk-smell									⊗	×						
printer-ink		×				○	×									
casino-money					⊗		⊗									
contact-phone	×				○		⊗						×			
lesson-learn		⊗			○			×					⊗			
discuss-talk	⊗	○					×						×		○	

Table 3.13 Selected relation classification results on unambiguous (top) and ambiguous WAX test instances, where each row shows the types of true (○) and predicted (×) relations when applied to the explanations for a cue-association pair.

capture this ambiguity which can be leveraged for model development and evaluation in future work.

Error Analysis To better understand the gap among relations, we conducted a class-wise performance analysis of the best model, BART. Table 3.14 presents the experimental results, revealing that it was accurate for taxonomic relations and well-defined attributes (e.g., {SYNONYM, ANTONYM, PARTOF, LOCATION }), which are well-established in the literature. In contrast, situational associations (e.g., RESULTIN, HASPREREQUISITE) are not captured by the -EXP model, but are predicted at much higher quality by +EXP. To further elucidate these findings, we provide examples of BART predictions from both -EXP and +EXP in Table 3.15.

Our findings concur with the open challenge of event representations in NLP (Sap et al., 2019a) and points to future work on tailoring models and relation sets. One promising direction for enhancing event representation in models is the incorporation of external event-centric knowledge graphs, such as ATOMIC (Sap et al., 2019a) or its expanded iteration, ATOMIC$^{20}_{20}$ (Hwang et al., 2021). Hosseini et al. (2022) has demonstrated that integrating

	Relation	BART -EXP			BART +EXP			
		P	R	F1	P	R	F1	Δ F1
(a)	SYNONYM	100.0	83.3	90.9	77.1	72.6	74.8	↓
	ANTONYM	100.0	100.0	100.0	75.0	100.0	85.7	↓
	ACTION	84.6	61.1	71.0	85.7	55.6	67.4	↓
	PARTOF	55.0	100.0	71.0	100.0	33.3	50.0	↓
	EMOTIONEVALUATION	50.0	100.0	66.7	42.9	60.0	50.0	↓
(b)	LOCATION	76.9	71.4	74.1	69.7	85.2	76.7	↑
	TIME	27.3	100.0	42.9	33.3	100.0	50.0	↑
	FUNCTION	23.5	26.7	25.0	63.6	48.3	54.9	↑
	HASPROPERTY	70.0	38.9	50.0	63.9	82.1	71.9	↑
	COMMONPHRASE	11.1	3.7	5.6	47.6	26.3	33.9	↑
(c)	THEMATIC	0.0	0.0	0.0	17.7	21.4	19.4	↑
	RESULTIN	0.0	0.0	0.0	50.0	33.3	40.0	↑
	HASPREREQUISITE	0.0	0.0	0.0	22.2	60.0	32.4	↑
	MATERIALMADEOF	0.0	0.0	0.0	16.7	100.0	28.6	↑
	CATEGORYEXEMPLAR	0.0	0.0	0.0	27.8	45.5	34.5	↑

Table 3.14 Class-wise relation classification performance of BART when fine-tuned on minimal templates (-EXP) and on full explanations (+EXP). Relations are grouped by change in F1 after adding explanations (Δ F1): (a) well-predicted without explanations, (b) improved with explanations, (c) reliant on explanations for more accurate predictions.

ATOMIC$^{20}_{20}$ into BERT notably improves performance on capturing cause-effect relationships. Similarly, incorporating these resources into models such as BART may improve the event related relations in word associations such as HASPREREQUISITE and RESULTIN.

3.5 Generating Relation Explanations

Natural language inference or commonsense reasoning is often framed as mapping a free text input (e.g., a paragraph) to a structured output (e.g., a relation, (c, a, r) triple, or a multiple-choice answer). The underlying reasoning steps behind models typically remain obscure. Constructing intuitive and faithful explanations for model predictions is an active research area of increasing impact (Ribeiro et al., 2016). Mapping structured representations to natural language explanations is one approach (Gardent et al., 2017, Bansal et al., 2022), which has been limited by a lack of suitable training data sets (Ke et al., 2021). WAX is

(c,a)	Explanation	Pred -EXP	Pred +EXP
	(a) **both correct**	✓	✓
clean-neat	the space is clean and neat enough.	SYNONYM	SYNONYM
military-army	the us army is one branch of our military.	PARTOF	PARTOF
pocket-wallet	i put my wallet in my pocket.	LOCATION	LOCATION
darkness-light	darkness is the opposite of light.	ANTONYM	ANTONYM
	(b) **only +EXP correct**	×	✓
casino-money	casinos are full of tons of money locked up.	FUNCTION	LOCATION
contact-phones	we use our phones to contact people.	THEMATIC	FUNCTION
disaster-hurricane	a hurricane can cause a large disaster.	CATEXEMP	RESULTIN
catcher-ball	the catcher threw the ball back to the pitcher.	FUNCTION	ACTION
payment-money	in order to make a payment you need money.	CATEXEMP	HASPREREQ
	(c) **both wrong**	×	×
chomp-eat	when you chomp at something, you eat it.	ACTION	SYNONYM
weeping-willow	a weeping willow is a type of tree.	ACTION	CATEXEMP
summer-hot	in summer it gets hot.	PROPERTY	PHRASE
salad-dressing	the secret to a great salad is the dressing.	EMOTION	EMOTION

Table 3.15 Case study on relation prediction by BART (-EXP and +EXP): (a) Both models correct (top); (b) Only +EXP correct with explanations (middle); (c) Both models incorrect (bottom).

a parallel data set of structured relational data, aligned with diverse, human-generated free text explanations. Here, we show that it can support models to generate explanations which capture the diversity of human reasoning. We fine-tune a generative LM to generate e given (c,a,r). This task is designed to test how well a model understands high-level relations and its ability to generate explicit explanations that align with those relations.

3.5.1 Dataset

We used the same train, dev and test split as Section 3.4.1. In addition, we augmented the training set to increase its size and balance with respect to the relation labels: for each (c,a,e,r) instance in the training data, we masked either c or a in the explanation and filled the blank with the top 10 candidates generated by BERT-large.[20] We down-sampled generated instances of overrepresented relation types, resulting in a balanced dataset of 12K

[20]We inspected a sample of 80 prompts for validity.

(c, a, e, r) tuples, which were used to fine-tune BART. The original validation data was used for model selection.

In testing, we modify the test set used in the relation classification task for this evaluation. The adaptation involved evaluating the fine-tuned model on its explanation generation using (c, a, r)-triples, considering relations r as either 'seen' (present in WAX) or 'unseen' (absent from WAX). We investigated two conditions: (a) prompting with human-created triples from WAX, e.g., (*dog, bark,* ACTION); and (b) replacing r in seen triples with unseen relations, e.g., (*dog, bark,* ANTONYM). This setup allowed us to examine and analyse how the model differentiates and responds to relations that they have previously encountered ('seen') in contrast to those they have not ('unseen'). Notably, although these 'unseen' relations do not exist in our test instances, they could be either plausible or implausible.

3.5.2 Method

Generating Explanations The task is to generate a free-text explanation from a given (c, a, r)-triple encoded into the sentence prompt "c and a have a r relation". The output is a short sentence supporting the prompt. For example, the input "*bucket* and *wash* have a *function* relation" could elicit the output "I use a bucket to wash my car". We chose BART as the base model for fine-tuning. BART is suitable for this task because its training is aligned with the objective of generating explanations: it is trained to complete missing context in an input (cf., Section 2.3.1.4). In this scenario, the (c, a, r)-triple is viewed as an incomplete or noisy input, and the explanation generated by BART serves to recover or fill in the missing contextual information pertinent to the relation. We used BART-large to fine-tune on 12K augmented training instances and then used it to generate explanations for the test set.

The hyper-parameters for the BART model, selected based on validation set, are: optimiser set to AdamW, a maximum training duration of 2000 steps, a learning rate of 2E-05, and a batch size of 4.

Evaluating Explanations Evaluating generated explanations is challenging due to their diversity and free-form nature, making comparisons to a reference difficult as multiple correct

variants may exist. Traditional evaluation metrics, such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), are unable to effectively capture high-level relation similarity, while recent semantic metrics like BERTScore (Zhang et al., 2020b) often overestimate the quality, making them suboptimal for evaluating the quality of generated relational explanations. In addition, our test set contains unseen relations where the human generated explanations are not available.

To assess the alignment between generated explanations and the input (c, a, r)-triple, we use perplexity from GPT-2 (Radford et al., 2019)[21] as a metric. This is inspired by the *forward perplexity* metric used in adversarial autoencoders by Zhao et al. (2018), which evaluates the fluency of model-generated data with another language model trained on real data. This metric has also been applied to the task of graph-to-text conversion, as demonstrated by Zhu et al. (2019), where a sequence-to-sequence neural model is trained to transform relational triples into coherent natural language sentences, which are evaluated using a separate language model.

In our evaluation, we: concatenate the relational prompt with the generated explanation; and input this concatenated text into GPT-2-XL to measure the perplexity of the entire sequence. This approach is also adopted because we observe that BART is not sufficiently relation-aware, often generating similar explanations regardless of the given relation. By concatenating the relational prompt with the explanation, we enable a more effective evaluation of their alignment. The underlying hypothesis is that a closer alignment between the explanation and the prompt relation should yield a more fluently and naturally resonating text. For example, "Cowgirls and boots have a 'part of' relation because cowgirls wear boots as part of their outfits" is perceived as more natural compared to "Cowgirls and boots have an 'emotion evaluation' relation because cowgirls laced up their boots under the stars."

While perplexity serves as our chosen metric for evaluation, we must acknowledge its limitations. It's a measure of how smoothly text reads rather than its factual correctness. Additionally, we are mindful that the biases in GPT-2 could influence the evaluation, poten-

[21] Refer to Section 2.3.1.2 for details on model architecture and training data.

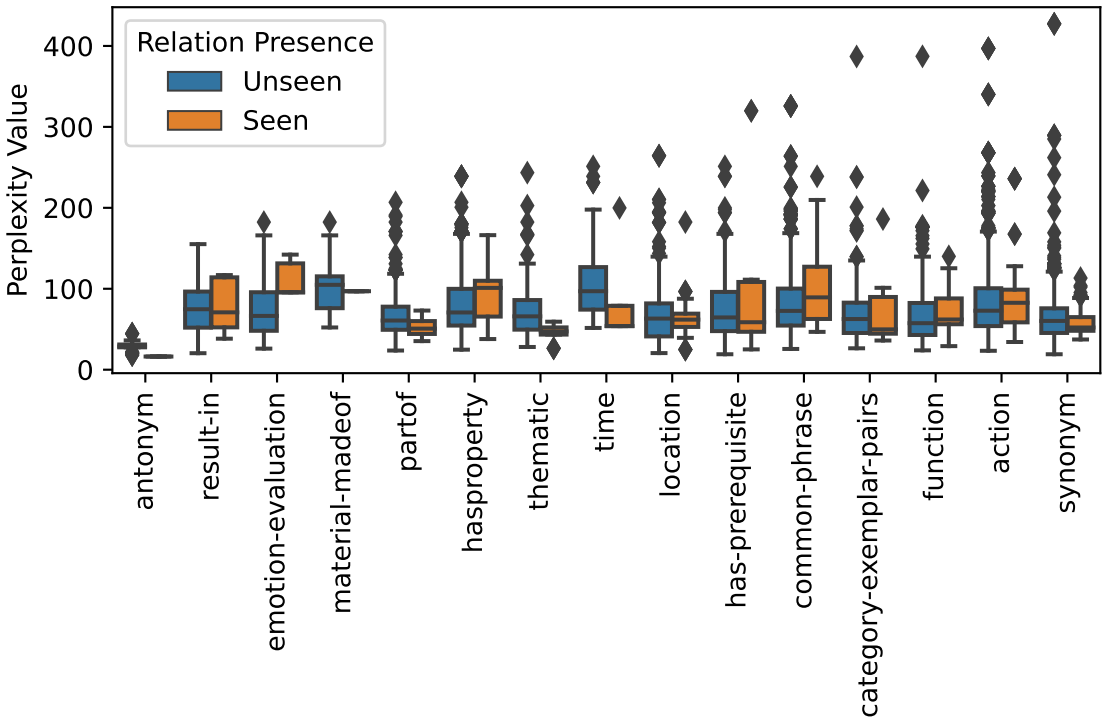


Figure 3.6 Model perplexity for seen and unseen relations, broken down into relations present in our labelled test set ('Seen') or not ('Unseen').

tially favoring explanations that align with its training data over those that are truly accurate. Therefore, we also conduct qualitative evaluations for generated explanations.

3.5.3 Results

Quantitative results are depicted in Figure 3.6, which presents the distribution of perplexity across all relations, subdivided into seen and unseen prompt relations. Generally, the seen group exhibits lower perplexity values, while the unseen group displays a wider, more outlier-prone distribution (denoted by ♦ in Figure 3.6). This aligns with our expectation as unseen relations might also be plausible, and therefore not all explanations for them are invalid, even though they are more likely to present with higher perplexity values. This pattern suggests that perplexity can, to a degree, distinguish between the two groups; however, this distinction fluctuates across relations. While it effectively differentiates certain relations (e.g., ANTONYM, PARTOF, THEMATIC, HASPREREQUISITE), it fails to do so for others (e.g.,

	Prompt Relation	Generated Explanation
(a)	LOCATION	a judge is a person who sits in a court.
	THEMATIC	the secret to a great salad is the dressing.
	SYNONYM	liberty has a similar meaning as freedom.
	ANTONYM	the opposite of darkness is light.
(b)	ACTION	i last week drove my car.
	PHRASE	my father is very fatherly figure.
	CATEXEMP	a chew is a type of eats.
	LOCATION	i learned the lesson in school.
	EMOTION	if someone is weeping they are likely sad.
(c)	ANTONYM	liberty is the opposite of freedom.
	ANTONYM	lonely is the opposite of alone.
	THEMATIC	there are a lot of drugs in the pharmacy.
	LOCATION	there is a lot of light in the darkness.
	PROPERTY	when you are writing a book, you are using a pen to type the words.
	RESULTIN	when you take your wallet out of your pocket, you put it in your back pocket.
(d)	EMOTION	the igloo was very cold.
	EMOTION	i use a tire to spare my spare parts.
	EMOTION	i last week drove my car.
	RESULTIN	phones make contact.
	FUNCTION	a chomp has teeth that eats.
	FUNCTION	i like to start my car first

Table 3.16 Illustrative examples of BART generated explanations in response to relation prompts of the form "*c* and *a* have a *r* relation." For each example, *r* is shown on the left and *c* and *a* are underlined in the generated explanation. Outputs are grouped to illustrate the following: (a,b) prompt relations seen in WAX labelled set with low and high perplexity, respectively; (c,d) prompt relations not seen in the WAX, also with low and high perplexity, respectively.

SYNONYM, RESULTIN, HASPROPERTY and COMMONPHRASE). Moreover, the overall perplexity distribution is smaller for some relations (e.g., ANTONYM, RESULTIN), yet dispersed for others (e.g., ACTION, SYNONYM), suggesting the model's inconsistent understanding of different relations. Future investigations focused on models that exhibit enhanced knowledge and language comprehension, such as KG-enriched models like COMET (Bosselut et al., 2019) or advanced language models like GPT-4 (OpenAI, 2023), will likely provide more comprehensive understanding into the generation and evaluation of explanation quality.

To better understand our results, we analyse both seen and unseen groups qualitatively, with the goal of comprehending the role of perplexity in evaluating explanation quality. Our inspection spans four scenarios: (a) seen relations with low perplexities; (b) seen relations with high perplexities; (c) unseen relations with low perplexities; and (d) unseen relations with high perplexities. Refer to Table 3.16 for representative examples, focusing on the variability in response quality across different perplexity levels.

We observed the following: explanations in (a) and (b) are generally relevant, factual, and of high quality, indicating that models can generate accurate explanations when prompted with factually correct relational triples; (c) explanations can connect concepts with a given relation, but they may introduce factual errors, as seen in examples 1 and 2; (d) high-perplexity explanations do not align with prompted relations, accurately indicating misalignment. While perplexity can serve as a metric for distinguishing alignment, it fails to detect factual errors, as seen in (c). This observation aligns with recent studies, which demonstrate that large language models suffer from hallucinations (Shuster et al., 2021, Dziri et al., 2022).

Our analyses suggest that the triple-explanation paralleled instances in WAX can be used for probing knowledge alignment between structure and free-text knowledge in PLMs, aiding further research on knowledge consistency in pre-trained language models. We leave the exploration of developing more advanced approaches to generate explanations and other evaluation metrics for future work. One potential direction could be jointly training KG-to-Text and Text-to-KG models to enhance their mutual alignment. Mousavi et al. (2023) pursued this idea in Wikipedia domain and introduced a new evaluation metric named cyclic evaluation to evaluate the generated text. This metric converts explanations back to triples and measures precision and recall against the original input triples. Such a method offers a qualitative measure for assessing knowledge alignment; however, its validity in broader domains requires further investigation.

3.6 Limitations and Discussion

In this research, we created and analysed WAX, the first large-scale word association explanation dataset. We explored associative rationales and computational approaches to automate classification and explanation generation. Amidst the rapid advancement of NLP techniques, we acknowledge the necessity to introspect and critically evaluate the constraints and challenges that surfaced during our journey. We now elucidate the challenges encountered, and shed light on the limitations inherent in our study, spanning ontology design, efficiency, linguistic expansion and discusses alternative modelling strategies.

Relation ontology Our relation ontology (cf., Table 3.4), which encompasses four broad categories and sixteen fine-grained relations, was designed with some strategic considerations in mind. These include the coverage and frequency of relations, annotation difficulty, and expressiveness in explanations. One key decision was to merge the directionality for all relations during our annotations, which was primarily driven by two reasons. First, the reverse directions are not always reflected in the explanations. Second, we reduced the number of relationships to ease the annotation process and ensure annotation agreement.

However, this approach did eliminate directional information. This is not an issue if the relation type itself is symmetric (e.g., SYNONYM, ANTONYM), however, a further step is necessary for asymmetric relations (e.g., HASPREREQUISITE and RESULTIN) as they inherently carry a sense of directionality. For such asymmetric relations, the absence of directional information may require additional context or inference to accurately interpret the relationship. In future work or applications, special attention should be given to these relations to ensure their correct interpretation and usage.

Additionally, throughout our experiments, we opted to discard instances for which there was no label agreement among at least two annotators (Table 3.5 bottom). These instances, however, possess potential utility for future research (Plank, 2022).[22] By introducing a

[22]For WAX and related resources, refer to the repository of (Plank, 2022) at https://github.com/mainlp/awesome-human-label-variation.

'none-of-the-above' category, subsequent studies might be able to encapsulate these instances for which a consensus on labelling could not be reached.

Scalability and Linguistic Expansion of the Dataset We acknowledge that our dataset is collected from a limited number of English native speakers, and it can serve as an initial work to understand the underlying associative reasons *within this group*. Caution should be exercised when drawing general conclusions about human conceptual knowledge, and an important direction for future work is an extension to other languages. Reasons for associations are likely more diverse than reflected in our data set. The efficiency of WAX's collection process (200 hours of crowd-sourcing) suggests that it can be scaled up or extended to other languages. Nevertheless, considering the complexity of this task, particularly in annotating the relation labels, providing adequate training is important to ensure that annotators fully comprehend the task.

The Relationship between WAX and SWOW WAX is essentially a subset of SWOW, with cues in WAX directly sampled from SWOW — the current largest word association network introduced in Section 2.2.2. However, what sets WAX apart is the enrichment with explanations and relation labels for its edges. Thus, establishing WAX serves not only as a foundational step and proof of concept, demonstrating that word associations encode commonsense knowledge but also paves the way for our subsequent exploration of larger-scale word associations in Chapters 4 and 5.

While WAX has been crucial in unearthing the relational knowledge within word associations, its relatively smaller scale, compared to SWOW, limits its usage as a commonsense knowledge resource. Specifically, WAX has 12 times fewer cues and 10 times fewer participants for each cue than SWOW. Consequently, despite the valuable insights provided by WAX, SWOW, with its more extensive knowledge base, emerges as a more potent resource for investigating the practical utility of commonsense knowledge within word associations for various downstream tasks. We have demonstrated that model performance on relation classification for word associations can be enhanced by providing explanations. Consequently, our focus will shift to SWOW in Chapters 4 and 5.

3.7 Summary

Word associations have long been used as a lens into human conceptual representations, however, the *types* and *reasons* underlying these associations have not been studied at scale. In this chapter, we tackled the first research question of why do people associate certain words via uncovering the contextual signals that participants rely on when forming these associations through self-explanation. We presented WAX, a large data set of English word associations with explanations and relation labels. WAX is both an opportunity to better understand the human mental lexicon, and a repository of relational commonsense knowledge, both structured as (c, a, r) tuples, and free-text through the associated explanations. We demonstrated the utility of WAX for supervised relation classification and explanation generation; and presented a detailed data set analysis including association diversity and data-driven relation types.

Gaining a deeper understanding of the rationales and relationships encoded in word associations through WAX, we are led to explore their practical significance. In the upcoming chapter, we focus on the understanding commonsense knowledge encapsulated in large-scale word associations and examine their benefits on NLP downstream tasks.

Chapter 4

Commonsense Knowledge in ConceptNet and Word Associations

Human word associations offer insights into human representations of semantic knowledge (A. Rodriguez and Merlo, 2020) and serve as a proxy to evaluate knowledge in computational models. They have been used to evaluate word embeddings (Thawani et al., 2019) and probe social biases in pre-trained language models (Kaneko and Bollegala, 2021, Bommasani et al., 2020).

In Chapter 3, we illuminated latent links within word associations, revealing how they can be explicit, explainable, and encode several high-level relation types. Notably, these types align closely with the relation ontology of `ConceptNet`, a widely-recognized commonsense knowledge graph, as introduced in Section 2.2.1. These findings provide evidence that word associations pairs, linked by semantic relations, encode commonsense knowledge, albeit in an implicit manner. Furthermore, word associations can be collected cheaply at large scale from diverse participants (Nelson et al., 2004, De Deyne et al., 2019, Cabana et al., 2023). This led us to ask: Can word associations provide a new approach for acquiring commonsense knowledge? As discussed in Chapter 1, commonsense knowledge is both implicit and diverse, spanning an expansive range of human experiences. It touches upon various aspects of our daily activities, from spatial and physical to social, temporal, and psychological dimensions (Liu and Singh, 2004). Word associations, by their spontaneous

and open-ended nature, emerge from our memories and lived experiences, reflecting a wide range of connections that people use or experience in their lives.

It is unclear how the knowledge in large-scale word associations like SWOW (De Deyne et al., 2019) differs from existing commonsense knowledge graphs, as well as its utility in downstream tasks. Understanding how commonsense knowledge is represented across different paradigms is crucial for both a deeper comprehension of human cognition and for augmenting automatic reasoning systems.

In light of this, our research question arises: *Can large-scale word associations improve performance on downstream commonsense reasoning tasks? How does the knowledge they encode differ from the existing largest commonsense knowledge graph* ConceptNet*?* This chapter delves into an in-depth comparison of two large-scale resources of general knowledge: ConceptNet, an engineered relational database, and SWOW, a knowledge graph derived from crowd-sourced word associations.[1] We examine the structure, overlap and differences between the two graphs, as well as the extent to which they encode situational commonsense knowledge. We finally show empirically that both resources improve downstream task performance on commonsense reasoning benchmarks over text-only baselines, suggesting that large-scale word association data, which has been obtained for several languages through crowd-sourcing, can be a valuable alternative to curated knowledge graphs.[2]

This chapter builds on the paper:

> Chunhua Liu, Trevor Cohn and Lea Frermann. 2021. Commonsense Knowledge in Word Associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CONLL 2021)*, pages 481–495, Online only.

[1] See the construction processes and properties of ConceptNet and SWOW in Chapter 2.2.
[2] Code available at https://github.com/ChunhuaLiu596/CSWordAssociation

4.1 Introduction

Recently, language models (Devlin et al., 2019, Radford et al., 2019) pre-trained on massive text corpora achieved promising results on commonsense reasoning benchmarks with fine-tuning, however, the lack of explicit reasoning capabilities and an understanding of underlying knowledge structures remains a problem. Augmenting them with external commonsense knowledge can provide complementary knowledge (Ilievski et al., 2021, Safavi and Koutra, 2021) and thus make models more robust (Lin et al., 2019a, Zhang et al., 2021b).

Such external commonsense knowledge resources are typically built through automated extraction from large text corpora (Tandon et al., 2014, 2017, Navigli and Ponzetto, 2012, Zhang et al., 2021a) or via manual curation with crowdsourcing (Liu and Singh, 2004, Sap et al., 2019a). While valuable, their main challenge lies in the limited coverage, as they often fail to encompass the vast array of human commonsense knowledge. This issue of sparsity, even in the largest commonsense knowledge graph like `ConceptNet` (Speer et al., 2017), was previously pointed out in Section 2.2.1. In contrast, word associations, derived from human spontaneous associations, emerge as a new resource capturing basic human knowledge. Studies of word associations have been scaled to thousands of cue words, tens of thousands of participants, and several languages, thereby providing a way of collecting diverse and comprehensive representations (see Section 2.2.2 for details).

However, the nature of commonsense knowledge it encodes and its divergence or alignment with established commonsense knowledge graphs remains to be explored. Does it merely mirror the knowledge present in curated commonsense graphs, or does it encapsulate unique insights reflective of direct human cognition? Moreover, does it possess any tangible utility when applied to commonsense reasoning tasks?

To answer these questions, we systematically compare the most comprehensive, domain general, curated commonsense knowledge base (`ConceptNet`) with the largest data set of English word associations (the "Small World of Words"; `SWOW`; De Deyne et al. (2019)). We evaluated their structures, contents, and performance in commonsense reasoning tasks. Our analysis spanned three experiments: examining graph properties such as density and node degree; aligning both knowledge bases with commonsense scripts detailing everyday

activities to assess their representational differences; and incorporating both graphs to commonsense question-answering benchmarks. Findings show that SWOW, despite its smaller size and limited knowledge overlap with ConceptNet, forms denser networks with common concepts. Importantly, both graphs enhance commonsense QA models, demonstrating comparable performance improvements across three datasets.

Our study in this chapter is important for three reasons. First, comparing explicitly engineered with spontaneously produced knowledge graphs can advance our fundamental understanding of the diversity and potential gaps between the paradigms, and suggest ways to combine them. Second, recent progress in automatic commonsense reasoning largely focused on English and relies heavily on the availability of large language models. These are, however, infeasible to train for all but a few high-resource languages. Word associations in SWOW, spanning across 18 languages,[3] offer new potential for examining low-resource languages. Finally, work has shown that the competitive performance of large language models on commonsense reasoning tasks is at least partially due to spurious correlations in language rather than genuine reasoning abilities; and that they perform close to random once evaluation controls for such confounds (Elazar et al., 2021b). Explicit representations of commonsense knowledge bases thus have the potential to promote robust and inclusive natural language reasoning models.

In summary, our contributions to this chapter are:

1. We are the first to conduct in-depth comparisons of the large-scale human psychological behavior database SWOW with the engineered database ConceptNet, distilling a number of systematic differences that can inform future theoretical and empirical work.

2. We analyse how much *commonsense* knowledge ConceptNet and SWOW encode, leveraging a human-created data set covering explicit situational knowledge. Our results suggest that SWOW represents this knowledge more directly and intuitively.

[3] https://smallworldofwords.org/en/project/stats

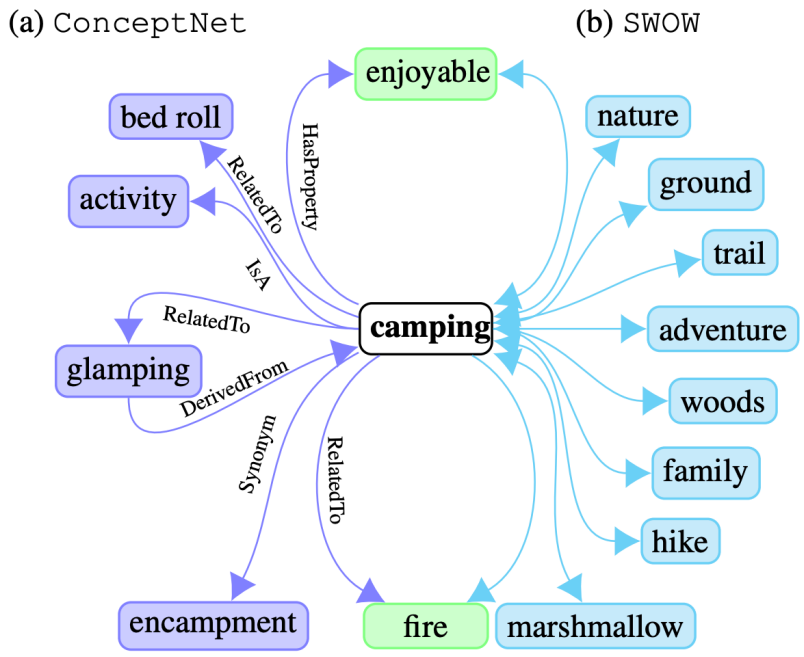


Figure 4.1 Sub-graphs centered around '**camping**' from (a) ConceptNet and (b) SWOW. Nodes in green are common to both KGs. Nodes on the left/blue (right/cyan) are unique to ConceptNet (SWOW). For further details on concepts linked with "camping" in ConceptNet and SWOW, see Figure 1.1 in Chapter 1.

3. We introduce SWOW as a commonsense resource for NLP applications and show that it achieves comparable results with ConceptNet across three commonsense question answering benchmarks.

In this chapter, we address the second research question posed in our thesis, revealing the dense and intuitive nature of commonsense knowledge encoded in the large-scale word associations SWOW, and highlighting its potential as a novel commonsense knowledge resource.

4.2 Background

We provide a brief background, highlighting aspects relevant to this chapter and the data pre-processing undertaken. For a comprehensive overview of commonsense knowledge graphs and word association networks, please refer to Section 2.2.

ConceptNet ConceptNet (Speer et al., 2017) is the largest domain-general commonsense knowledge graph.[4] In this chapter, we use the most recent version ConceptNet v5.6 (Speer et al., 2017), which is a directed graph comprising over 3M nodes (aka concepts). Related concepts are connected with directed edges, which are labelled with one of 47 generic relation types. Figure 4.1a shows a small subgraph, centered around the concept *camping*. Nodes are represented as free-text descriptions, which leads to a large node inventory and a sparsely connected graph. We filter out nodes that are not in English, lowercase all descriptions, and remove punctuation. Row 1 in Table 4.1 shows statistics of the resulting knowledge graph.

SWOW Word associations, which are spontaneous associations made by humans, encapsulate implicit semantic knowledge in our minds (see Sections 2.1.2 and 2.2.2). However, in order to consider these associations as a general knowledge resource, it should (a) cover a large set of diverse cues, and (b) include a large number of responses which are both diverse and reliable. Recently, the *Small World of Words* (De Deyne et al., 2019) project has significantly scaled up its word association collection, involving more than 90K participants and 12K cues.[5]

In our use of SWOW, we adopt the official, pre-processed release.[6] We removed "NA" responses, lowercased all node descriptions, and removed punctuation. Notably, SWOW edges are not labelled with explicit relation types. In our experiments, we introduce a basic set of two relations using the associative directions, namely *forward associations* from a cue to a response (e.g., *camping* $\rightarrow$ *fire*), and *mutual associations* for pairs where the reverse is also included in SWOW (e.g., *camping* $\leftrightarrow$ *hiking*). In this way, the resulting cue-association pairs can be compiled into a knowledge graph, as illustrated by a small excerpt in Figure 4.1b, and the corresponding graph statistics are presented in Row 2 of Table 4.1.

In the remainder of this chapter, we conduct three experiments to compare SWOW and ConceptNet from different dimensions, ranging from the intrinsic graph properties (Sec-

[4]For detailed information about the construction of ConceptNet, refer to Section 2.2.1
[5]See Section 2.2.2 for details on SWOW construction process and properties.
[6]https://smallworldofwords.org/en/project/research

KG	#Triples	#Nodes	#Relations	Density	Degree	H_N
ConceptNet	3,009,636	1,080,759	47	3.00×10^{-6}	2.78	23.28
SWOW	1,593,564	124,626	2	1.03×10^{-4}	12.78	18.07

Table 4.1 Statistics of ConceptNet and SWOW considered as directed graphs. Density is the graph density, Degree indicates the average node degree. H_N indicates the node entropy.

tion 4.3), to the coverage of encoded commonsense knowledge (Section 4.4), and their utility for downstream commonsense reasoning tasks (Section 4.5).

4.3 Intrinsic Comparisons

A relational knowledge graph $\mathcal{G}$ consists of a set of concept nodes $\mathcal{C}$ and edges $\mathcal{E}$, comprising triples (c_1, r, c_2) to denote a directed edge from head node c_1 to tail node c_2 labelled with relation $r \in \mathbb{R}$. We denote $\mathcal{E}(c)$ as the incoming and outgoing edge set for node c, $\mathcal{E}(r)$ as the set of edges with relation r, and $|\cdot|$ as the size of a set. Here, we consider the specific knowledge graphs ConceptNet and SWOW, and begin by comparing the intrinsic properties: their typology and content encoded.

4.3.1 Knowledge Graph Structure

ConceptNet is a substantially larger graph than SWOW, with about eight times as many nodes and 1.9× as many edges (cf., Table 4.1). We compare sparsity in terms of (1) *graph density*,

$$\frac{|\mathcal{E}|}{|\mathcal{C}|(|\mathcal{C}|-1)},$$

and (2) *node degree* as the average total of incoming and outgoing edges (Malaviya et al., 2020). Table 4.1 shows that SWOW has 39× the density and 4× the average node degree of ConceptNet: Even though SWOW is smaller than ConceptNet, it is substantially more densely connected.

Node Distribution. To better understand the distribution of nodes in the KGs, we measure node diversity via the entropy of the node distribution (H_N; Pujara et al. (2017)):

$$H_N = \sum_{e \in N} -P(c) \log P(c),$$

where $P(c) = |\mathcal{E}(c)|/|\mathcal{E}|$ is the fraction of edges incident on the node c.

Higher H_N indicates a more uniform node distribution where many nodes are connected, whereas lower entropy suggests a skewed distribution, where few nodes are highly connected. Table 4.1 shows a lower H_N for SWOW, i.e., nodes are less uniformly connected. This is because SWOW is by construction more structured than ConceptNet: its 12K *cue* nodes are densely connected to a much larger number of (sparsely connected) *response* nodes.

Node and Edge Overlap. How large is the overlap between ConceptNet and SWOW? We quantify the overlap of individual nodes in the two KGs, based on exact string match.[7] We find 58% (71K) of the nodes in the smaller SWOW are present in ConceptNet (conversely, 7% of the nodes in ConceptNet are present in SWOW). Over 40% of the concepts in SWOW are not present in ConceptNet, which is perhaps expected given their very distinct methods of construction, but motivates further in-depth comparison (Section 4.3.2). Moving on to edge overlap, which we measure over undirected head-tail pairs,[8] we find that 6% of edges in ConceptNet are present in SWOW, and 15% of the edges in SWOW are present in ConceptNet. This low overlap demonstrates that human associations indeed elicit connections among words missed in the large database ConceptNet. For the 71K overlapping nodes, we further find that 691K connections exist in ConceptNet, and 1.5M connections in SWOW, covering 95% of all SWOW edges. This again suggests that SWOW is more comprehensive than ConceptNet.

[7] String matching is arguably a simplistic way of matching concepts across two KGs. Advanced methods of concept resolution could leverage embedding methods. We leave this interesting direction for future work.

[8] Edge comparison ignores direction as many relations can naturally be inverted, e.g., PART OF and HAS PART. Consequently linking concepts in either direction is considered to be correct.

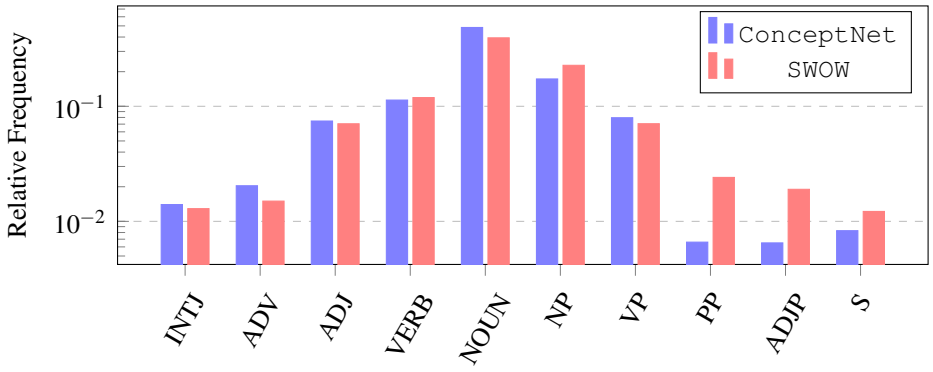


Figure 4.2 The distribution of syntactic tags on ConceptNet and SWOW for the 10 most frequent tags.

4.3.2 Knowledge Graph Content

Having established the structural characteristics of ConceptNet and SWOW, we will now focus on their respective encoded knowledge.

4.3.2.1 Conceptual Content

Nodes in ConceptNet and SWOW express concepts as words or short phrases. We compare: (1) the distributions of the syntactic categories for concepts over the two knowledge bases; (2) the occurrences of concepts in two KGs in large text corpora.

We use a constituency parser to predict the syntactic phrase or part of speech (POS) tag for a concept string.[9] The relative prevalence of the 10 most frequent syntactic types is shown in Figure 4.2. While the overall distribution is similar in both KGs, two patterns emerge. First, even though both KGs are dominated by nominal nodes, SWOW's distribution over POS types is less skewed, suggesting concepts are more diverse. Secondly, the proportion of phrasal concepts compared to single-word concepts tends to be higher in SWOW compared to ConceptNet.

Next we examine the corpus frequency of concepts in ConceptNet and SWOW using the Google N-gram corpus.[10] Many ConceptNet concepts were not present in this large

[9] We use the parser of Kitaev and Klein (2018) as implemented in Spacy.
[10] https://books.google.com/ngrams

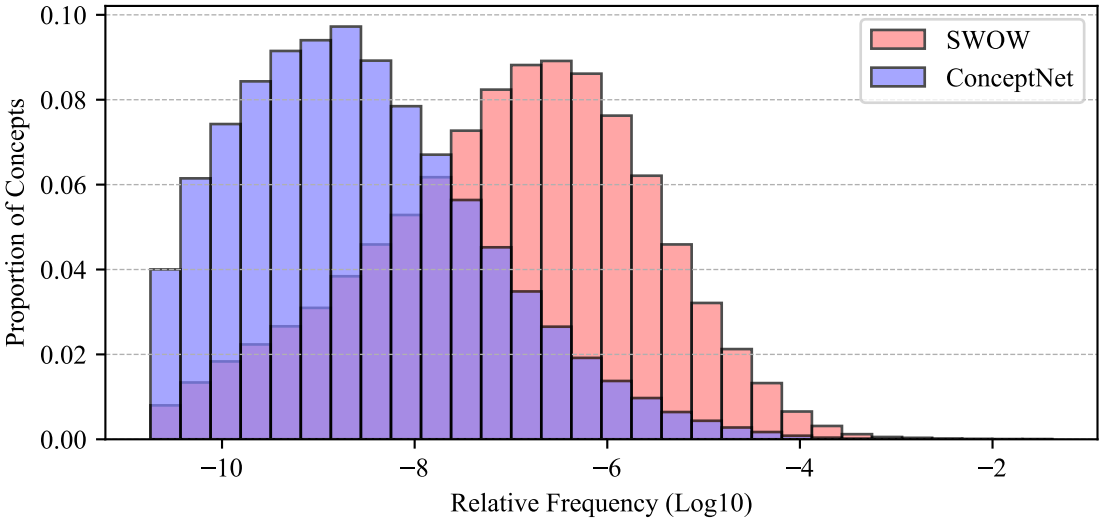


Figure 4.3 Distribution of relative corpus frequencies for concepts in ConceptNet and SWOW, derived from the Google N-gram corpus.

corpus (27%), versus 15% for SWOW. Of those concepts that could be found, concepts in SWOW are on average 7× more common than those in ConceptNet. As illustrated in Figure 4.3, in the high-frequency areas (to the right of the figure), the proportion of concepts is dominated by SWOW, not ConceptNet. Therefore, we conclude that SWOW concepts are generally common, while ConceptNet includes more obscure concepts due to the phrasal nature of its nodes.

4.3.2.2 Relational Content

Relations are crucial for shaping and understanding concept meanings as we introduced in Section 2.1.1. Motivated by this, we are interested in comparing the relational knowledge in ConceptNet and SWOW. Specifically, we aim to explore in how relations from human spontaneous associations in SWOW vary from the structured, engineered knowledge in ConceptNet. Our prior work in Chapter 3 studied relations within word association explanations, employing both a top-down ontology for labelling (Section 3.3.2) and a data-driven clustering approach (Section 3.3.3.1).

In this section, we provide another perspective on the relations in the larger SWOW network, examining how its relation distribution aligns with WAX, whose subset is manually labelled

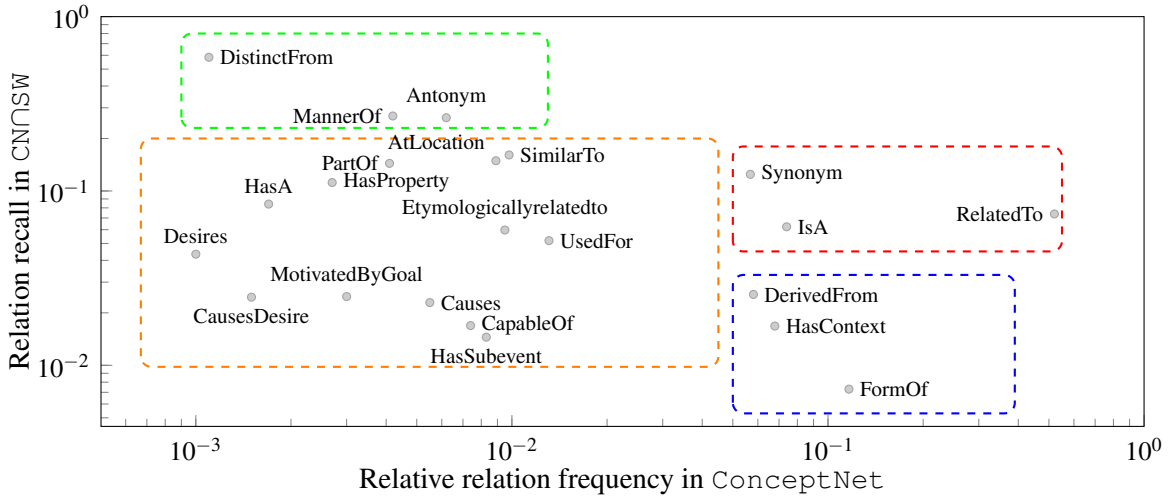


Figure 4.4 The correlation between the relative frequency of ConceptNet relations and their recall in the overlap subgraph, CN∩SW.

with relation types, offering an estimation of relational distribution in word associations. Direct comparison of relation distributions between ConceptNet and SWOW is challenging due to SWOW's lack of labelled relations. To address this, we turn to the intersecting graphs (CN∩SW), which encompass shared head-tail pairs between the two graphs ($|\mathcal{E}_{\text{CN}\cap\text{SW}}|$=190K), labelled by their ConceptNet relations. We leverage CN∩SW as a proxy of SWOW to assess: (a) the relation distribution contrast with ConceptNet and (b) how this distribution aligns with WAX.

Firstly, for each relation type r, we compare its relative frequency in the full ConceptNet (f^r_{CN}) against its recall in CN∩SW (recall^r), where

$$f^r_{\text{CN}} = \frac{|\mathcal{E}_{\text{CN}}(r)|}{|\mathcal{E}_{\text{CN}}|}, \quad \text{recall}^r = \frac{|\mathcal{E}_{\text{CN}\cap\text{SW}}(r)|}{|\mathcal{E}_{\text{CN}}(r)|}.$$

Secondly, we compare the relative relation frequency in CN∩SW to that in the 1.5K labelled WAX data using a similar approach:

$$f^r_{\text{CN}\cap\text{SW}} = \frac{|\mathcal{E}_{\text{CN}\cap\text{SW}}(r)|}{|\mathcal{E}_{\text{CN}\cap\text{SW}}|}, \quad f^r_{\text{WAX}} = \frac{|\mathcal{E}_{\text{WAX}}(r)|}{|\mathcal{E}_{\text{WAX}}|},$$

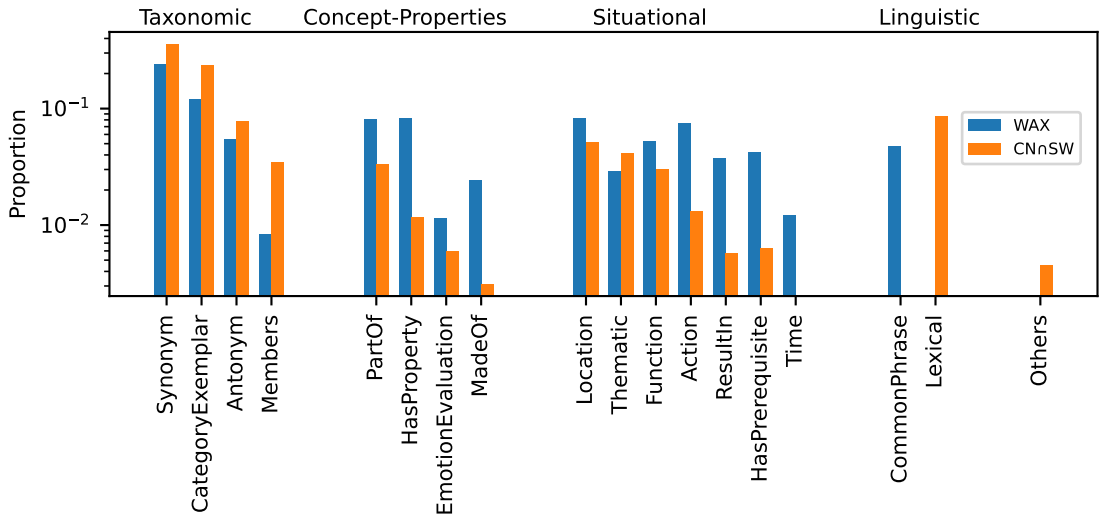


Figure 4.5 Comparison of relation distribution between CN∩SW and the labelled data in WAX.

We align the ConceptNet relation ontology into WAX relation ontology,[11] introducing a new LEXICAL relation to cover morphological relations such as FORMOF. An additional OTHERS relation captures relations in ConceptNet that do not have a direct counterpart in WAX. Importantly, the RELATEDTO relation, constituting 61% of CN∩SW, has been excluded for the second analysis due to its broad and complex nature (it might correspond to multiple or none of WAX relations), and its dominant presence skews the distribution of other relations.

In Figure 4.4, we present the correlation between the f^r_{CN} and their recall in CN∩SW. Additionally, Figure 4.5 shows the distribution of $f^r_{\text{CN}\cap\text{SW}}$ and f^r_{WAX}. First, none of the six most frequent relation types in ConceptNet (Figure 4.4 right part), are highly recalled in CN∩SW. Out of these, relations indicating (near) synonymy retain a medium recall (Figure 4.4 red box on middle right), while morphosyntactic relations are less prevalent (Figure 4.4 blue box on bottom right). This aligns with the high-prevalence of taxonomic relations (including SYNONYM and CATEGORYEXEMPLAR) in WAX (Figure 4.5 first group) in contrast to the low-prevalence of morphosyntactic/linguistic relations (Figure 4.5 LINGUISTIC). Morphosyntactic relations are prevalent in ConceptNet, because they are largely derived

[11]The full list of relation mapping can be found in the Appendix Table B.1

ConceptNet and SWOW shared triples	
(relatedto, neither, nor)	(relatedto, none, nothing)
(synonym, no one, nobody)	(synonym, break down, fail)
(similarto, continuous, never ending)	(similarto, innocent, not guilty)
(distinctfrom fail success)	(distinctfrom, always, never)
ConceptNet negated triples	**SWOW negated triples**
(antonym, still with us, no longer with us)	(forwardassociated, love, non tangible)
(antonym, able, cannot)	(forwardassociated, real, not fake)
(synonym, zero, nothing)	(forwardassociated, everything, nothing)
(antonym, both, neither)	(forwardassociated, unimportant, nothing)
(capableof, clues, lead nowhere)	(forwardassociated, broke, no money)
(causes, going to sleep, never waking up)	(forwardassociated, delayed, not on time)
(hassubevent, eat quickly, barely chew)	(forwardassociated, blank, nothing)
(causes, dying, non existence)	(forwardassociated, give up, fail)
(hassubevent, eat healthily, dont eat junk food)	(forwardassociated, without, empty)
(relatedto, dare, you wont)	(forwardassociated, not allowed, rules)

Table 4.2 Comparative examples of negated triples in ConceptNet and SWOW. Top: pairs that are shared by both KGs with relationships sourced from ConceptNet. Bottom: triples that are exclusive to each individual KG.

from structured linguistic resources like Wiktionary.[12] Word associations, on the other hand, are known to be dominated by relations pertaining to concept meanings (Mollin, 2009). This implies that word associations reflect how concepts are interlinked by meaning and potentially provide insights into cross-cultural understanding in future work.

Second, we observe a discenerable correlation between the majority of low- to medium frequency relations in ConceptNet and their recall in CN∩SW (Figure 4.4 orange box on bottom left). These relations cover largely semantic associations pertaining to the appearance, use or situational contexts of concepts. The distribution of these relations in CN∩SW and word associations are largely aligned with SITUATIONAL in WAX (see Figure 4.5).

Third, the relations with highest recall in CN∩SW tend to be infrequent in ConceptNet (green box on top left). Two of these relations focus on *differences* indicating that humans associate contrasting concepts (such as 'hot'→'not cold'; Deese (1964), Clark (1970)). We inspected edges, referred to as 'negated edges', where at least one of the nodes represents a

[12] 74% of ConceptNet edges originate from Wiktionary. See Figure 2.8 in Section 2.2.1.1 for details of the knowledge source distribution in ConceptNet.

concept that is negated. Negated nodes were identified based on a list of negation markers.[13] We found that SWOW has higher proportion of negated nodes: 8% (N=2833) for SWOW[14] and 0.3% N=4116) for ConceptNet. Furthermore, we discovered that the proportion of negated edges in SWOW (0.7%; N=2.3K) is more than double that in ConceptNet (0.3%; N=11.5K), with only a 15% overlap. The main reason for this small overlap is that there are more interconnections between these negated nodes in SWOW. For pairs that do overlap, we observed that their relations in ConceptNet are mostly categorised as Synonym, Antonym, and DistinctFrom. Some negated triples sampled from ConceptNet and SWOW are presented in Table 4.2. Representations of antonyms and negations have traditionally been difficult to infer from text, suggesting that word associations could be a promising task to elicit more such kind of knowledge.

In summary, our analysis of the relational content suggests that word associations have discrepancies with ConceptNet in morphosyntactic relations, while their relation distribution in semantic relations is largely aligned.

4.4 Coverage of Commonsense Knowledge

In the prior section, we compared the knowledge structures and content of the two graphs, but we did not specifically probe how well they align with the commonsense knowledge that people use daily. Discerning this alignment is important as it offers insights into the difference between the everyday commonsense knowledge people use and what/how knowledge graphs capture. In this section, our focus shifts to examining ConceptNet and SWOW for their alignment and coverage of (situational) *commonsense* knowledge within scripts describing daily activities. These activities such as *doing the laundry* or *visiting the doctor* involve a wealth of general knowledge touching on causal, temporal, physical, or social knowledge which is rarely explicitly stated (Mostafazadeh et al., 2016, Rashkin et al., 2018, Ostermann et al., 2018, 2019).

[13]The negation markers includes: no, not, none, nor, no one, nobody, nothing, neither, nowhere, never, hardly, barely, scarcely, non, without, fail, cannot, can't, no longer, don't, won't.

[14]Associations provided by at least two participants were used.

To this end, we leverage the MCScript2.0 data set (Ostermann et al., 2019), a large collection of *explicit* descriptions of everyday scenarios, and investigate whether ConceptNet and SWOW encode the commonsense knowledge underlying these situations. In particular, we test whether the knowledge graph structure underlying MCScript scenarios is retained in ConceptNet and SWOW. To illustrate our investigation process, an example is presented in Figure 4.6 to visualise the pipeline. Within this figure, (a) provides a script text elaborating a scenario from the MCScript2.0 dataset, titled 'growing vegetables'. This knowledge is distilled and transformed into structured graphs in (b), capturing essential key concepts and their connections within (a). These structured graphs represent the commonsense knowledge that people commonly use in daily activities. To assess how well this knowledge aligns with ConceptNet and SWOW, we use a graph mapping method, which will be introduced in Section 4.4.2, to map the knowledge represented in (b) to the knowledge encoded within each of the KGs. Subsequently, Figure 4.6 shows a subset of the retrieved path-knowledge from ConceptNet (c) and SWOW (d).[15]

4.4.1 Dataset

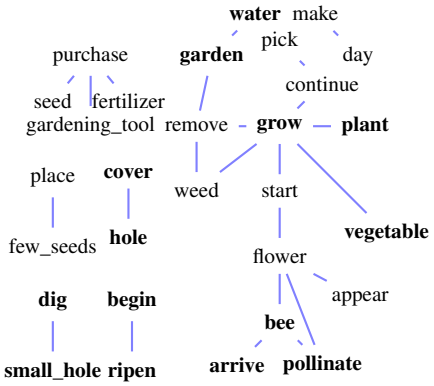
MCScript2.0 is a collection of 3,487 short narrative descriptions covering 200 every-day scenarios of varying complexity (e.g., *cleaning the floor* vs *growing vegetables*) (Ostermann et al., 2019). The descriptions were crowd-sourced, and authors were instructed to describe the underlying scenarios "as if talking to a child" (Ostermann et al., 2018). Thus by design MCScript narratives spell out commonsense knowledge more explicitly than most text corpora. Following prior work on modelling narrative scripts, we posit that narrative chains are fundamentally characterized in terms of their events and participants (Chambers and Jurafsky, 2009, Frermann et al., 2014). We recover this information using semantic role labelling (SRL) ,[16] and identify predicates, ARG0s and ARG1s in each narrative. We use the resulting set of spans and their relations to transform each narrative (Figure 4.6a) into a *script*

[15]For a full list of conversions from (b) to (c) and (d), refer to Table 4.3.

[16]We use the SRL model of Shi and Lin (2019) implemented in AllenNLP https://demo.allennlp.org/semantic-role-labelling.

Growing vegetables In the spring I went to the garden center to **purchase seeds, fertilizer**, and **gardening tools**. Back at home I **dug small holes** into the dirt in my garden, **placed** a **few seeds** in each hole, and **covered** the **holes** with dirt. I then **watered** the **garden**, and **made** sure to water it every **day**. Over the next several weeks I **removed** any **weeds** that were **growing** in the **garden**, and watched as the **plants continued** to **grow**. **Flowers** then **appeared** on the plants, and **bees arrived** to **pollinate** the **flowers**. Then, **started** to **grow** where the **flowers** had been. After a while they **began** to **ripen** and **continued** to **grow** until they were big enough to **pick**.

(a) script text



(b) script graph

grow — fruit — vegetable
grow — plant

water — earth — garden

bee — pollen — pollinate
bee — branch — leave — arrive

cover — opening — hole

begin — action — change — ripen

dig — hole — mouse — small_hole

(c) ConceptNet paths

grow — vegetable
grow — plant

water — garden

bee — pollinate
bee — fly — arrive

cover — band — hole

begin — blossom — ripen

no path retrieved

(d) SWOW paths

Figure 4.6 Illustration of a pipeline to assess situational knowledge in ConceptNet and SWOW using MCScript2.0: (a) Narrative describing the scenario of 'growing vegetables' from MCScript2.0; (b) derived SRL graph over predicates ARG0 and ARG1s; (c) shortest paths in ConceptNet for a subset of connected nodes (bolded) in (b); (d) corresponding paths retrieved from SWOW. See Table 4.3 for full directed connected paths for (b) and their shortest paths from ConceptNet and SWOW.

graph (Figure 4.6b). Spans constitute the nodes of the script graph,[17] and edges correspond to (undirected) SRL dependencies.

4.4.2 Method

We assume that predicate-argument relations in detailed descriptions of everyday events (e.g., *water* and *garden*) are instances of commonsense knowledge, and should hence be directly encoded in a KG. We posit that the shorter the paths, the more directly the necessary scenario-specific commonsense knowledge is encoded in the target KG. E.g., we find a direct 1-hop connection between *water* and *garden* in SWOW (Figure 4.6d), while the path between the concepts in ConceptNet is less direct (2-hops; Figure 4.6c).

[17] Pronouns and stopwords are excluded.

	MCScript2.0	ConceptNet	SWOW
1	(purchase, seed)	(purchase, sale, full, seed)	(purchase, need, seed)
2	(purchase, fertilizer)	(purchase, chain, garage, fertilizer)	(purchase, product, fertilizer)
3	(hole, cover)	(hole, opening, cover)	(hole, band, cover)
4	(flower, appear)	(flower, visit, appear)	(flower, become, appear)
5	(flower, pollinate)	(flower, pollen, pollinate)	(flower, bee, pollinate)
6	(flower, start)	(flower, open, start)	(flower, green, start)
7	(flower, bee)	(flower, bee)	(flower, bee)
8	(grow, continue)	(grow, carry, continue)	(grow, extend, continue)
9	(grow, vegetable)	(grow, fruit, vegetable)	(grow, vegetable)
10	(grow, plant)	(grow, plant)	(grow, plant)
11	(grow, weed)	(grow, field, weed)	(grow, weed)
12	(continue, pick)	(continue, carry, pick)	(continue, stick, pick)
13	(pollinate, bee)	(pollinate, pollen, bee)	(pollinate, bee)
14	(ripen, begin)	(ripen, change, action, begin)	(ripen, blossom, begin)
15	(garden, water)	(garden, earth, water)	(garden, water)
16	(garden, remove)	(garden, cricket, remove)	(garden, wart, remove)
17	(weed, remove)	(weed, remove)	(weed, remove)
18	(small hole, dig)	(small hole, mouse, hole, dig)	-
19	(bee, arrive)	(bee, branch, leave, arrive)	(bee, fly, arrive)
20	(day, make)	(day, clear, make)	(day, present, make)
21	(few seed, place)	-	-
22	(gardening tool, purchase)	(gardening tool, tool, lever, purchase)	-

Table 4.3 Full list of paths from ConceptNet and SWOW for example shown in Figure 4.6. - indicates path are not retrieved in the target KG.

Graph Mapping More specifically, we project the MCScript graphs onto SWOW and ConceptNet as follows. For all directly connected node pairs in a MCScript graph ($\mathcal{G}_s$), we identify their corresponding concepts in a target graph $\mathcal{G}_t$, which can be either SWOW or ConceptNet, through exact string matching. In our current study we are primarily interested in *whether* (not how) node pairs from the MCScript graph exist in the target KGs. We therefore treat all graphs as undirected, and retrieve the shortest path between the two nodes in the target KGs. For example, for a connected pair (*begin*, *ripen*) in the $\mathcal{G}_s$, the corresponding paths retrieved from ConceptNet and SWOW are (*begin*, *action*, *change*,*ripen*) and (*begin*, *blossom*, *ripen*), respectively. Figure 4.6c and 4.6d show a subset of the shortest paths retrieved from ConceptNet and SWOW, respectively, for the MCScript graph in 4.6b (bolded nodes).

We employ two metrics to evaluate the coverage and alignment of MCScript graphs within ConceptNet and SWOW. Firstly, we utilize the concept of edge recall to gauge the

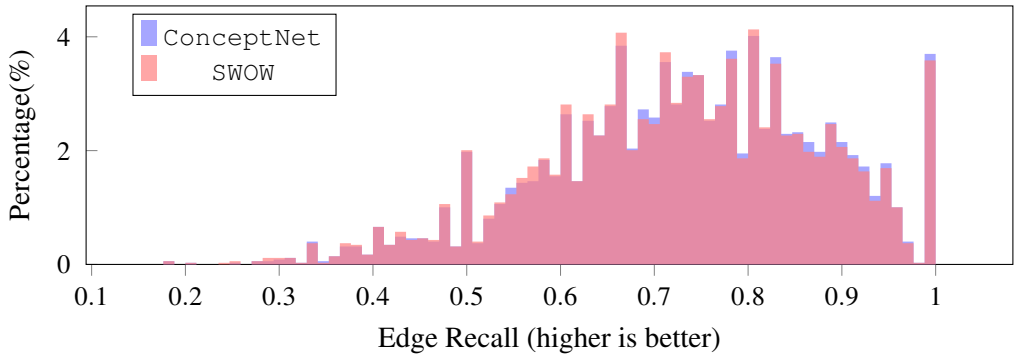


Figure 4.7 The edge recall distribution of ConceptNet and SWOW for MCScript graphs.

extent to which a target KG ($\mathcal{G}_t$) encompasses the content of a source KG ($\mathcal{G}_s$). Edge recall is expressed as a ratio, denoting the fraction of edges in $\mathcal{G}_s$ that can be identified in $\mathcal{G}_t$ by tracing the shortest paths between their corresponding nodes. A higher edge recall value signifies a larger percentage of edges in $\mathcal{G}_s$ that can be "recalled" or matched in $\mathcal{G}_t$, as formulated below:

$$\text{EdgeRecall} = \frac{|E(\mathcal{G}_s) \cap E(\mathcal{G}_t)|}{|E(\mathcal{G}_s)|}, \tag{4.1}$$

where $E(\mathcal{G}_s)$ and $E(\mathcal{G}_t)$ represents the set of edges in the source and target knowledge graph, respectively.

Secondly, for each graph $\mathcal{G}_s$, we calculate the average path length of the retrieved shortest paths from $\mathcal{G}_t$. A shorter path length indicates a more direct alignment between $\mathcal{G}_s$ and $\mathcal{G}_t$.

4.4.3 Results

Figure 4.7 presents the distribution for the edge recall, showing that ConceptNet and SWOW have similar recall for these graphs. In both KGs, a substantial proportion of edges are successfully recalled, with an average recall rate of approximately 70%. This observation suggests that both ConceptNet and SWOW encode the commonsense knowledge relevant to daily scenarios in MCscript2.0.

Figure 4.8 presents the distribution over shortest path lengths, averaged over the full MCScript2.0 data set. We observe that paths, on average, are shorter in SWOW compared

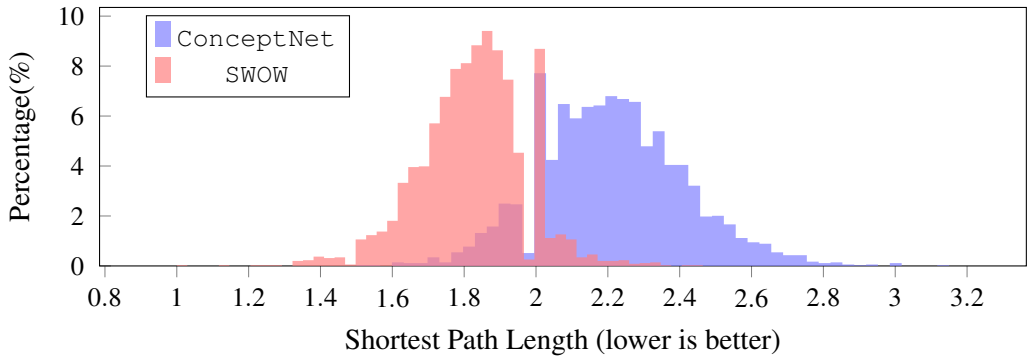


Figure 4.8 The length distribution of average shortest paths in ConceptNet and SWOW for edges from MCScript graphs.

to ConceptNet, suggesting that the situational commonsense associations in MCScript are more directly encoded in SWOW. The example shortest paths shown in Figure 4.6c (ConceptNet) and d (SWOW) further illustrate the associations in the two commonsense resources. The associations imposed in paths of length >1 are meaningful across the board, but differ across the KGs: for example, the connection between *bee* and *arrive* is further elaborated in ConceptNet by explaining that in order to arrive, the bee needs to leave (from a plausible location *branch*); SWOW on the other hand imputes information on the mode of travel (*flying*).

To further understand whether shorter paths in SWOW are seen as feasible, we randomly sampled paths from all MCScript stories, as presented in Table 4.4.[18] We observed that paths in SWOW are largely reasonable and align more closely with our direct daily interactions. For example, cue-associations such as (*tooth*, *brush*) and (*clothe*, *wear*), which we encounter directly every day, are not directly connected in ConceptNet. Our analyses suggest that SWOW encodes a comparable amount of situational commonsense knowledge as ConceptNet but tend to encode it more compactly.

[18]Note that multiple shortest paths between concepts may exist; we randomly chose one for analysis.

MCScript	ConceptNet	SWOW
(stewardess, soda)	(stewardess, stew, cow, milk, soda)	(stewardess, soda)
(carrot, chop)	(carrot, stick, tree, chop)	(carrot, chop)
(bike, trail)	(bike, car, road, trail)	(bike, trail)
(sew, mother)	(sew, cut, slang, mother)	(sew, mother)
(lego, play)	(lego, toy store, action figure, play)	(lego, play)
(grandma, visit)	(grandma, relatives house, relative, visit)	(grandma, visit)
(stylist, towel)	(stylist, stylistics, linguistics, dialect, towel)	(stylist, hair, towel)
(coffee, sip)	(coffee, drink, sip)	(coffee, sip)
(face, wash)	(face, make, wash)	(face, wash)
(tooth, brush)	(tooth, brushwheel, brush)	(tooth, brush)
(clothe, wear)	(clothe, dress, wear)	(clothe, wear)
(egg, scoop)	(egg, person, music, scoop)	(egg, spoon, scoop)
(painting, think)	(painting, pick, decide, think)	(painting, work, think)
(pizza, warm)	(pizza, cheese, yellow, warm)	(pizza, food, warm)
(suitcase, get)	(suitcase, airport, arrive, get)	(suitcase, grip, get)
(fruit, want)	(fruit, box, need, want)	(fruit, food, want)
(breakfast, prepare)	(breakfast, dinner, provide, prepare)	(breakfast, food, prepare)

Table 4.4 Examples of paths that are longer in ConceptNet than SWOW. Top: ConceptNet paths are more than 2 hops longer than those in SWOW; Bottom: paths are one hop longer.

4.5 Word Associations for Commonsense QA

In this section, we explore the utility of commonsense knowledge from ConceptNet and SWOW in tasks necessitating commonsense reasoning. We focus on the commonsense question answering (CQA) task, where a question like "Why might someone send flowers to a friend who is feeling down?" is provided. To answer such questions, a model is expected to use commonsense knowledge that goes beyond the direct context of the posed question, as detailed in Section 2.4.2. The performance in this task can provide insights into a model's commonsense reasoning capability, which is shaped by factors like model architectures, inherent commonsense knowledge, and task difficulty.

In this study, we use ConceptNet and SWOW as two external commonsense KGs. We incorporate them into various CQA models that were considered representative and competitive at the time of conducting our research (Wang et al., 2020, Feng et al., 2020), and apply them to three benchmark data sets. We emphasise that the goal of this study is not

Dataset	Train / Dev / Test split	Example
CSQA	8,500/1,221/1,241	What do all humans want to experience in their own home? (a) **feel comfortable**, (b) work hard, (c) fall in love, (d) lay eggs, (e) live forever
OBQA	4,957/500/500	What is a source of energy? (a) bricks, (b) **grease**, (c) cars, (d) dirt
MCScript2.0	14,191/2,020/3,610	When did small plants grow? (a) two days, (b) **after seeds were planted**

Table 4.5 Details on the benchmarks CSQA, OpenbookQA and MCScript2.0: One example QA-pair per dataset (correct answer in boldface) and sizes of the respective train/dev/test splits. The paragraph of MCScript2.0 example is shown in Figure 4.6.

competing on leaderboards. Recent models leverage very large language models with billions of parameters (Khashabi et al., 2020), and often draw on additional external resource such as Wiktionary (Xu et al., 2021). Instead, we explore the utility of `SWOW` and `ConceptNet` in a selection of representative knowledge graph encoding models.

4.5.1 Datasets

Benchmarks We consider three standard multiple-choice CQA benchmark datasets. **CommonsenseQA** (CSQA; Talmor et al. (2019)) contains commonsense questions generated by crowd workers on the basis of sub-graphs in `ConceptNet`, giving `ConceptNet` an inherent advantage over `SWOW`. The QA-pairs in this dataset require various commonsense skills, and distractor answers were carefully selected to share semantic associations with the key concepts in a question. Each question in CSQA is provided with five answer choices. **OpenBookQA** (OBQA; Mihaylov et al. (2018)) consists of question-answer pairs along with science facts from elementary-level science books. Following previous work (Wang et al., 2020, Feng et al., 2020), we disregard the facts, and apply our models to question-answer pairs directly. Four answer options are given for every question in OBQA. Building on our analysis in Chapter 4.4, we also apply our models to the **MCScript2.0** QA benchmark (Ostermann et al., 2019). Each task consists of a story, and a question paired with two answer options. Since the questions lose their meaning without the context of the story, especially

17 relation types			
1	atlocation, locatednear	10	usedfor
2	capableof	11	receivesaction
3	createdby	12	madeof
4	desires	13	partof, hasa
5	hascontext	14	notdesires
6	hasproperty	15	notcapableof
7	antonym, distinctfrom	16	isa, instanceof, definedas
8	relatedto, similarto, synonym	17	causes, causesdesire, motivatedbygoal
9	hassubevent, hasfirstsubevent, haslastsubevent, hasprerequisite, entails, mannerof		

7 relation types	
1	capableof
2	usedfor, receivesaction
3	atlocation, locatednear, hascontext, similarto
4	causes, causesdesire, motivatedbygoal, desires
5	antonym, distinctfrom, notcapableof, notdesires
6	isa, hasproperty, madeof, partof, definedas, instanceof, hasa, createdby, relatedto, synonym
7	hassubevent, hasfirstsubevent, haslastsubevent, hasprerequisite, entails, mannerof

Table 4.6 Conflation of relation types in ConceptNet to either 17 (left) or 7 (right) coarser grained groups. We start with grouping the 31 original ConceptNet relation types into 17 clusters following (Wang et al., 2020). We further group the 31 relation types into 7 by following (Liu and Singh, 2004), which grouping relation types into 7 categories, including {things, spatial, events, causal, affective, functional, agents}.

when referencing names or objects from the story, we concatenate the story with each question to form a single sequence during modelling. We use the in-house data split by Lin et al. (2019a) for CSQA and the official splits for the other data sets. Table 4.5 presents data set statistics, as well as an example from each dataset.

Knowledge Graphs Following previous work on CQA (Lin et al., 2019a, Wang et al., 2020), we select 31 out of ConceptNet's 47 relation types which proved helpful for CQA, and merge the remaining relations into 17 types (cf., Table 4.6 left). We use forward- and mutual associations for SWOW. The number of relations (17 for ConceptNet and 2 for SWOW) are used to train TransE for knowledge graph embeddings. In align with previous work (Malaviya et al., 2020, Wang et al., 2020), we densify both SWOW and ConceptNet by introducing the reverse relation r^{-1} for each $r \in \mathbb{R}$, prefixed with the symbol _. For instance, alongside the relation r=*UsedFor*, we add r^{-1}=*_UsedFor*. Consequently, if an original KG contains a triple like (*television* , USEDFOR, *watching*), we would add an new triple of (*watching*, _USEDFOR, *television*) into the KG.

4.5.2 Method

To compare the effectiveness of `ConceptNet` and `SWOW` on the task of commonsense question answering, we experiment them with KG-augmented QA systems. As briefly introduced in Section 2.4.2, such systems typically consist of three modules: a text encoder, a KG encoder, and a scoring module. Specifically, for a question-answer pair $\{q, a\}$, a text encoder learns a language embedding $\mathbf{c} \in \mathbb{R}^{d_c}$ using the $\{q, a\}$ pair. A KG encoder then generates a knowledge embedding $\mathbf{k} \in \mathbb{R}^{d_k}$ by modelling the grounded question concepts $\mathcal{C}^q$ and answer concepts $\mathcal{C}^a$ (also the relations connecting them). The $\mathcal{C}^q$ and $\mathcal{C}^a$ are obtained through lexical matching of q and a with nodes in a KG by following Lin et al. (2019a). Then, the scoring module transforms the concatenation of $\mathbf{c}$ and $\mathbf{k}$ into a single scalar s with a linear transformation: $s_a^q = \mathbf{w}^\top [\mathbf{c}, \mathbf{k}] + b$. The process is repeated for all answer options a, and the final answer is predicted as the maximum scoring answer.

Text-Encoder We use ALBERT-xxlarge-v2 (Lan et al., 2020)[19] as our text encoder, which performed competitively in Wang et al. (2020). We use this model as a text-only baseline to gauge the improvements of adding external knowledge from two graphs with different KG encoders.

KG-Encoders We experiment with four KG encoders,[20] broadly grouped into static and dynamic models. The static models, **GconAttn** (Wang et al., 2019) and **Relation Networks (RN)** (Santoro et al., 2017), focus on retrieving pre-existing KG knowledge. For dynamic KG encoder, we test **PG-Global** (Wang et al., 2020), which incorporates knowledge embedding $\mathbf{k}$ dynamically generated for concept pairs from a $\{q, a\}$ pair. Additionally, we employ **PG-Full** (Wang et al., 2020), which unifies the static knowledge embedding from RN and the dynamic knowledge embedding from PG-Global. We introduce these models below.

[19] Refer to Section 2.3.1.3 for a more detailed description of ALBERT.
[20] For further details on KG encoder background, please refer to Section 2.3.2.

GconAttn represents question concepts $\mathcal{C}^q$ as well as answer concepts $\mathcal{C}^a$ with pre-trained concepts embeddings, then aligns them with concept-level attention and max pooling.

$$\mathbf{k} = g([\mathcal{C}^q, \mathcal{C}^a]) = \text{MAX}(\text{Attn}([\mathcal{C}^q, \mathcal{C}^a])), \tag{4.2}$$

The Attn(.) is a two-way attention (Seo et al., 2017) that aligns each $\mathcal{C}^q_i$ with each $\mathcal{C}^a_j$ with dot-product attention and captures their differences and similarities. GconAttn is a relation-free model, leveraging only mentioned concepts from a KG.

To disentangle the impact of relation labels, we also experiment with **RN**, which generates a knowledge embedding $\mathbf{k}$ by performing context-aware path-level attention over path embeddings. Each path embedding $\mathbf{p}_i$ encodes a path, retrieved from a KG, that connects some question concept $\mathcal{C}^q_i$ and answer concept $\mathcal{C}^a_j$. We compute attention weights α_{p_i} by combining text- and path representations, and apply them to $\mathbf{p}_i$ to obtain the final KG embedding. Formally,

$$\mathbf{k} = \sum_{p_i \in \mathcal{P}} \alpha_{p_i} \mathbf{p}_i, \tag{4.3}$$

$$\mathbf{p}_i = \text{MLP}([\mathcal{C}^q_i; r_1, \cdots, r_T; \mathcal{C}^a_j]), \tag{4.4}$$

$$\alpha_{p_i} = \mathbf{c}^\top \tanh\left(\mathbf{W}_{att} \cdot \mathbf{p}_i + \mathbf{b}_{att}\right), \tag{4.5}$$

where, the $\mathbf{W}_{att}$ is used to project $\mathbf{p}_i$ into the same space as $\mathbf{c}$ and MLP is used to encode the path representation for the path sequence. Following Wang et al. (2020), we use one-hop and two-hop paths retrieved from KGs in our experiments.

PG-Global, structurally similar to RN, differentiates itself by dynamically generating path embeddings using a path generator, which is a fine-tuned GPT-2 (Radford et al., 2019) on sampled paths from a KG (e.g., *news programming*, ATLOCATION, *television*, USEDFOR, *watching*). During fine-tuning, the model learns to connect two concepts with a multi-hop path. For each question concept $\mathcal{C}^q_i$ and answer concept $\mathcal{C}^a_j$, the path generator will generate the most likely path (with intermediate concepts and relations) to connect them. Each path embedding is obtained by mean pooling over the hidden states of the last layer of the path

generator:[21]

$$\mathbf{p}_i = \text{MEAN}(\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_t), \tag{4.6}$$

where the $\mathbf{h}_i$ is the hidden state of each token in the path. The final KG embedding $\mathbf{k}$ is calculated the same as Equation (4.3) across all paths. For **PG-Full**, the $\mathbf{k}$ is the concatenation of the output from RN and PG-Global. Note that PG-Global and PG-Full are also relation-aware encoders as they use relational paths.

4.5.3 Experimental Setup

4.5.3.1 Pre-processing

Both GConAttn and RN require *node embeddings*. We use RoBERTa-Large (Liu et al., 2019) to obtain a node embedding matrix, separately for `ConceptNet` and `SWOW`. Specifically, for each concept node $\mathcal{C}_i \in \mathcal{C}$, we feed the sequence of `[CLS]` $+\mathcal{C}_i+$ `[SEP]` to RoBERTa and use the last layer representation of `[CLS]` as its embedding. RN also requires *relation embeddings*. For each of our KGs (`ConceptNet`, `SWOW`), we obtain a separate relation embedding matrix $\mathbf{R}$ with TransE to initialize the relation matrix in RN. We use OpenKE [22] to train TranE model on our KGs. Note that we do not use TransE node embeddings in our experiments, as they were outperformed by RoBERTa embeddings in preliminary tests. For a detailed introduction on TransE, please refer to Section 2.3.2.1.

4.5.3.2 Training

We use cross-entropy as the loss function with RAdam (Liu et al., 2020) as optimizer to train all models. We use as GELU (Hendrycks and Gimpel, 2016) as the activation function. We report the most important hyper-parameters in Table 4.7.

All models are run six times and we report results using the best three models, as judged by training loss; this method is used to remove outliers resulting from instability of training.

[21] For path generator's pre-processing and fine-tuning details, see Section 2.3.2.2.
[22] https://github.com/thunlp/OpenKE

Type	Hyperparameter	Value
General	batch size	32/16/16
	dropout	0.1/0.2/0.1
	early stopping patience	2 epochs
	max sentence length	80/84/300
	weight decay	0.01
ALBERT-xxlarge-v2	learning rate	1e-05
GconAttn	learning rate	3e-04/3e-04/1e-03
	MLP layers	2
	hidden units	{256, 128}
	concept embedding dimension	1024
RN	learning rate	1e-03/3e-04/1e-03
	MLP layers	3
	hidden units	{256, 256, 128}
	concept embedding dimension	1024
	relation embedding dimension	100
PG-Global	learning rate	1e-03/3e-04/1e-03
	MLP layers	3
	hidden units	{256, 256, 128}
	path embedding dimension	768
PG-Full	learning rate	1e-03/3e-04/1e-03
	MLP layers	3
	hidden units	{256, 256, 128}
	concept embedding dimension	1024
	relation embedding dimension	100
	path embedding dimension	768

Table 4.7 Hyperparameters for various models and data sets. Values split by "/" follow the order of CSQA/OBQA/MCScript2.0.

All results are our own re-runs using the official implementations from Feng et al. (2020),[23] and are largely comparable with those reported in the literature. Our reproduction results are presented in Appendix C.

4.5.4 Results

We assessed all models on the test sets of three datasets, with accuracy as the chosen evaluation metric. Experimental results in Table 4.8 show that all knowledge-augmented models outperform the language baseline (ALBERT) for all data sets except RN on MCScript2.0. For static KG models, the path-aware RN achieves better performance for both CSQA and OBQA. The dynamic KG model PG-Global achieves comparable results with RN. Addi-

[23] https://github.com/INK-USC/MHGRN

Models	CSQA		OBQA		MCScript2.0	
	ConceptNet	SWOW	ConceptNet	SWOW	ConceptNet	SWOW
ALBERT	73.78 (±0.79)		63.47 (±1.42)		93.62 (±0.44)	
+ GconAttn	74.03 (±0.46)	74.05 (±0.50)	65.13 (±2.16)	65.87 (±1.21)	93.91 (±0.50)	93.84 (±0.35)
+ RN	75.64 (±0.70)	74.40 (±0.37)	64.73 (±2.10)	66.40 (±1.00)	93.53 (±0.11)	93.49 (±0.22)
+ PG-Global	74.40 (±0.17)	74.38 (±0.66)	66.20 (±0.92)	67.27 (±0.81)	94.28 (±0.04)	94.34 (±0.02)
+ PG-Full	**76.85** (±0.61)	74.78 (±1.38)	67.80 (±2.03)	**67.93** (±0.12)	94.50 (±0.16)	**94.71** (±0.38)

Table 4.8 Test accuracy on CSQA, OBQA and MCScript2.0. We report performance of ALBERT as text-only baseline, and augment it with four KG-aware models using either ConceptNet or SWOW. Results are averages of the best three out of six runs (based on dev set performance); standard deviations reported in brackets. KG-augmented models significantly outperform the text-only model ($p<0.05$), with the exception of ALBERT vs RN with ConceptNet on OBQA ($p=0.07$). No significant difference between ConceptNet-based and SWOW-based models observed ($p>>0.05$).

tionally, PG-Full achieves the best performance, suggesting that the dynamic and static knowledge are complementary.

All models, including the text-only baseline, show comparative and high performance on MCScript2.0. Our models outperform the state-of-the art on MCScript2.0 by up to 4.11% (absolute), according to the COIN leaderboard.[24] Our results suggest that MCScript2.0 is a simpler task compared to the other two. This may be attributed to the rich story information providing context signals and the two-way classification. In our subsequent analyses, we will focus on CSQA and OBQA.

More importantly, models incorporating either ConceptNet or SWOW achieve similar performance across the board. Recall that CSQA is derived from ConceptNet edges, putting SWOW at a disadvantage. SWOW performs best on OBQA and MCSCript2.0 which are independent of both KGs. We measure the significance of differences in performance between the text-only baseline and the KG-augmented models using Student's t-test. We find that the KG-augmented models outperform the text-only model significantly ($p<0.05$) with both ConceptNet and SWOW as underlying KG.[25] There is no significant difference between the ConceptNet-based and the SWOW-based models ($p>>0.05$). These results

[24] https://coinnlp.github.io/task1.html

[25] The only exception is ALBERT vs RN with ConceptNet on OBQA where $p=0.07$.

PG-KG	RN-KG	CSQA		OBQA	
		dev	test	dev	test
ConceptNet	ConceptNet	80.59 (±0.36)	76.85 (±0.61)	69.00 (±0.87)	67.67 (±1.68)
SWOW	SWOW	79.23 (±0.45)	74.78 (±1.38)	70.07 (±0.70)	68.13 (±0.83)
SWOW	ConceptNet	80.34 (±0.50)	76.66 (±0.20)	70.07 (±1.33)	67.40 (±1.00)
ConceptNet	SWOW	79.28 (±0.46)	74.32 (±1.24)	69.40 (±0.53)	68.73 (±2.32)
Ensemble (ConceptNet)		**81.98**	**78.24**	71.00	70.20
Ensemble (SWOW)		79.85	76.47	72.20	69.60
Ensemble (ConceptNet + SWOW)		**81.98**	77.92	**72.60**	**71.60**

Table 4.9 Impact of combining ConceptNet and SWOW in ALBERT+PG-Full for CSQA and OBQA. PG-KG is the dynamic, and RN-KG the static component of PG-Full. Top: KG-specific models for comparison. Middle: combining KGs in different components. Bottom: ensembling of KG-sepecific models.

provide initial evidence that SWOW can be a valuable alternative source of commonsense knowledge to ConceptNet for downstream NLP tasks.

Combining ConceptNet and SWOW ConceptNet and SWOW contain different knowledge biases, which naturally leads to the question of whether their knowledge is complementary and their combination can lead to additional improvements to downstream tasks. We considered a variety of ways of incorporating both KGs into PG-Full+ALBERT. First, we use different KGs in different components of the dynamic and static KG encoders; second, we ensemble models from separate runs, which included (a) three SWOW-based, (b) three ConceptNet-based PG-Full and (c) a combination of the six models from both (a) and (b).[26] We averaged the probabilities across of runs and selected the label with the highest probability.

Table 4.9 presents the results. Similar to Table 4.8, we found that SWOW does not bring as much benefits to CSQA as ConceptNet across the board, because CSQA is derived from ConceptNet and hence can be distracted by adding additional external knowledge. For OBQA, combining different KGs in different PG-Full components (middle) achieves comparable results with the models based on a single KG (top). Although ensembling six models (ConceptNet and SWOW) leads to the best overall results, the improvements

[26]Initial experiments with re-training the QA models on the union of both KGs did not produce encouraging results.

compared to ensembled single KG models are not substantial. Our results suggest that SWOW could serve as an alternative to ConceptNet, both as a dynamic KG and a static KG. However, this also indicates that for the task of CSQA, ConceptNet and SWOW do not effectively complement each other.

Impact of the Numbers of Relation Types The competitive performance of SWOW may be surprising, particularly with the relation-aware static KG encoder RN that solely retrieves the existing knowledge in SWOW. ConceptNet has access to a rich typed relation inventory while SWOW does not. We investigate the impact of labelled relation types on CQA for two KG encoders (RN and PG-Full) that use retrieved relational knowledge from a KG, by ablating the number of relation types accessible to ConceptNet. We grouped the 17 ConceptNet relation types used in the models above into (a) seven coarse types (plus reverse relations) using the relational ontology of Liu and Singh (2004), see Table 4.6 for details of 17 and 7 relation types; (b) a single generic relation type (plus reverse relation). This version still contains one-hop and two-hop paths as well as reverse relations; and (c) removing all relation information from the model. We removed the relation information by excluding the relation embeddings in Equation (4.4) when calculating the path embedding **k** for RN, which is originally derived from both concept embeddings and relation embeddings. For PG-Full, we retain the path generator trained with 17 relation types but use the **k** derived from RN with different number of relations.

Figure 4.9 presents the experimental results. All models consistently outperformed the ALBERT baseline, highlighting the crucial role of external KG knowledge. The least performance gain observed when relation information was entirely absent, showing the important nature of relational knowledge, irrespective of its granularity. When varying the number of relation types, there is no clear pattern across RN and PG-FULL. This is interesting as it implies that the granularity of relations does not linearly affect the efficacy of a KG in commonsense question answering tasks. This might also explain why SWOW achieved comparable results as ConceptNet.

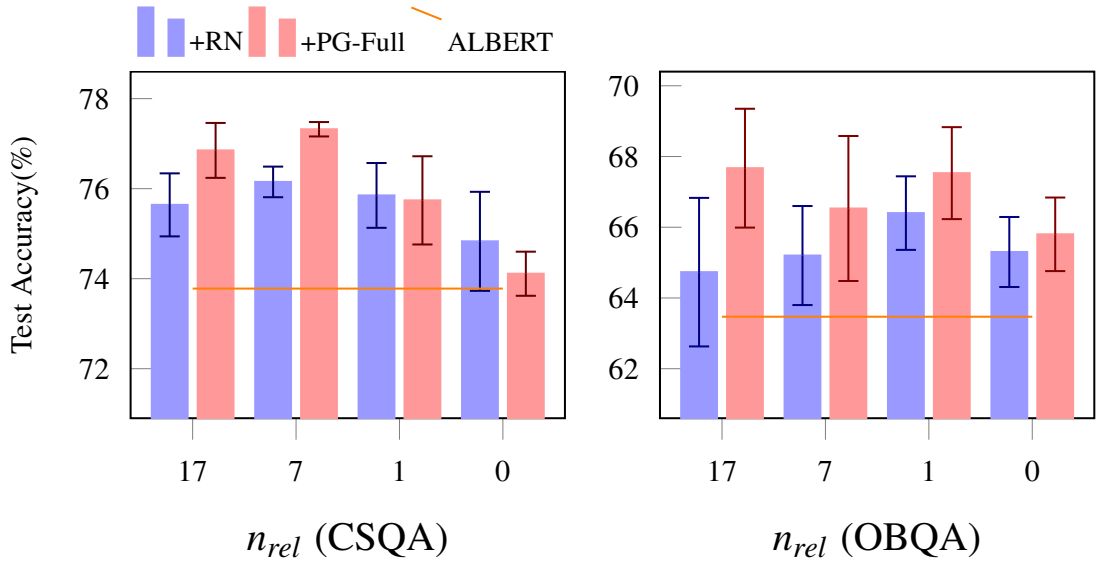


Figure 4.9 Test set accuracy under different numbers of relation labels for `ConceptNet` on CSQA (left) and OBQA (right).

Interestingly, when models were provided with a singular generic relation type, RN achieved the best on both datasets. In contrast, PG-Full's performance was on par with scenarios using 17 and 7 relation types for OBQA, but slightly lower on CSQA. Notably, this single generic relation in `ConceptNet` mirrors the structure of `SWOW`. This suggests that the effectiveness of a KG lies not just in the variety of its relations, but its internal structure and connectivity. Our results suggest that augmenting `SWOW` with a rich label inventory may not be necessary for it to be used as a commonsense resource in downstream commonsense reasoning models. Nevertheless, labelling `SWOW` with cognitively valid relation (or commonsense type) information in order to better understand the types of spontaneous associations humans express is an exciting avenue for future work. Although our WAX dataset, introduced in Chapter 3, serves as an exploratory study to understand the relations underlying word associations using self-explanations, labelling `SWOW` requires scaling up explanations, which is difficult to obtain. We further discuss this direction in Chapter 6.

4.6 Limitations and Discussion

In this study, we conducted a series of experiments aimed at understanding the knowledge encapsulated in ConceptNet and SWOW, as well as exploring their respective utility. In this section, we discuss certain limitations in our studies, potential alternatives and avenues for future exploration.

Graph Mapping in MCScript2.0 In Section 4.4, we presented an investigation into the breadth of situational commonsense knowledge encapsulated within commonsense KGs. To this end, we transformed textual narratives from the MCScript2.0 dataset into graph structures using semantic role labelling (SRL). These graph representations were then aligned with both ConceptNet and SWOW to facilitate a comparative analysis of path coverage and lengths. During the transformation process from text to graph, we used the SRL model proposed by (Shi and Lin, 2019), a BERT-based probabilistic model that exhibits an accuracy of approximately 86% on dependency-based SRL tasks in out-of-domain evaluations. Such accuracy suggests the possibility of frame identification errors, potentially leading to inaccuracies in the graph conversion. The focus on primary semantic roles—predicates, ARG0s, and ARG1s—may also omit complex information such as locations and temporality.

Future enhancements should address these limitations. The script graphs themselves could be improved and enriched with more semantic roles, or higher-level narrative structure as captured for instance in Rhetorical Structure Theory (Taboada and Mann, 2006). Replacing string mapping to commonsense knowledge with embedding-based methods would provide flexibility, and adding edge directions and labels to mapped graphs could enhance structural clarity. Crucially, paths longer than one hop require validation and human interpretation to ensure reliability. We believe that leveraging explicit human-created commonsense data sets, like MCScript2.0, opens interesting avenues to understand the commonsense knowledge present in word associations.

ConceptNet vs. SWOW: Complementary or Alternative In this chapter, we aim to understand the relationship between SWOW and ConceptNet, and whether they are com-

plementary or alternative. In Section 4.3, we focused on the knowledge within both KGs and observed limited overlap. However, in Section 4.5, we found that they brought comparable improvements when applied to commonsense question answering tasks. We also conducted experiments to explore their roles as dynamic and static KGs (see Table 4.9) and found that using both of them did not yield substantial improvements.

To understand this gap, we inspect the coverage of concepts in the two data sets in SWOW and ConceptNet, counting the questions and answer with at least one retrieved KG concept node (by exact match). For both KGs and data sets at least one concept is retrieved for 97% of the questions and answers. We also analyse the overlap of covered nodes in ConceptNet and SWOW to capture the difference in covered knowledge. We found that grounded concept coverage is higher in ConceptNet for both data sets, with 24,636 (CSQA) and 18,764 (OBQA) concepts from ConceptNet and 16,536 (CSQA) and 14,179 (OBQA) from SWOW. Notably, there is high overlap in covered nodes between the KGs, with $>$ 89% of nodes in SWOW are also in ConceptNet. This could also be one factor contributing to the comparable results in Table 4.9 in OBQA. Considering their similar downstream performance (especially on GConAttn), this suggests that both KGs contribute similar concept-level commonsense knowledge to the two datasets. Based on our findings, we believe that SWOW serves as an alternative commonsense resource to ConceptNet for downstream commonsense reasoning tasks.

KG Comparisons In Section 4.3, we investigated the knowledge content overlap of ConceptNet and SWOW using the directed connected triples (or one-hop paths). Considering that 58% of SWOW concept nodes are present in ConceptNet, examining whether cue-association pairs are connected in ConceptNet beyond one-hop would provide extra information about the degree of knowledge content overlap between the two KGs. A subsequent work (Yao et al., 2022) followed this idea, linking cue-association pairs in SWOW with two KGs, WordNet (see Section 2.1.2.1) and ASCENT++ (Nguyen et al. (2023); an enriched version of ConceptNet). Specifically, they retrieve the shortest path within a KG that connects each cue-association pair in SWOW. Interestingly, they found that: (a) 20%

of SWOW pairs are not connected in either of the KGs, and (b) of those retrieved paths, the one-hop connections make up 9%, while the majority (57%)[27] are connected with two and three hops. However, we currently lack a deeper understanding of two aspects: why 20% of the pairs are not connected in word associations and why the majority of pairs are connected through longer paths in other KGs. Are these phenomena due to the sparsity of these target KGs, or do they related to the high-level cognitive process of generating word associations, causing the unique shortcut phenomenon? This also aligns with our finding in Section 4.4 that word associations align closer with this commonsense knowledge being used to describe daily activities. This opens future opportunities on further investigating underlying factors that lead to the multi-hop connections and labelling large-scale word associations like SWOW, shedding light on the interplay between human cognition, knowledge representation, and the structure of knowledge graphs.

4.7 Summary

In this chapter, we addressed the second research question of how large-scale word associations encode commonsense knowledge and their utility in downstream tasks. We presented an in-depth analysis of the general and commonsense knowledge encoded in human word association norms (SWOW), versus a traditional curated commonsense knowledge graph (ConceptNet). We showed that the two knowledge resources differ systematically in their structure and content. We also showed that SWOW encodes situational commonsense knowledge as encoded in the human-created MCScript2.0 narratives more directly than ConceptNet; and that both KGs impose meaningful additional relations between concepts that were left implicit in the descriptions.

Finally, we illustrated that SWOW and ConceptNet bring comparable gains compared to using the text-only pre-trained language models, when applied to three commonsense question answering benchmarks. This finding is important as word associations are simpler than structured relations, and accordingly can be created more cheaply via crowd-sourcing

[27] Source from https://github.com/U-Alberta/WordTies/blob/main/data/swow-full-merged.jsonl.gz

and without the need for experts. It shows that large-scale word associations can serve as an alternative resource to `ConceptNet` and provide complementary knowledge to existing pre-trained models for commonsense reasoning tasks.

In the following chapter, we investigate the connection between knowledge in word associations and pre-trained language models, and discover that word association can aid in the extraction of implicit knowledge from these models.

Chapter 5

Robust Hypernym Extraction from BERT with Anchors and Word Associations

Our exploration so far has delved into the relational structure of word associations and their potential as a commonsense knowledge graph. By using pre-trained language models to capture the relations in word association explanations (Chapter 3), we discovered their ability to predict relations between associated word pairs, particularly for specific relation types (e.g., SYNONYM and LOCATION). Moreover, augmenting pre-trained language models with word associations enhances their performance in commonsense question-answering tasks (Chapter 4). This suggests that the knowledge encapsulated within pre-trained language models and word associations may be complementary.

Pre-trained language models, as discussed in Section 2.3.1, have been shown to capture linguistic properties and broader knowledge from vast text corpora (Petroni et al., 2019, Durrani et al., 2020, AlKhamissi et al., 2022). However, this knowledge remains implicit, hidden within model parameters, which makes achieving a comprehensive understanding and robust extraction of it challenging. Understanding the embedded knowledge in pre-trained language models is a crucial research focus (Petroni et al., 2019, Ettinger, 2020, Pan et al., 2023), as it deepens our understanding of the role of pre-trained language models and how they compare with other knowledge representations, particularly contrasted with structured knowledge graphs, which offer explicit, interpretable, and reliable knowledge. Conversely,

word associations reflect human mental lexica, encoding representations humans derive from their environment. Thus, comparing the relational knowledge between pre-trained language models and word associations can shed light on divergent paradigms of knowledge representations.

Prior work (Petroni et al., 2019) has shown that prompting pre-trained language models in a zero-shot setting (where the model is used as off-the-shelf) is a straightforward approach to extract relational knowledge and evaluate their ability to generalize to unseen tasks (see 2.4.1.3 for more details). However, existing work (Jiang et al., 2020, Ravichander et al., 2020) finds that pre-trained language models are sensitive to the format of prompts in zero-shot prompting, and consequently, the results are fragile. Therefore, achieving robust access to and extraction of relational knowledge remains a challenge.

This drives the primary focus of this chapter: *How to better understand and robustly extract implicit relational knowledge encoded in pre-trained language models? Can word associations contribute to the knowledge extraction?* In this chapter, we shift our attention to the internal relational knowledge encoded in pre-trained language models, using prompts in a zero-shot probing setting. Our specific interests lies in extracting hypernym knowledge, as it forms the foundation for concept categorization and ontology building.

We review research on hypernym extraction from text corpora using surface patterns that represent the relationships between hyponyms and hypernyms (e.g., *Y such as X*). We identified a collection of effective patterns with the potential to address the challenges in current prompt-based approach. To test the efficiency of these patterns, we propose a framework that combines the pattern-based approach and prompt-based approach to identify the presence and absence of hypernym knowledge in PLMs. Using this framework, we: (a) investigate the effects of pattern structure on hypernym extraction (Section 5.6.2); (b) examine challenging scenarios such as abstract, infrequent hyponyms and syntax variations (Sections 5.5 - 5.6.2); and (c) explore the potential of utilising word associations as an external source to improve the quality of prompts (Section 5.7).

This chapter incorporates the following paper:

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2023. Seeking Clozure: Robust Hypernym extraction from BERT with Anchored Prompts. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 193–206, Toronto, Canada. Association for Computational Linguistics.

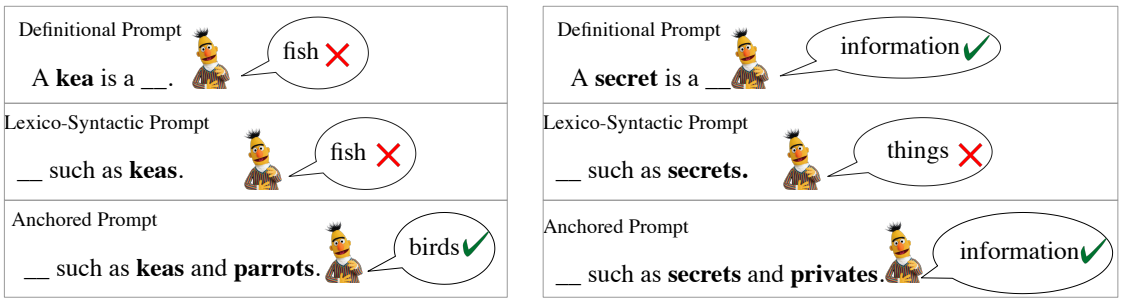
5.1 Introduction

In Section 2.1.1, we established that human memory is fundamentally structured around concepts and their interconnections. The relationships connecting these concepts are important for the organization, retrieval, and understanding of concepts, and enabling reasoning over them. This principle remains important in our model development, particularly with the recent advancements in PLMs. It is essential to evaluate how these models represent concepts and semantic relations in comparison to our existing knowledge. Such an assessment can offer insights into the efficacy and interpretability of these models.

As the backbone of semantic relations, hyponymy/hypernymy relations express a hierarchical relation between a specific concept (the hyponym; e.g., dog) and a general one (the hypernym; e.g., mammal), and form the foundation of human concept understanding (Yu et al., 2015) and relation reasoning (Green et al., 2002, Lyons, 1977). Given the fundamental role of hypernym-hyponym relation, the automatic extraction of hypernym knowledge from large texts (Hearst, 1992, Roller et al., 2018) or pre-trained language models (LMs) (Ravichander et al., 2020, Takeoka et al., 2021, Jain and Espinosa Anke, 2022) is an active area of research (Peters et al., 2019). More background on the evolution of relevant datasets and task formulations can be found in Section 2.4.1.1.

In the task of hypernym extraction, given a hyponym, the objective is to retrieve potential hypernyms either from large corpora using patterns or from PLMs via prompts.[1] The unsupervised extraction of hypernyms from PLMs by prompting with a pattern like *A dog is a type of [MASK]* and retrieving the most likely filler words from the model has been used (Ettinger, 2020, Jain and Espinosa Anke, 2022, Weir et al., 2020). Results were mixed: while

[1] More details on the approach development on hypernym extraction. see Section 2.4.1.3.



(a) infrequent hyponym (b) abstract hyponym

Figure 5.1 Example prompts for infrequent hyponyms (a) and abstract (b) hyponyms for hypernym prediction, derived from established pattern structures.

PLMs can reliably predict hypernyms of concrete and frequent hyponyms (Ettinger, 2020, Weir et al., 2020), experiments on more challenging data sets show a quick deterioration in the face of rare concepts (Schick and Schütze, 2020), and a lack of response consistency across paraphrased prompts (Ravichander et al., 2020, Elazar et al., 2021a). How to alleviate these issues and extract more reliable hypernyms from PLMs remain open questions.

In this work, we draw connections between prompting for hypernyms and pattern-based hypernym extraction (Hearst, 1992, Snow et al., 2004) (see Figure 5.1 and Table 5.1). We conduct a case study using BERT as the representative of masked pre-trained language models. Our choice of masked pre-trained language models like BERT over auto-regressive models such as GPT-2 is aimed at maintaining consistency with prior studies (Petroni et al., 2019, Ettinger, 2020, Ravichander et al., 2020) and ensuring compatibility across patterns.[2] We systematically investigate the utility of different styles of patterns as BERT prompts, and use them as diagnostic tools to better understand the conditions under which probing for hypernyms is effective and consistent. Following this, we explore how large-scale word associations can be used to improve the prompt construction process to acquire better prompts for hypernym extraction from BERT.

Pattern-based hypernym extraction from raw text has a long history, starting from Hearst (1992), which promoted lexico-syntactic patterns (*Y such as X and Z*)[3] as more effective

[2]Specifically, auto-regressive PLMs typically require the hypernym to be the last token, but some patterns position the hypernym before the hyponym, making masked language models more suitable.

[3]We will use Y to denote hypernyms, X for hyponyms and Z for the co-hyponym/anchor of X.

than definitional patterns (*X is a type of Y*) for hyponym-hypernyms extraction from text. Specifically, a follow-up work (Hovy et al., 2009) expanded the utility of the *Y such as X and Z* pattern to extract hypernyms from the Web, reminiscent the task setup of zero-shot prompting PLMs, where only X is given. Their method involved an iterative process: (a) filling the placeholder Z, termed 'anchors', which are sibling words of X and retrieved from the Web, and then (b) retrieving Y from the Web. These anchors (instantiates of Z), provide additional context signals for *X* to extract reliable hypernym *Y*. Figure 5.1 illustrates this, where the anchor *parrot* provides additional information to facilitate the prediction of the correct hypernym of *kea*. This method of iterative 'anchoring' improved the quality and coverage of automatically extracted hypernym knowledge from large corpora. However, the potential of this method in the context of prompting PLMs remains unexplored.

We leverage these established patterns from the hypernym extraction literature in the context of language model prompting, and systematically study the existence and gaps of hyponym-hypernym knowledge in BERT. We conduct experiments on six English data sets and address four questions:

How to effectively construct anchored prompts? We devise a precise and scalable method to automatically retrieve anchors (co-hyponyms) to construct anchored prompts. Anchors are mined from PLMs with well-known co-hyponym patterns (e.g., *such as X and* ___) and evaluated with WordNet (Miller, 1995).

How do different pattern structures compare as prompts under different data conditions? We ground our prompts in hypernym patterns from the pre-neural area which have been applied to raw text successfully, and investigate their effectiveness for zero-shot PLM hypernym retrieval. We find strong and consistent benefits of anchored prompts, especially in the context of rare or abstract concepts.

Robust extraction of hypernym knowledge. Much recent work has shown that PLM prompting results are brittle under prompt paraphrases, calling into question whether prompting surfaces robust knowledge encoded in the PLMs or rather shallow associations. We compare the robustness of different patterns under paraphrasing, and find, again, a benefit of anchored prompts for retrieving more consistently correct hypernyms.

How can large-scale word associations be used to construct effective anchored prompts? Large-scale word associations contain semantic relatedness implicitly and as such can be a resource for retrieving high quality anchors. It is, however, unclear whether and how they can aid the prompting process. We propose to incorporate knowledge from word association networks into the anchor extraction phase to: (a) extract anchors directly; or (b) re-rank the LM anchors. We compare the effects of different anchors selection strategies, and find that anchors selected by combining LM probability and similarity score estimated from word association network yield the best performance on hypernym extraction. This suggests the knowledge complementary between large-scale word associations and BERT.

In essence, we contribute to the on-going research on hypernym extraction by unifying the long-standing work of pattern-based and prompt-based approaches, demonstrating that anchoring prompts can unlock a wealth of hidden knowledge within BERT, especially for challenge scenarios discovered in prior work. Furthermore, we show that large-scale word associations can be used to construct more effective prompts to strengthen deeper and more robust hypernym knowledge extraction from pre-trained language models.

5.2 Background

In Section 2.4.1.3, we provide a detailed review on approaches for hypernym extraction. Here, we elaborate on the two approaches for hypernym extraction that we will use in our study: pattern-based (Section 5.2.1) and prompting PLMs (Section 5.2.2).

5.2.1 Pattern-based Hypernym Extraction

The pattern-based approach applies hyponym-hypernym patterns on large corpora to extract hypernyms. Two widely-used pattern structures have been identified: lexico-syntactic and definitional.

DFP	A(n) X is a Y. A(n) X is a type of Y. A(n) X is a kind Y.	DFP^{+A}	A(n) X or Z is a Y. A(n) X or Z is a type of Y. A(n) X or Z is a kind Y.
LSP	Y such as X. Y, including X. Y, especially X. X or other Y. X and other Y. such Y as X.	LSP^{+A}	Y such as X and Z. Y, including X and Z. Y, especially X and Z. X, Z or other Y. X, Z and other Y. such Y as X and Z.

Table 5.1 Four types of hyponym-hypernym pattern structures: definitional patterns (DFP; top) and lexico-syntactic patterns (LSP; bottom); and their anchored versions: DFP^{+A} and LSP^{+A} (right).

5.2.1.1 Lexico-Syntactic Patterns (LSP)

Lexico-syntactic patterns (LSP; Tab 5.1 bottom left), e.g., *such Y as X*, were first introduced by Hearst (1992) and have since been used to mine hyponym-hypernym pairs or build ontologies from large corpora (Pasca, 2004, Pantel and Pennacchiotti, 2006, Etzioni et al., 2005, Roller et al., 2018). The six LSP (1) all indicate the hyponym-hypernym relation with explicit signals (e.g., *such as, especially*), (2) frequently occur in text, and (3) are applicable to nouns or noun-phrases.

Anchored LSP (LSP^{+A}) Hovy et al. (2009) proposed an 'anchored' version of LSP to mine hypernyms (LSP^{+A})[4], using patterns like *Y such as X and Z*, where Z is an 'anchor' which reduces ambiguity and assists the extraction of Y (Tab 5.1, bottom right).[5] These patterns have been shown to be effective in extracting reliable hypernyms from the text corpora. However, its efficiency has not been studied in the context of extracting knowledge from PLMs. We address this gap by using LSP^{+A} to mine hypernyms from PLMs and examine the benefit of anchored prompts.

[4] LSP^{+A} is referred to as DAP^{-1} in the original paper.

[5] See Section 2.4.1.3 for its detailed discussion on the anchored hypernym extraction patterns with examples.

5.2.1.2 Definitional Patterns (DFP)

In contrast to LSP that conveys the hypernym relation implicitly, definitional patterns (DFP; Tab 5.1 top left), e.g., *A(n) X is a type of Y*, explicitly define an *Is-A* relation between X and Y (Lyons, 1977). A common use of DFP is to mine sentences for definition extraction (Borg et al., 2009, Navigli et al., 2010) or ontology/dictionary building (Muresan and Klavans, 2002). Recently, DFP has been widely used in prompting studies (Schick and Schütze, 2020, Ettinger, 2020, Ravichander et al., 2020, Hanna and Mareček, 2021) to probe hypernym knowledge in PLMs by prompting PLMs with a prefix and using its generation (see Section 5.2.2).

Anchored DFP (DFP^{+A}) Analogous to LSP^{+A}, we augment DFP with anchors for disambiguation (Tab 5.1 top right), e.g., *A(n) X or Z is a type of Y*. To the best of our knowledge, Hanna and Mareček (2021) is the only work which uses anchored definitional patterns (i.e., *A(n) x is a Y. So is a(n) Z.*) to prompt PLMs for hypernyms, described in more detail below.

5.2.2 Prompting-based Hypernym Extraction

Recently, with the advances of PLMs, rich knowledge is captured into models. A stream of research aims at automatically extracting this knowledge, e.g., by probing PLMs for hypernym knowledge (Ettinger, 2020, Weir et al., 2020, Peng et al., 2022). Hanna and Mareček (2021) examined the effectiveness of single hypernym patterns (e.g., *X is a Y*, *Y such as X*) on prompting PLMs and showed that performance varies with patterns. Similarly, Ravichander et al. (2020) found that BERT fails to retrieve consistent knowledge when prompted with singular versus plural variants of the same concept under DFP. For instance, the model produces different hypernyms for the singular *car* prompt, *A car is a* ___, compared to the plural *cars* prompt, *Cars are* ___. These findings motivated our decision to employ a collection of patterns, enhancing the reliability of hypernym extraction. Furthermore, we delve into examining the consistency across these pattern groups in Section 5.6.3.3.

Most previous work on prompting was conducted in relatively simple conditions with one pattern structure and a single data set. We systematically investigate the effects of

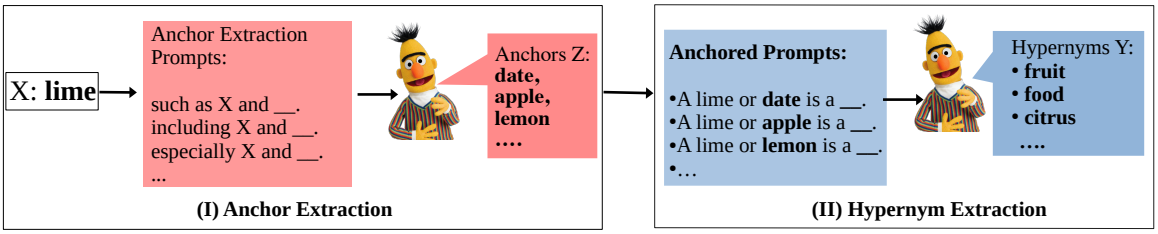


Figure 5.2 The framework of (I) constructing anchored prompts and (II) extracting hypernyms from PLMs.

well-established patterns (LSP/LSP^{+A} and DFP/DFP^{+A}) on extracting hypernyms across six widely-used datasets and paint a more nuanced picture of hypernym knowledge in BERT by explicitly studying the challenging cases of rare or abstract concepts.

5.3 Method: Anchored Prompts

We now introduce our framework of extracting hypernyms from a PLM by constructing sets of prompts given a hyponym X and a pattern type $\in$ {DFP, DFP^{+A}, LSP, LSP^{+A} }. The framework consists of two modules: (I) prompts construction and (II) hypernym extraction . We illustrate the workflow in Figure 5.2, exemplified by LSP^{+A}.

Prompt Construction For each pattern type, we construct a set of prompts by instantiating each of its assigned patterns (cells in Table 5.1) with a concept in positions X and Z, and a [MASK] token in position Y. For DFP and LSP we can construct prompt sets directly given a hyponym X of interest. To construct prompts for LSP^{+A} and DFP^{+A} we need to additionally provide meaningful anchors $\mathcal{Z}$. We propose two approaches to automatically mine such anchors: (a) prompting PLMs, which will be introduced below; and (b) using large-scale word associations (SWOW) as an external resource for anchor mining, which will be introduced in Section 5.7. Next, we describe a way to effectively mine such anchors from language models, as illustrated in Figure 5.2 (I).

Anchor Extraction Given X, we seek to use a PLM to automatically extract a set of anchors $\mathcal{Z}$, i.e., concepts Z that share a hypernym with X. To acquire such anchors, we again

ID	Patterns	ID	Patterns
1	such as X and Z. such as X or Z. such as X, Z,	2	including X and Z. including X or Z. including X, Z,
3	especially X and Z. especially X or Z. especially X, Z,	4	X, Z or other X, Z and other

Table 5.2 Co-hyponym patterns used for anchor extraction. The four groups are adopted from the LSP patterns (Hearst, 1992), which are frequently used in texts to express the co-hyponym relation.

adopt a set of established lexico-syntactic patterns that indicate the fact that X and Z share a common hypernym (Hearst, 1992, Snow et al., 2004, Etzioni et al., 2005). Table 5.2 presents the full list of patterns we used to mine anchors. Each pattern is converted into a prompt by filling in X and replacing Z with a [MASK] token, which is fed into the PLMs, and its top-k outputs are used as anchor candidates. Specifically, we retrieve the 10 most likely filler words according to language model probability for each co-hyponym prompt $\mathrm{C}_i \in \mathrm{C}$. We score candidates z by their average probability across the prompts that contained z among the top 10 fillers:

$$s_{LM}(z \mid x, \mathrm{C}) = \frac{1}{|\mathrm{C}_z|} \sum_{i=1}^{|\mathrm{C}|} P_{LM}(z \mid x, \mathrm{C}_i), \tag{5.1}$$

where $P_{LM}(z \mid x, \mathrm{C})$ is the probability of z in the i^{th} pattern instantiated with x and $|\mathrm{C}_z|$ is the number of patterns that predicted z. We finally keep the M highest scoring concepts as anchors, and instantiate M copies of both $\mathtt{LSP}^{+\mathtt{A}}$ and $\mathtt{DFP}^{+\mathtt{A}}$, each with the different anchors, respectively.

Hypernym Extraction Being able to construct sets of prompts for vanilla ($\mathrm{P}_{\mathtt{DFP}}, \mathrm{P}_{\mathtt{LSP}}$) and anchored prompts ($\mathrm{P}_{\mathtt{DFP}^{+\mathtt{A}}}, \mathrm{P}_{\mathtt{LSP}^{+\mathtt{A}}}$), we are now in a position to prompt PLMs for hypernyms (Figure 5.2 II). Separately for each prompt set $\mathrm{P} \in \{\mathrm{P}_{\mathtt{LSP}}, \mathrm{P}_{\mathtt{DFP}}, \mathrm{P}_{\mathtt{LSP}^{+\mathtt{A}}}, \mathrm{P}_{\mathtt{DFP}^{+\mathtt{A}}}\}$, we score hypernyms candidates y by averaging their log probabilities over all the individual patterns

Dataset	#Hypon	#Hyper	#Pairs	WordNetCov.(%)	Conc.
BLESS (Baroni and Lenci, 2011)	200	85	935	99.8	100 / 91.4
EVAL (Santus et al., 2015)	621	348	953	99.8	88.1/ 83.4
LEDS (Baroni et al., 2012)	1073	364	1262	100	83.7/ 79.2
SHWARTZ (Shwartz et al., 2017)	11061	1101	12724	44.1	66.4/ 92.3
DIAG (Ravichander et al., 2020)	576	9	576	100	97.9/ 100
CSLB (Devereux et al., 2014)	508	232	1079	98.1	100/ 98.2

Table 5.3 The statistics of datasets. WordNetCov. is the coverage of hyponym-hypernym that are connected in WordNet on hypernyms hierarchy. Conc. is the percentage of concrete hyponyms/hypernyms, measured by the concreteness rating from Brysbaert et al. (2014) for the shared vocabulary.

P_i within that prompt set P.

$$s_{LM}(y \mid x, \mathrm{P}) = \frac{1}{|\mathrm{P}|} \sum_{i=1}^{|\mathrm{P}|} \log P_{LM}(y \mid x, \mathrm{P}_i), \tag{5.2}$$

where x denotes a given hyponym and y denotes the predicted hypernym. $\log P_{LM}(y \mid x, \mathrm{P}_i)$ refers to the log probability, as predicted by the BERT model, of y given x. The hypernyms ranked by $s_{LM}(y|x, \mathrm{P})$ and the top K are retained as hypernym candidates.

5.4 Datasets

In this section, we introduce the six datasets used in our study. For consistency and comparability with past work, which has predominantly focused on English in pattern-based studies, all datasets are in English. This alignment also corresponds with the pre-training language of BERT, the model used for our experiments. Now, we introduce the details of these datasets.

- **BLESS** (Baroni and Lenci, 2011) comprises 200 concrete, unambiguous and frequent nouns as hyponyms, with hyponym-hypernym pairs derived from various sources, including WordNet (Fellbaum, 1998), `ConceptNet` (Liu and Singh, 2004), property norms (McRae et al., 2005) and Wikipedia.

- EVALution (**EVAL**; Santus et al. (2015)) includes both concrete and abstract concepts, constructed by combining `ConceptNet` and WordNet with further human judgements to filter out noisy pairs.

- Hyponym-hypernym pairs in **LEDS** (Baroni et al., 2012) are generated by extracting nouns in text corpora and then labeling them with hyponym-hypernym relation using WordNet, provided a chain of hypernym hierarchical relationships exists within WordNet.

- **SHWARTZ** (Shwartz et al., 2016), which features a larger scale of data, is created directly from multiple knowledge bases, including WordNet and other encyclopedic resources (DBPedia Auer et al. (2007), Wikidata Vrandečić (2012) and Yago Suchanek et al. (2007)).

- LM DIAGNOSTIC (**DIAG**; Ravichander et al. (2020)) was recently constructed by retrieving hyponyms from WordNet for nine common superordinate categories (Battig and Montague, 1969), including bird, insect, fish, vehicle, tool, building, tree, flower, and vegetable.

- **CSLB** (Devereux et al., 2014) is a behavioral dataset derived from a property generation task where participants list features for given concepts (see Section 2.1 for more details). These features encompass diverse semantic relations, including possession (HasA) and hypernymy (IsA), as detailed in Table 3.1. In our experiments, we use only the hyponym-hypernym pairs for test.

The first four datasets are widely-used test sets for hypernym extraction more generally (Levy et al., 2015, Roller and Erk, 2016, Shwartz et al., 2017, Roller et al., 2018).[6] The two other datasets, DIAG and and CSLB have been recently used to probe for hypernym knowledge in PLMs (Weir et al., 2020, Ravichander et al., 2020).

Dataset statistics are reported in Table 5.3. We only consider NOUN-NOUN hyponym-hypernym pairs from the datasets to align with the nature of established patterns. We exclude

[6]All four datasets, except for SHWARTZ, have been manually verified by human annotators.

hypernyms that are not included as single tokens in BERT vocabulary.[7] These data sets vary widely in terms of their corpus size, the ratio of abstractness and concreteness, concept frequency and their construction methods, and hence underlying knowledge sources. While most data sets are based on WordNet, SHWARTZ builds on a wider set of resources and includes more obscure concepts. EVAL stands out with a relatively high proportion of abstract concepts, unlike the other data sets which are predominantly concrete. Section 5.6.2 explores performance using these data conditions.

5.5 Experimental Setup

Model All our experiments are based on BERT-large-uncased (Devlin et al., 2019) from Huggingface[8] and use a zero-shot approach to probe the model. Regarding the output vocabulary, prior work (Ravichander et al., 2020) used a *closed-vocabulary* approach, constraining the set of candidates y to hypernyms in a particular data set. However, this approach oversimplifies tasks for datasets like DIAG, which has only nine hypernyms, and demands specific adaptations for each dataset. To allow for comparability of results across data sets, we adopt an *open vocabulary* approach throughout, considering the whole BERT vocabulary as hypernym candidates. We remove test instances where the hypernym is not in the BERT vocabulary as only apply a single `[MASK]` token for each prompt. Note that there is no such restriction on hyponyms, so that results in Section 5.6.3.1 include both seen and unseen hyponyms as single tokens. We set the number of anchors in anchored prompts to $M = 5$, which was optimized on LEDS dataset (see Figure 5.3).

To estimate the upper bound of anchored prompts, we treat siblings from WordNet (Miller, 1995) as oracle anchors and evaluate their effects on hypernym extraction. The detail of retrieving siblings is described in Section 5.6.1. We select top five siblings with the highest rank of their path similarity calculated from WordNet, i.e., $\frac{1}{p(x,z)+1}$, where p is the length of

[7] The ratio of discarded (x, y) pairs is lower than 1% for most datasets, except for BLESS (30% is discarded) and CSLB (17% is discarded). This limitation is further discussed in Section 5.8.

[8] https://huggingface.co/bert-large-uncased

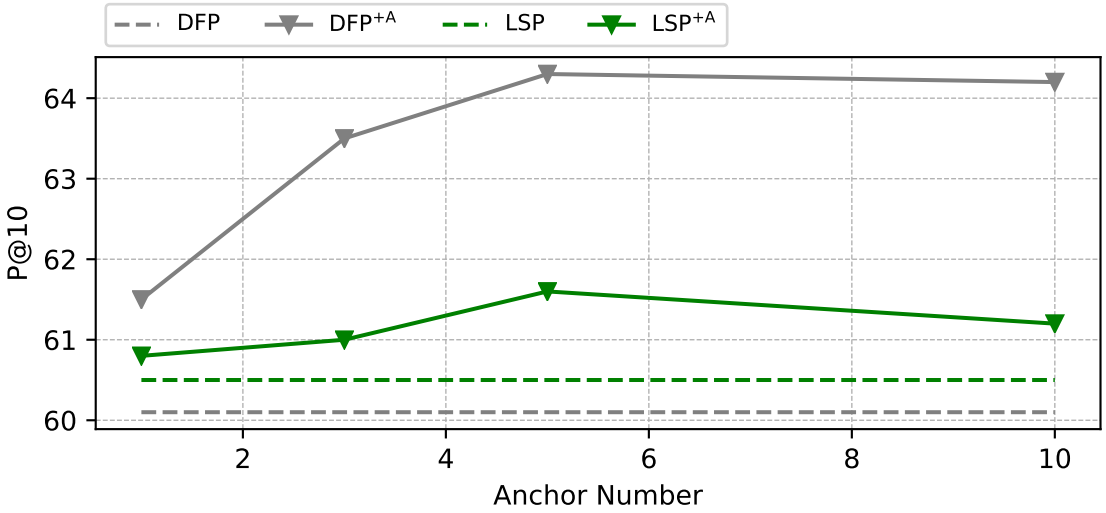


Figure 5.3 Ablation study on the number of anchors for the LEDS dataset. We selected 5 anchors for the remainder of our experiments. Note that the performances of `LSP` and `DFP` are independent of the number of anchors, and these are shown here for reference purposes only.

the shortest path between the x and z among their top two synsets. We use random sampling among siblings with the same score to select up to five anchors.

Evaluation Metrics Following previous work (Petroni et al., 2019, Qin and Eisner, 2021), we retain the $K = 10$ hypernym candidates and report Precision at 10 (P@10) to measure the extent to which correct hypernyms are included in the top 10 model predictions ranked by Equation 5.2. We also report mean reciprocal rank (MRR), which measures the rank of the true label. We evaluate model predictions at the *concept level*, normalizing predictions into their canonical form, i.e., accepting any inflection of the correct hypernym,[9] and exclude punctuation, stop words, numbers and the hyponym x from the predictions. We measure the significance of differences using two-sided t-tests with threshold $p<0.05$.

Analyses In addition to the main results, we aim to understand underlying factors that might affect the performance of hypernym extraction. We analyse the performance of pattern types on different types of concepts. We distinguish sets of hyponyms and hypernyms

[9]We used pyinflect 0.5.1 `https://github.com/bjascob/pyInflect`.

Dataset	MRR	P@1	P@5	P@10
BLESS	73.9	66.0	86.6	89.6
DIAG	34.9	28.6	43.8	48.8
CSLB	60.3	51.2	73.2	77.7
SHWARTZ	23.7	16.8	33.1	39.8
EVAL	33.6	26.1	44.1	49.4
LEDS	45.8	35.7	59.7	66.3

Table 5.4 Anchor evaluation results, where predicted anchors z for a concept x are validated by checking whether x and z share a hypernym in WordNet.

in terms of their **frequency** (Section 5.6.3.1) and **abstractness** (Section 5.6.3.2) and test **consistency** of predictions across prompt paraphrases (Section 5.6.3.3).

5.6 Results

In this section, we present the experimental results on anchor extraction (Section 5.6.1) and hypernym extraction (Section 5.6.2), along with our analyses conducted under various data conditions (Section 5.6.3).

5.6.1 Anchor Validation

How accurate are the automatically mined anchors? We qualitatively and quantitatively inspect retrieved anchor concepts. We leverage WordNet[10] for this purpose, and follow Schick and Schütze (2020) to consider a candidate z to be a valid anchor of x if they share a common ancestor, within two levels above x and four levels above z. We exclude hyponym-hypernyms that cannot be retrieved in WordNet in this analysis.[11]

Table 5.4 reports the results across six datasets using P@K and MRR. For three of the data sets (BLESS, CSLB, LEDS), a correct anchor is predicted as top 1 result more than 35% of the time, and contained among the top 10 predictions close to 70% of the time. This is likely because these datasets have both high coverage in WordNet and contain frequent and

[10]We did not consider ConceptNet (Speer et al., 2017) as the evaluation resource due to its low coverage on co-hyponyms.

[11]Table 5.3 WordNetCov. column presents the coverage.

x	Top 5 predicted anchors ($\mathcal{Z}$)
car	**truck**, **motorcycle**, **boat**, yes, **bike**
apple	**grape**, **pear**, nuts, vegetable, **date**
train	**bus**, plane, **car**, **tram**, **truck**
corn	bean, potato, **barley**, **wheat**, pea
panzer	**tank**, infantry, gun, artillery, panther
motel	hotel, yes, sure, restaurant, actually
daisy	rose, yes, lavender, rush, fern
murre	dog, bird, fox, crow, rabbit
trireme	warship, frigate, ship, ferry, battleship

Table 5.5 Examples of mined anchors ($\mathcal{Z}$) for hyponyms that share ≥ 1 (top) or zero (bottom) co-hyponyms with WordNet. Anchors confirmed in WordNet in bold.

concrete concepts. The other data sets are overall challenging due to diversity and/or low frequency of concepts as described in the Datasets section (Section 5.4).

Table 5.5 presents our qualitative inspection. We noticed a significant overlap between BERT anchors and WordNet siblings for common words (Table 5.5 top). For instance, with *car* as the hyponym, BERT returns anchors like *truck*, *motorcycle*, and *boat*, all of which are listed within WordNet siblings. However, for less common or ambiguous words (Table 5.5 bottom), BERT predicted anchors and WordNet siblings may not align, yet the relevance often remains. For instance, the hyponym *trireme*, defined as *an ancient vessel and a type of galley*, has predicted anchors like *warship*, *frigate*, and *ship*. Though absent from WordNet siblings, these anchors aid in inferring the hypernym *galley*. As we shall see in Section 5.6.2 the utility of anchors does not seem to hinge on them being siblings in WordNet, and that the predicted BERT anchors effectively improve hypernym extraction.

5.6.2 Hypernym Evaluation

Here, we first examine the effectiveness of `LSP` versus `DFP` and the added value of anchoring on our six data sets (Section 5.6.2). Afterwards, we inspect specifically rare (Section 5.6.3.1) and abstract (Section 5.6.3.2) concepts as well as the well-known issue of inconsistency of responses in the face of prompt paraphrases (Section 5.6.3.3), exploring different patterns in these contexts.

	BLESS		DIAG		CSLB		SHWARTZ		EVAL		LEDS	
Pattern	MRR	P@10	MRR	P@10	MRR	P@10	MRR	P@10	MRR	P@10	MRR	P@10
DFP	23.6	42.4	42.6	66.8	39.8	67.5	6.3	12.8	**24.0**	**46.7**	32.6	60.1
DFP^{+A}	25.7^{+}	47.2^{+}	$\mathbf{45.5}^{+}$	$\mathbf{67.2}^{+}$	$\mathbf{42.3}^{+}$	$\mathbf{70.5}^{+}$	5.9^{+}	13.6^{+}	22.1^{+}	43.3^{+}	$\mathbf{35.7}^{+}$	$\mathbf{64.3}^{+}$
$\text{DFP}^{+A}_{\text{oracle}}$ ♯	23.9	41.9	65.4	84.6	41.2	68.2	8.9	15.6	23.3	45.1	37.7	66.2
LSP	**27.1***	**53.9***	**45.5***	66.1	40.8	68.2	6.4	**15.2***	17.3^{*}	39.5^{*}	33.4	60.5
LSP^{+A}	26.5	53.2	42.8^{+}	62.7^{+}	40.4	67.7	**6.5**	14.9	17.0	38.1	34.0	61.6
$\text{LSP}^{+A}_{\text{oracle}}$ ♯	26.2	49.6	65.6	85.8	41.9	68.3	9.1	18.8	18.7	40.5	37.1	66.3

Table 5.6 Main results on six hypernym extraction datasets. Bold number indicates the highest score per data set and metric (except for oracle anchors). * indicates significant difference of LSP vs. DFP; $^{+}$ indicates significant difference wrt. the non-anchored counterpart (i.e., LSP versus LSP^{+A} and DFP vs. DFP^{+A}). $\text{DFP}^{+A}_{\text{oracle}}$ and $\text{LSP}^{+A}_{\text{oracle}}$ use the oracle anchors form WordNet. The ♯ symbol denotes that we report the average over 3 runs on sampled anchors from WordNet.

Main Results Table 5.6 presents the main results. Performance over datasets varies widely, with SHWARTZ standing out with particularly low performance. SHWARTZ is dominated by proper noun hyponyms (e.g., city and person names), and includes a very broad range of hypernyms (1.1K). Performance on the other data sets are more comparable.

Do LSP and DFP differ? Comparing DFP (row one) and LSP (row three) in Table 5.6, we see no consistent trend. While performance is often comparable, on BLESS LSP outperforms DFP, and the reverse is true for EVAL. BLESS contains frequent and largely unambiguous hyponyms which are presumably more frequently discussed in natural patterns as comprised by LSP. EVAL is dominated by ambiguous and abstract concepts, which are perhaps more commonly described by formal, definition-style language.

Do anchors help retrieve more accurate hypernym knowledge? Table 5.6 reveals a consistent improvement of adding anchors for DFP (row 1 vs. 2) but not for LSP (row 3 vs. 4): definitional patterns benefit from anchoring via co-hyponyms while lexico-syntactic patterns don't. Additionally, incorporating oracle anchors from WordNet largely improves the performance on datasets (DIAG and LEDS) that are directly created based on WordNet, while those datasets contain commonsense knowledge (e.g., CSLB) benefit more from BERT anchors. Another observation is that incorporating WordNet anchors hinders the performance on hypernym extraction on BLESS, a dataset consisting solely of unambiguous hyponyms, due to the presence of multiple senses in anchors from WordNet, leading to a mismatch

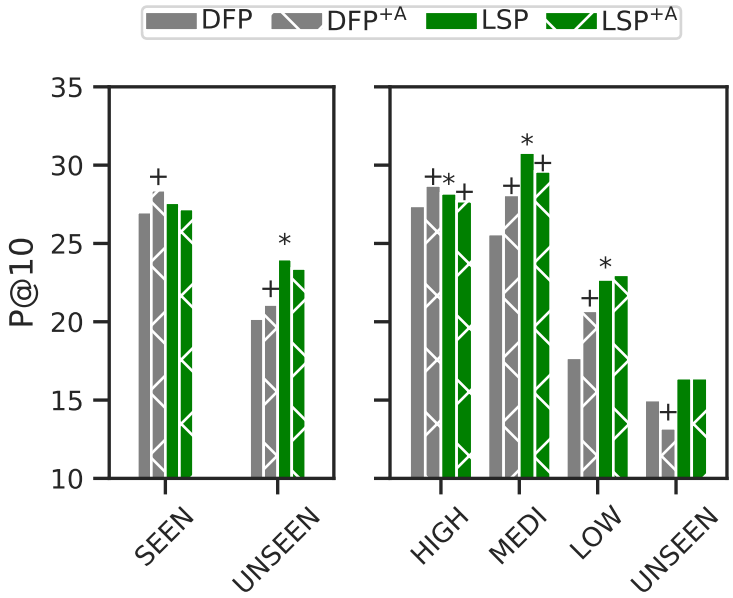


Figure 5.4 Performance of different pattern structures: rare versus common hyponyms. Left: hyponyms seen in BERT vocabulary and not. Right: hyponyms frequency of different frequency bands estimated from large corpora. $^{+}$ and * as in Table 5.6.

between senses. This demonstrates that the quality of anchors is a crucial factor for the hypernym extraction. We delve deeper into improving the quality of anchors in Section 5.7.

5.6.3 Analysis

Next, we disentangle the main results from above, considering a range of conditions which have been identified as challenging in prior work, and examine whether different patterns and/or anchoring can improve hypernym retrieval from BERT in these contexts. Specifically, we analyze the impact of frequency (Section 5.6.3.1), concept abstractness (ssec:abstracenes) and paraphrase consistency Section 5.6.3.1 across patterns and prompts.

5.6.3.1 The Impact of Frequency

Previous work (Ravichander et al., 2020, Hanna and Mareček, 2021, Schick and Schütze, 2020) found that *BERT often fails to predict hypernyms for uncommon hyponyms.* Here, we examine whether incorporating anchors can alleviate this issue. This is driven by the intuition that humans often draw on surrounding context signals to help understand the relationship

between concepts. For example, even if we are unfamiliar with the concept of *kea*, knowing an anchor like *parrot* can help us infer that *bird* is one of the hypernyms. We anticipate that anchors will provide additional linking context between hyponyms and hypernyms, thereby improving hypernym extraction performance, especially when the hyponyms are rare.

To test this, we look into two aspects that reflect frequency. Firstly, we use existence in the BERT vocabulary – hyponyms that are included as single-tokens are frequent.[12] Based on this criterion, we categorize hyponyms into Seen and Unseen. BERT employs the WordPiece tokenizer (Wu et al., 2016), which builds its 30,000-token vocabulary by learning frequent words and subwords in the training data. Thus, common (frequent) words are typically represented as single tokens, whereas rarer words might be divided into multiple subtokens.

Secondly, we obtain term frequency from WorldLex (Gimenes and New, 2016) and examine hyponyms frequency in the corpus. We categorize frequency into four levels based on absolute count: High ($>$ 100), Medium (10-100), Low (1-10), and Unseen (0). Note that our analysis exclusively focuses on hyponym frequency, as hypernyms in our datasets are all frequent, though the methodology is equally applicable for hypernym study.

For this analysis, we aggregate instances from all datasets to increase statistical power. Figure 5.4 presents experimental results. We find that rare hyponyms have lower performance in general, aligning with previous work (Ravichander et al., 2020, Hanna and Mareček, 2021). More interestingly, unlike in the main results, `LSP` exhibits a significant advantage over `DFP` on unseen and low frequency hyponyms (solid bars in UNSEEN and LOW blocks in Figure 5.4). However, in line with the main results, we observe that incorporating anchors only brings improvements for `DFP` but not for `LSP`.

Looking closely, we see that incorporating anchors into `DFP` significantly improves the performance on low frequent hyponyms (solid gray vs. dashed gray). This confirms our hypothesis that anchors are beneficial for uncommon hyponyms by guiding BERT to predict hypernyms. As illustrated in Table 5.7, *dray* is an uncommon hyponym defined as *a low flat vehicle pulled by horses and used in the past for carrying heavy loads, especially barrels of beer*. Without anchors, the `DFP` predictions, such as *boat* and *machine* are

[12] All hypernyms in our test set are single tokens in BERT's vocabulary, as others were excluded.

x	DFP Predictions	DFP^{+A} Predictions	Top 5 predicted anchors ($\mathcal{Z}$)
terebinth	stone, sculpture, rock	**tree**, plant, sculpture	fern, shell, plant, shrub, tree
dray	boat, machine, tool	**vehicle**, cart, wagon	wagon, tractor, cart, horse, yes
gannet	computer, net, network	**bird**, fish, dolphin	seal, dolphin, herr, whale, penguin
swordtail	tail, sword, whip	sword, weapon, **fish**	carp, sword, pike, runner, dragon
tragopan	drum, umbrella, pitcher	**bird**, dog, fish	indian, native, hybrid, relative, dog

Table 5.7 Examples of rare hyponyms x, predicted hypernyms, and predicted anchors. Correct hypernyms are in **bold**.

generated. However, when provided with mined anchors like *wagon* and *tractor*, BERT successfully predicts the accurate hypernym *vehicle*. This finding is of practical importance as it demonstrates that anchored prompts help for uncommon hyponyms, which can be applied to hypernym extraction in domain-specific or low-resources scenarios.

5.6.3.2 The Impact of Concreteness

Previous work on distributional semantics has shown that abstract words have higher contextual variability and are more difficult to predict than concrete concepts (Naumann et al., 2018). Here, we examine whether the degree of concept abstractness of hyponyms and hypernyms affect hypernym extraction accuracy, as well as the impact of different patterns and anchoring in this context.

To obtain the concept concreteness level, we use the Brysbaert dataset (Brysbaert et al., 2014),[13] which covers abstractness ratings for 40K well-known English concepts. Each concept was scored by at least 25 human annotators on a scale from 1 (most abstract) to 5 (most concrete). We use the median score to represent the abstractness of each word and bin them into Abstract (< 3) and Concrete (≥ 3).[14] We inspect all four possible combinations of {concrete, abstract} $\times$ {hypernym, hyponym}, and evaluate the performance using the P@10 metric for each combination. As before, we aggregate instances across all datasets to increase statistical power.

Figure 5.5 shows that hypernyms of hyponyms at same abstraction levels (e.g., Conc–Conc and Abs–Abs, where Conc means Concrete and Abs means Abstract) are predicted with

[13] We exclude hyponyms and hypernyms that are not in the Brysbaert dataset.

[14] We set the categorization threshold at 3, which is the mean score of all concepts in the Brysbaert dataset.

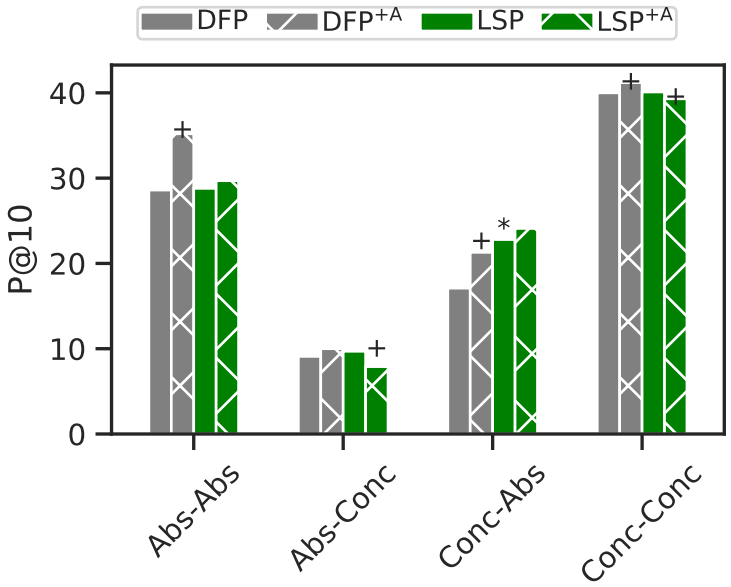


Figure 5.5 Performance of different pattern structures on: abstract hypo- and hypernym (Abs-Abs); abstract hypo- concrete hypernym (Abs-Conc); concrete hypo- abstract hypernym (Conc-Abs); and concrete hypo- and hypernym (Conc-Conc). $^{+}$ and * as in Table 5.6.

higher accuracy than those under different levels (e.g., Abs–Conc). This result is intuitive as words in same abstraction level tend to co-occur more (Bhaskar et al., 2017, Frassinelli et al., 2017). Overall, concrete hyponym-hypernym pairs are predicted with higher accuracy than pairs involving an abstract concept, indicating that abstract knowledge is more difficult to retrieve from BERT.

In terms of anchor benefits, we observe enhanced performance exclusively in the case of DFP, with no consistent improvement noted for LSP. Moreover, we find that DFP^{+A} brings substantial improvements on abstract hypernyms, effectively reducing the gap between abstract and concrete hypernyms. A closer look at abstract hypernyms that failed with DFP but succeed on anchored prompts reveals failure on abstract hypernyms such as {*emotion, organization, language*}. For example, for the prompt *excitement is a* ___ BERT predicts {*thrill, fear, rush*}. However, by incorporating anchors like *surprise* into the prompt *excitement or surprise is a* ___, BERT predicts the correct hypernym *emotion*. More examples are presented in Table 5.8. This finding is encouraging because it points to the weakness of using hyponyms alone to prompt PLMs for abstract hypernyms and can potentially inform

x	DFP Predictions	DFP^{+A} Predictions	Top 3 predicted anchors ($\mathcal{Z}$)
happiness	joy, life, pleasure	joy, feeling, **emotion**	love, joy, good
principle	rule, law, concept	rule, law, **value**	practice, rule, procedure
snoopy	toy, pigeon, mouse	toy, puppet, **character**	peanut, snoop, batman
excitement	thrill, rush, fear,	feeling, **emotion**, fear,	surprise, love, fear
wrath	curse, vengeance, demon	demon, **emotion**, curse	human, self, john,
anger	rage, fury, **emotion**	**emotion**, reaction, feel	fear, hurt, hatred
military	army, soldier, **organization**	**organization**, alliance, navy	army, alliance, institution
kokborok	elephant, indian, island	**language**, indian, christian	indian, chinese, burmese
crusade	adventure, invasion, army	**event**, invasion, campaign	campaign, war, invasion,

Table 5.8 Examples of abstract hyponyms x (top) and abstract hypernyms (bottom), along with predicted hypernyms; and predicted anchors. Correct hypernyms are in **bold**.

future work on prompt design for retrieving specific types of knowledge (e.g., concrete or abstract) and building ontologies.

5.6.3.3 Consistency

Despite the success of prompting, a persistent challenge is an inconsistency of responses under rephrasing of the prompt (Elazar et al., 2021a). In the context of hypernm prediction, Ravichander et al. (2020) specifically showed that compared to singular prompts (*a car is a* ___.), plural versions (*cars are* ___.) returned different (worse) results. This indicates that BERT is sensitive to slight alterations in surface text and fails to robustly retrieve concept-level knowledge. We study consistency more systematically by including different paraphrases, and exploring the utility of anchoring on the robustness of results. We investigate: (a) consistency across prompts paraphrased with singular and plural hyponyms; and (b) consistency over prompts paraphrased with four pattern type instantiations, thus DFP, DFP^{+A}, LSP and LSP^{+A} (see cells in Table 5.9). We only score the prediction for a test instance as correct, if it was correctly predicted by *all* prompt paraphrases.

Pairwise Number Consistency Following Ravichander et al. (2020), we construct pairs of prompts for singular and plural hyponyms, obtaining one representative pair for each of our for pattern types (listed in Table 5.10, left). The results in Table 5.10 show that consistency very strongly correlates with the choice of patterns: DFP prompts lead to highly inconsistent

DFP	A(n) X is a Y. A(n) X is a type of Y. A(n) X is a kind Y.	DFP^{+A}	A(n) X or Z is a Y. A(n) X or Z is a type of Y. A(n) X or Z is a kind Y.
LSP	Y such as X. Y, including X. Y, especially X. X or other Y. X and other Y. such Y as X.	LSP^{+A}	Y such as X and Z. Y, including X and Z. Y, especially X and Z. X, Z or other Y. X, Z and other Y. such Y as X and Z.

Table 5.9 Replicating Table 5.1 for convenience. Four types of hyponym-hypernym pattern structures: DFP (top), LSP (bottom), and their anchored versions: DFP^{+A} and LSP^{+A} (right). We regard the patterns within each cell as paraphrases of one another and employ them to investigate group consistency across four types.

	Singular Probes	Plural Probes	BLESS	DIAG	CSLB	SHWARTZ	EVAL	LEDS
DFP	A(n) X is a(n) Y.	X are Y.	2.7	4.5	3.5	0.4	1.7	4.8
DFP^{+A}	A(n) X or Z is a(n) Y.	X or Z are Y.	0.2	2.3	0.3	0.0^{+}	0.1	0.6^{+}
LSP	Y such as a(n) X.	Y such as X.	**51.2**	46.0	60.9	4.4	26.2	40.6
LSP^{+A}	Y such as a(n) X or Z.	Y such as X or Z.	**51.2**	**51.6**$^{+}$	**65.0**$^{+}$	**10.4**$^{+}$	**32.5**$^{+}$	**52.5**$^{+}$

Table 5.10 Experimental results (P@10) on pairwise number consistency. X/Z in singular probes are instantiated as singular (e.g., car), and in plural probes as plural (e.g., cars). The $^{+}$ notation, as defined in Table 5.6, indicates a significant difference between the anchored and non-anchored counterparts.

results, while LSP shows strong potential for retrieving consistent knowledge. One reason is ambiguity in the plural DFP: the prompt *Xs are [MASK]* tends to return verbs and adjectives as candidates (e.g., *carrots are {grown, eaten, cultivated}.*), as plausible completions. Contexts are much more restrictive in LSP. Moreover, the consistency improves significantly for all but one data set when incorporating the anchors into LSP.[15] This finding is important as it identifies a promising means of retrieving consistent knowledge from PLMs.

Group Consistency Our sets of pattern-type specific prompts suggest a natural, stricter consistency evaluation, namely to test whether BERT reliably predicts the same, true hypernym for all prompts associated with a pattern type (i.e, each of the cells of Table 5.9). Table 5.11 presents the results. What stands out in the table is that anchored prompts signifi-

[15] Indeed, when comparing against the less strict evaluation in Table 5.6, LSP^{+A} incurs the smallest performance drop.

	BLESS	DIAG	CSLB	SHWARTZ	EVAL	LEDS
DFP	21.9	42.5	44.7	4.6	23.7	34.3
DFP^{+A}	**31.7**$^{+}$	**49.0**	**53.8**$^{+}$	**8.3**$^{+}$	**28.3**$^{+}$	**42.2**$^{+}$
DFP^{+A}	47.2	67.2	70.5	13.6	43.3	64.3
LSP	26.8	32.8	45.8	2.6	10.2	29.0
LSP^{+A}	**31.7**$^{+}$	**39.9**$^{+}$	**52.5**$^{+}$	**4.7**$^{+}$	**13.3**$^{+}$	**36.1**$^{+}$
LSP^{+A}	53.2	62.7	67.7	14.9	38.1	61.6

Table 5.11 Experimental results (P@10) on group consistency. $^{+}$ as in Table 5.6. We include aggregated results across all patterns within a type for reference, highlighted in gray rows, taken from Table 5.6.

cantly improve group consistency, which aligns with our observation in the pairwise number consistency tests above. In summary, our results show that anchors, in particular DFP^{+A}, can help retrieve more robust and consistent hypernyms from PLMs. This is not only important for downstream tasks which rely on (automatic) high-quality hypernym knowledge, such as taxonomy creation, but could also inform strategies to probe BERT for genuine, systematic hypernymy knowledge, rather than superficial associations.

In addition to the positive results, we also recognize that the quality of anchors is a crucial factor in constructing effective prompts. We observe that some anchors remain noisy, as they often contain irrelevant common words like *yes* and *actually* or predict overly general terms like *things* and *objects*. As a result, we aim to enhance anchor quality by leveraging information from word associations, which are more likely to yield meaningful associative words that humans produce. Next, we will describe how we utilize word associations to improve the quality of anchors.

5.7 Improving Anchor Quality with Word Associations

Large-scale word associations networks (WAN) encode rich relational structure, which can be used as a source of providing anchors. Although relation types in WAN are not explicitly provided, our study in Chapter 3 shows that they implicitly encode various semantic relationships, which aligns with prior work (Kolers, 1963, Rosenzweig, 1961, McRae et al., 2012). Importantly, we found that relations most similar to co-hyponyms, specifically

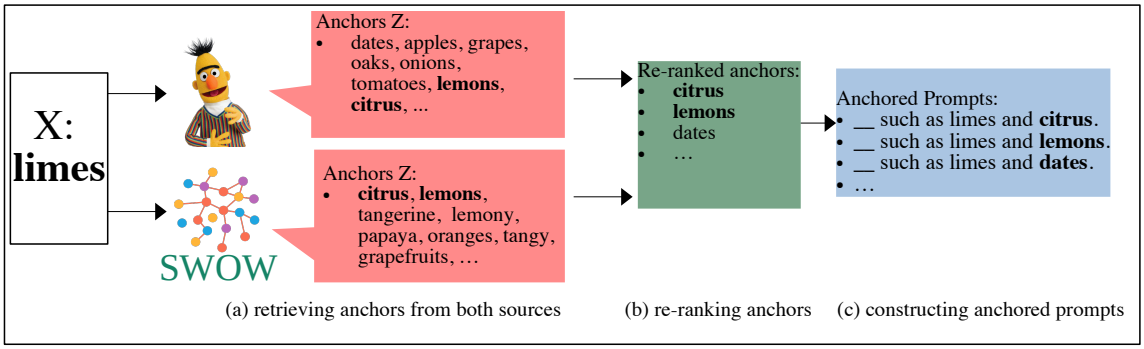


Figure 5.6 The framework of constructing anchored prompts with the incorporation of word associations.

SYNONYM and SAMECATEGORY, constitute 24.7% of our WAX labeled subset. Motivated by this, we hypothesise that associations can be a valuable source for selecting anchors. The core idea is to utilise SWOW as an external source to improve the quality of anchors by retrieving anchors or re-ranking anchors from BERT. However, how to select relevant anchors from the diverse associations is a challenge as they are not labelled. In our study, we propose heuristic rules to filter noisy anchors. In addition to direct cue-association links in WAN, we also consider similar words, which often share a considerable number of associations, as another source of acquiring anchors. Figure 5.6 (a) illustrates the framework of incorporating SWOW in the procedure of constructing anchored prompts.

We next describe the selection and incorporation of the two types of knowledge into the anchor extraction module (Section 5.7.1). We then conduct experiments comparing the benefits of using anchors from WAN with those from PLMS (see Section 5.3) in Sections 5.7.2-5.7.4.

5.7.1 Method: Anchor Extraction

We introduce two types of anchors that can be extracted from word associations: mutual associations and similar words. Mutual associations reveal bidirectional eliciting relationships between words, whereas similar words assist in pinpointing semantically akin words. We hypothesise that both provide the contextual relevance of information for input hyponyms to extract more accurate hypernyms from PLMS.

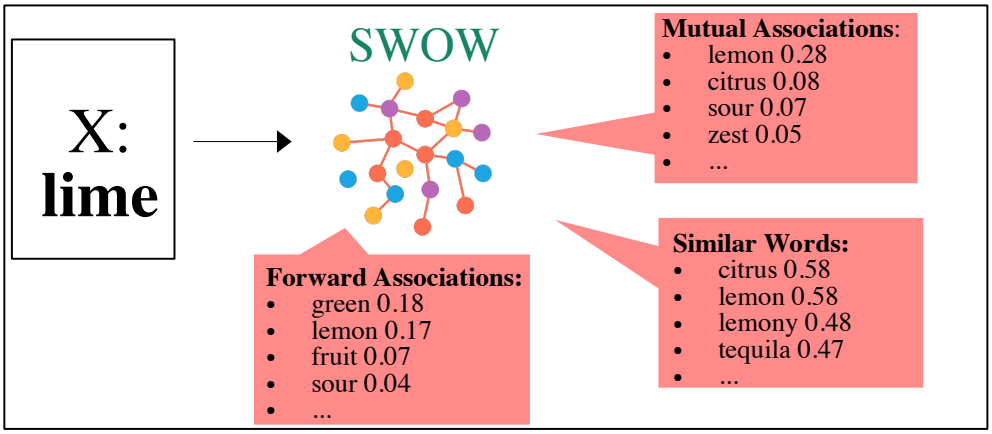


Figure 5.7 Example of forward associations (bottom), mutual associations (top right), and similar words (bottom right) in SWOW. The numbers represent strengths: forward associations use the random walk measure, mutual associations use Equation (5.3), and similarity scores are based on the cosine scores from Equation (5.6).

Mutual associations in WAN WAN consists of a basic set of associative directions, namely *forward associations* from a cue to a response (Figure 5.7 bottom), and *mutual associations* for pairs where the reverse is also included in the WAN (Figure 5.7 top right). Forward associations tend to be varied due to the open-ended nature of word associations, whereas mutual associations are more constrained since they signify words that elicit each other. These words often co-occur in the same context or share properties, making them more likely to be anchors. In our study, we choose candidate anchors from those mutual associations. We posit that two words with strong mutual associations (e.g., *lime* $\leftrightarrow$ *sour*), implying that links exist in both directions, are likely to be co-hyponyms (e.g., Figure 5.7 top). To measure how strongly two words x and z are mutually associated, we use the total association strength from both directions to rank:

$$s_{\text{WA}_{\text{bir}}}(z,x) = p(z \mid x) + p(x \mid z), \tag{5.3}$$

where $p(z \mid x)$ and $p(x \mid z)$ denote the association strength from x to z and the reverse. To estimate association strength between two words in a WAN, we use the decaying random walk measure (Griffiths et al., 2007a). In this approach, when a word (node) in a graph is activated, it activates its immediate neighbors iteratively, creating a ripple effect. Consequently, words

connected by shorter and a greater number of paths are considered more similar. The random walk has been demonstrated to outperform other measurements, such as raw response frequency or positive pointwise mutual information (PPMI) that normalizing the frequency of a response (De Deyne et al., 2016b, 2019). Unlike other measures that only consider the direct responses, it captures the distribution overlap of all direct and indirect neighbours of cues, which can be formulated as follows:

$$G_{\text{rw}} = (\mathbb{I} - \alpha P)^{-1} \tag{5.4}$$

where $\mathbb{I}$ is the identity matrix and α is the decaying parameter in random walk for controlling the importance of short paths and long paths. P is the associative matrix based on PPMI calculated from the response frequency. For a specific cue word c and its associated response a, the PPMI is defined as:

$$PPMI(a \mid c) = \max\left(0, \log_2\left(\frac{p(a \mid c)}{p(a)}\right)\right), \tag{5.5}$$

where $p(a \mid c)$ indicates the relative frequency of a response when presented with a specific cue word, while $p(a)$ signifies the overall frequency of that response across all cue words. The G_{rw} is normalized by row to sum to 1, resulting in a distribution $p(a \mid c)$ which measures the association strength between a and c. After calculating the G_{rw} matrix, we use it to determine the association strengths as specified in Equation (5.3). For instance, the $p(z \mid x)$ in Equation (5.3) corresponds to the association strength between words x and z, where x is treated as the cue word c and z as the response a.

After obtaining the anchor candidates along with their scores via eq (5.3), we consider two uses: (a) the direct use, which involves selecting the top K anchors after ranking by using $s_{\text{WA}_{\text{bir}}}(z,x)$ as mutual association scores; and (b) providing an auxiliary score function to adjust anchor scores from s_{LM}.

Similar Words in WAN Word associations have been shown to be an effective source for estimating word similarity (Rensbergen et al., 2016, Buades-Sitjar et al., 2021). Two words are similar if they share high proportion of responses or paths in a word association

network (Deese, 1966, De Deyne et al., 2019), thus are likely to share common properties (e.g., Figure 5.7 bottom). We hypothesise that similar words in WAN are co-hyponyms and thus serve as anchors. The semantic similarity between two words can be estimated by cosine distance between two word vectors obtained from the above random walk measure (De Deyne et al., 2016d, 2019) as follows:

$$s_{\text{WA}_{\text{sim}}}(z \mid x) = \frac{\sum_{i=1}^{N} p(a_i \mid x) p(a_i \mid z)}{\sqrt{\sum_{i=1}^{N} p(a_i \mid x)^2} \sqrt{\sum_{i=1}^{N} p(a_i \mid z)^2}}, \tag{5.6}$$

where $s_{\text{WA}_{\text{sim}}}(z \mid x) \in [0, 1]$ is cosine distance between x and z, reflecting the how similar two words are in a WAN. The index i ranges from 1 to N, with N being the total number of cue words in SWOW. $p(a_i \mid x)$ and $p(a_i \mid z)$ are the probability of association word a_i connected to x and z respectively, which are obtained via a random walk in the WAN (c.f., eq (5.4)).

Analogously to mutual associations, we consider two usages of $s_{\text{WA}_{\text{sim}}}$: (a) using it alone to identify anchor candidates; and (b) incorporating the score together with LM scores to improve the anchor quality, as described below.

Anchors from WAN and PLMs As seen in Section 5.6.1, anchors obtained from BERT can be noisy, including irrelevant or meaningless words (e.g., *yes*), while those derived from WAN may be more focused on content words, yet more diverse in the relations they express. We hypothesise that the two approaches are complementary, and thus integrating anchors from both sources can improve the quality of anchors. We propose two methods to combine both sources for acquiring anchors: (a) by intersecting their candidates, i.e., $\mathcal{Z}_{\text{LM}}(x) \cap \mathcal{Z}_{\text{WA}}(x)$; (b) by aggregating their scores, i.e., $s_{\text{LM}}(z \mid x) + s_{\text{WA}}(z \mid x)$. The first approach is expected to obtain reliable and high quality anchors, however, it might suffer sparsity issues when there is no overlap between top K LM candidates and associations. In cases with no shared anchors, we default to using BERT anchors. The second method potentially overcomes this limitation by ranking anchors based on combined scores from both sources, offering a more comprehensive approach.

5.7.2 Dataset

We use SWOW (De Deyne et al., 2019), the largest English word association network,[16] to obtain the candidates of mutual associations and similar words as anchors. We construct the test set by unifying the datasets we used in Section 5.6.2 and filter instances whose hyponyms do not occur as cues in SWOW and WordNet, resulting in 2K out of 4.8 K[17] test instances. Note that the results in this section cannot be directly compared to those in Section 5.6 due to differences in the test sets.

5.7.3 Experimental Setup

We apply the proposed framework (Section 5.3) to construct anchored prompts for hypernym extraction and evaluate anchors with WordNet as Section 5.6.1. For the pattern structure of anchored patterns, we decide to use the best-performing anchor structure: $\texttt{DFP}^{+\texttt{A}}$. Apart from anchors from SWOW, we include anchors from WordNet as an alternative source. Each anchor has a score $s_{\text{WN}} = \frac{1}{p(x,z)+1}$, where p is the shortest path length between the hyponym and the anchor. We use s_{WN} the same way as s_{WA} (cf., 5.7.1). In each scenario, we choose top 5 anchors based their scores. We randomly break ties and report the average results across three runs. Note that we consider anchors from WordNet as oracle anchors as they are extracted by following WordNet hypernym hierarchy.

5.7.4 Results

5.7.4.1 Anchor Validation

Table 5.12 shows the results of anchor validation on WordNet with various anchor selection strategies. Compared to s_{LM}, anchors from SWOW alone have less coverage in WordNet due to diverse nature and the limited number of associations. The performance of taking the intersection lies between two individual sources, which is reasonable as the number of anchors are reduced under the strict conditions. Adding scores from SWOW boosts the

[16] See more background information about SWOW in Section 2.2.2.1.
[17] We exclude the SHWARTZ dataset due to its high portion of named entities.

Anchor Selection	MRR	P@1	P@5	P@10
$s_{\text{WA}_{\text{bir}}}$	19.1	9.9	31.4	41.5
$s_{\text{WA}_{\text{sim}}}$	24.3	13.9	39.7	48.3
s_{LM}	52.7	44.6	64.0	68.9
$\mathcal{Z}_{\text{LM}} \cap \mathcal{Z}_{\text{WA}_{\text{bir}}}$	46.3	41.1	53.1	55.2
$\mathcal{Z}_{\text{LM}} \cap \mathcal{Z}_{\text{WA}_{\text{sim}}}$	48.3	42.3	55.6	58.6
$s_{\text{LM}} + s_{\text{WA}_{\text{bir}}}$	54.2	45.4	66.8	70.6
$s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$	**55.3**	**47.0**	**67.7**	**72.4**

Table 5.12 Anchor evaluation results across 2K test instances, where predicted anchors z for a concept x are validated by checking whether x and z share a hypernym in WordNet. Values starting with $\mathcal{Z}$ in Anchor Selection refer to anchors that are sampled from the intersection of two sources.

Pattern	**Anchor Selection**	MRR	P@10
DFP	-	23.2	46.8
DFP$^{+\text{A}}$	$s_{\text{WA}_{\text{bir}}}$	24.1	49.3
DFP$^{+\text{A}}$	$s_{\text{WA}_{\text{sim}}}$	24.4	49.0
DFP$^{+\text{A}}$	s_{LM}	24.5	49.1
DFP$^{+\text{A}}$	$s_{\text{WN}}^{\sharp}$	24.3	47.8
DFP$^{+\text{A}}$	$\mathcal{Z}_{\text{LM}} \cap \mathcal{Z}_{\text{WA}_{\text{bir}}}$	24.3	49.8
DFP$^{+\text{A}}$	$\mathcal{Z}_{\text{LM}} \cap \mathcal{Z}_{\text{WA}_{\text{sim}}}$	24.5	50.1‡
DFP$^{+\text{A}}$	$\mathcal{Z}_{\text{LM}} \cap \mathcal{Z}_{\text{WN}}$	24.7	49.6
DFP$^{+\text{A}}$	$s_{\text{LM}} + s_{\text{WA}_{\text{bir}}}$	25.1‡	50.0‡
DFP$^{+\text{A}}$	$s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$	**25.2**‡	**50.2**‡
DFP$^{+\text{A}}$	$s_{\text{LM}} + s_{\text{WN}}$	25.3	50.3

Table 5.13 Experimental results of hypernym extraction, augmented with word association in anchor extraction. Bold number indicates the highest score per metric, except for anchors from WordNet, which are considered as oracle. ‡ indicates adding scores from SWOW bring significant difference compared to the pure LM anchors. The symbol $^{\sharp}$ indicates the average of three runs on randomly selected top k anchors when their scores are the tied.

precision, indicating that including similarity scores raises the ranks of anchors that occur in both sources. Furthermore, combining similarity scores from SWOW and a LM yields the best performance, suggesting a significant portion of "similar words" exist in word associations are co-hyponyms.

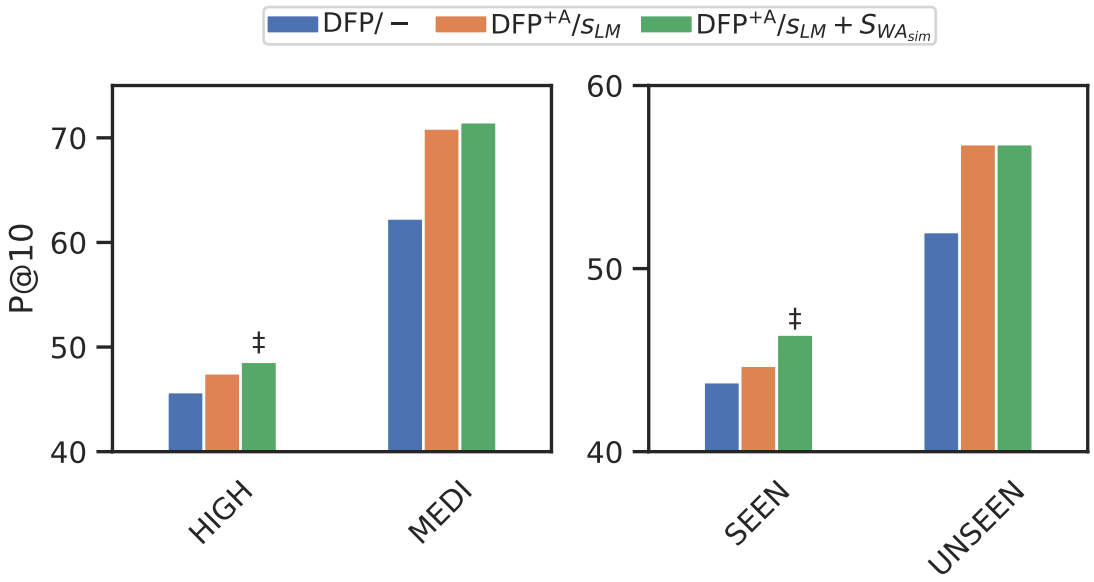


Figure 5.8 Performance of different anchor sources, varying with (a) frequency of hyponyms and (b) existence in BERT vocabulary. ‡ as in Table 5.13.

5.7.4.2 Hypernym Evaluation

Table 5.13 presents the experimental results on hypernym extraction. Using similar words from word associations alone achieves comparable improvements as automatically extracted anchors (i.e., $s_{\text{WA}_{\text{sim}}}$ versus s_{LM}), although Table 5.13 shows that anchors from SWOW alone have less overlap with co-hyponyms in WordNet. This indicates that word associations alone are a valuable resource for acquiring anchors. Importantly, adding SWOW into anchor selection boosts the performance in general, especially with word similarity scores from SWOW ($s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$), suggesting that SWOW is complementary to BERT. Interestingly, anchors from SWOW and WordNet achieve comparable results, suggesting SWOW as a valuable resource for obtaining anchors. The overall significance of this finding is meaningful, as it broadens the utility of word associations for retrieving more accurate hypernyms from PLMs.

5.7.4.3 Analysis

To assess the impact of anchor quality on different data conditions, we conduct the same analysis as Section 5.6.2 on the best performing model ($s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$), considering the

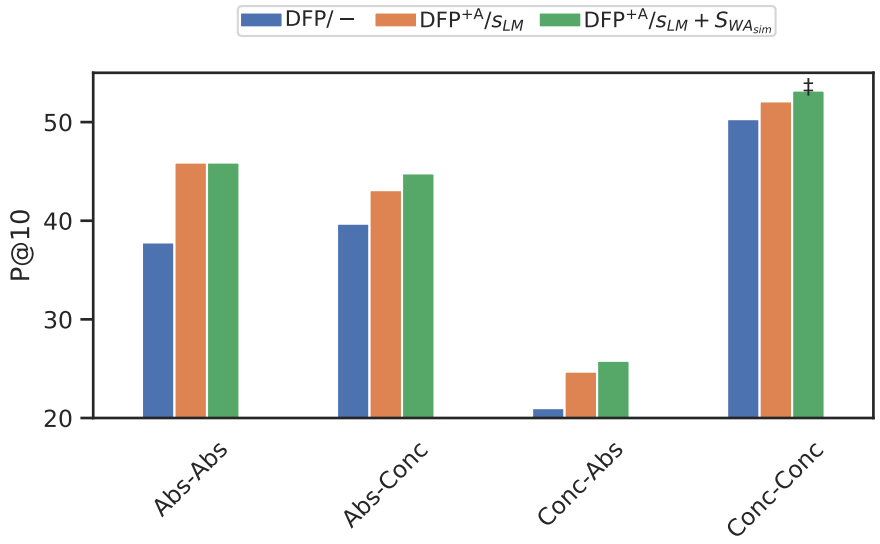


Figure 5.9 Performance of different anchor sources on the concreteness of hyponym-hypernyms, ranging from no anchors (left blue bar) to anchors from pure LM (middle orange bar) and to anchors combining LM and SWOW similarity scores (right green bar). ‡ as in Table 5.13.

frequency (see Figure 5.8),[18] concreteness (Figure 5.9) and consistency (Table 5.14). We find that incorporating word similarity from SWOW significantly enhances performance for the two dominant groups: highly frequent hyponyms (representing 92% of test instances) and concrete-concrete groups (representing for 83% of test instances), but no discernible effects on minor groups such as medium frequent and abstract-abstract groups. Additionally, we find that adding SWOW knowledge further improved the consistency of retrieved hypernyms, including both pairwise probe and group probe.

To illustrate the impact of incorporating word similarity, Table 5.15 presents examples of how anchors from s_{LM} are refined by adding $s_{\text{WA}_{\text{sim}}}$. For instance, for the hyponym *beetle*, the most relevant LM anchor is *moth*. However, after adding $s_{\text{WA}_{\text{sim}}}$, more relevant anchors like *bug* and *spider* are obtained, leading to the successful extraction the true hypernym *creature*. These examples show how adding SWOW knowledge can improve the quality of anchors and retrieved hypernyms.

[18]We exclude groups from our analysis if they have less than two instances, including low-frequency and unseen groups.

Anchor Scorer	PairConsistency	GroupConsistency
-	34.2	27.0
s_{LM}	42.5	33.2
$s_{\text{LM}} \cap s_{\text{WA}_{\text{bir}}}$	42.3	34.0
$s_{\text{LM}} \cap s_{\text{WA}_{\text{sim}}}$	43.4‡	**34.8**
$s_{\text{LM}} + s_{\text{WA}_{\text{bir}}}$	42.7‡	34.7
$s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$	**44.4**‡	**34.8**‡

Table 5.14 Experimental results (P@10) on pair and group consistency probe with anchors selected by various criteria. The optimal pattern structure is used for PairConcistency (definitional patterns) and GroupConsistency (lexico-syntactic patterns), respectively.

x | y_{true}	s_{LM} Anchors	s_{WA} Anchors	s_{LM} + $s_{\text{WA}_{\text{sim}}}$ Anchors
rabbit | creature	fox, deer, squirrels, mouse, rodent	bunny, hare, breeding, rodent, breed	**hare**, rodent, mouse, **mammal**, **dog**
scarf | clothing	tie, hat, coat, glove, maybe	shawl, knit, sweater, necktie, knitting	**sweater**, **wool**, hat, coat, glove
beetle | creature	fly, chip, moth, bark, ball	ladybug, insect, bug, ladybird, centipede	**insect**, **bug**, fly, **spider**, **slug**
axe | implement	sword, knife, spear, hammer, shovel	hatchet, chainsaw, lumberjack, machete, pickax	knife, sword, **saw**, **blade**, spear
onion | food	garlic, tomato, yes, pepper, potato	garlic, pickle, potato, tomato, leek	garlic, tomato, potato, **vegetable**, **vinegar**

Table 5.15 Positive examples where the use of s_{LM} + $s_{\text{WA}_{\text{sim}}}$ anchors successfully extract y_{true} from BERT. The $s_{\text{WA}_{\text{sim}}}$ effectively adjust the s_{LM} anchors. Bolded words in the last column highlight these adjustments.

Despite these improvements, not all instances benefit from the incorporation of SWOW scores. Through a manual inspection of these instances, particularly the abstract-abstract hyponym-hypernym pairs, we find some limitations. We discover that 40% of the hypernyms are incorrectly predicted in all systems, regardless of the existence and source of anchors (e.g., BERT always fails to predict the hypernym of *emotion* for the hyponym *thrill*), which we attribute to the limited capacity of the model. Additionally, the overlap between the top LM anchors and the top SWOW anchors for abstract hyponyms is low. As a result, incorporating SWOW similarities has a minimal impact on the ranking of LM anchors. See examples in Table 5.16 (bottom) for illustration.

Conversely, for concrete hyponyms, the anchor candidates exhibit a higher overlap, indicating a closer alignment between BERT's representation of concrete concepts and the word associations from SWOW. For example, for the hyponym *trumpet*, shared anchors include *trombone*, *saxophone*, and *clarinet*. However, this overlap does not necessarily yield

x | y_{true}	s_{LM} Anchors	s_{WA} Anchors	$s_{\text{LM}} + s_{\text{WA}_{\text{sim}}}$ Anchors
bowl | artifact	**cup**, pitcher, goal, spoon, **plate**	crockery, **plate**, dish, **cup**	**cup**, spoon, **plate**, dish, pot
sparrow | creature	crow, hawk, owl, **robin**, eagle	finch, bird, songbird, blackbird, **robin**	finch, **robin**, bird, lark, crow
pear | food	**apple**, plum, **peach**, cherry, grape	nectarine, **apple**, juicy, fruit, **peach**	**apple**, **peach**, plum, fruit, ripe
tiger | beast	**lion**, leopard, liam, carly, sean	**lion**, cougar, panther, puma, jaguar	**lion**, leopard, panther, jaguar, bear
pine | evergreen	oak, **cedar**, maple, **fir**, spruce	**fir**, evergreen, redwood, **cedar**, Forrest	oak, **fir**, **cedar**, redwood, tree
saxophone | device	flute, **clarinet**, piano, trumpet, guitar	sax, trombone, baritone, **clarinet**, tuba	flute, **clarinet**, trombone, tuba, oboe
trumpet | artifact	**trombone**, piano, **saxophone**, guitar, **clarinet**	**trombone**, tuba, **saxophone**, **clarinet**, bassoon	**trombone**, **saxophone**, tuba, **clarinet**, oboe
therapy | care	treatment, drug, medication, pharmaceutical, medicine	psychiatric, therapist, psychiatrist, psychiatry, psychotherapist	treatment, drug, medication, pharmaceutical, medicine
thrill | emotion	sexual, **excitement**, yes, real, food	thriller, **excitement**, suspense, adrenaline, elated	sexual, **excitement**, yes, real, food
analogy | inference	stereotype, metaphor, parallel, quote, reference	simile, metaphor, comparative, similarity, synonym	stereotype, metaphor, parallel, quote, reference
agreement | statement	contract, treaty, convention, obligation, negotiation	agreed, agree, pact, accord, treaty	contract, treaty, convention, obligation, negotiation
pride | satisfaction	yes, legend, horn, daughter, clan	proud, boastful, egotistical, boast, hubris	yes, legend, horn, daughter, clan

Table 5.16 Illustration of two scenarios where s_{WA} anchors *cannot* effectively adjust s_{LM} anchors: (1) concrete hyponym-hypernyms exhibit certain overlap between top five LM and WA anchors (top); (2) abstract hyponym-hypernyms have less overlap between top five LM and WA anchors (bottom). Bold anchors occur in three types of anchors.

additional benefits, hinting at a potential redundancy in the information provided (see more examples in Table 5.16 top).

Moreover, we observe an unexpected trend where the test performance is higher for the less frequent groups, with MEDI outperforming HIGH and UNSEEN surpassing SEEN. This outcome appears to contradict with our findings in Section 5.6.3.1. The underlying cause of this discrepancy can be traced to the composition of the high-frequency group, which has a significantly higher percentage of abstract concepts, manifesting either as hyponyms or hypernyms. For instance, 15.7% of all test pairs (N=325) in the HIGH group contain abstract concepts, against only 0.8% (N=17) in the MEDI group. As Figure 5.9 illustrates, pairs with abstract concepts have lower performance than those with only concrete terms, clarifying why the HIGH group lags behind.

In summary, we find that incorporating prior knowledge from SWOW improve the quality of anchors and performance on hypernym extraction, extending the utility of large-scale word associations for prompting accurate and robust hypernyms. However, our findings reveal that BERT's representation of concrete concepts aligns more closely with word associations,

whereas there exists a discrepancy with abstract concepts, pointing to a future direction for further study.

5.8 Limitations and Discussion

Our proposed framework aims to integrate hypernym patterns that have been successfully applied on raw text into pre-trained language model prompting. We show that incorporating anchors leads to significant improvements on the task of hypernym extraction, particularly in challenging scenarios where anchors were previously absent. However, there are several limitations that should be considered when using our framework, and we point to interesting opportunities for extension.

Effectiveness beyond noun-noun concepts We apply our method to hyponym-hypernym pairs over nouns. This idea of anchored prompts can also be extended to mine hypernyms for named entities (Pantel and Ravichandran, 2004, Pasca, 2004) or other parts-of-speech (verbs; Chklovski and Pantel (2004)) using patterns developed for text corpora.

Extending to Other Semantic Relations We apply the concept of 'anchoring' to extract hypernyms from pre-trained language models. This methodology is adaptable to other semantic relations explored in corpus mining. For example, Girju et al. (2003) investigated the PART-WHOLE relation, identifying lexico-syntactic patterns such as *X is part of Y*. These can be expanded to include anchored versions, like *X or Z is part of Y*. Additionally, Kozareva and Hovy (2010) applied anchored patterns to the casual relation with patterns like *X and Z cause Y*. However, adapting this approach of mining and incorporating anchors might require further refinement.

Approaches to Anchor Incorporation In our study, we combined anchors with the original variable X to prompt for Y. An alternative strategy might involve replacing X with another variable, Z, in order to maintain a consistent pattern throughout. To illustrate, consider the pattern *X is a part of Y*, which often signifies the PART-WHOLE relationship between X and

Y. Notably, this relation is observed more frequently than the pattern *X or Z is a part of Y*. Using prompts like *"An engine is a part of a [MASK]"*, and substituting *engine* with anchors (*wheel*, *bonnet*, *tire*), helps deduce that `[MASK]` refers to a car. However, determining the most effective method for selecting accurate `[MASK]` predictions based on anchors requires further investigation.

Hypernym diversity Current work on extracting hypernyms with BERT predominantly considers single-word hypernyms and does not consider multi-word hypernyms or hypernyms that are not in the BERT vocabulary. Our work is no exception. Exploring autoregressive language models like GPT-4 (OpenAI, 2023) or Llama 2 (Touvron et al., 2023) could mitigate this issue.

Scale of Language Models We focus on comparing different pattern structures with a single model, BERT-large. A recent work (Shani and Vreeken, 2023) has shown that GPT-based models have a better understanding of hypernyms than BERT. The behaviour of anchored patterns under larger language models, such as GPT-4 (OpenAI, 2023), remains to be examined.

5.9 Summary

In this chapter, we explored the effective extraction of hypernym knowledge from BERT using anchored prompts and the utilisation of large-scale word associations. We proposed a framework of unifying two powerful techniques in hypernym extraction: the pattern-based and prompt-based approach. Using this framework, we conducted a thorough investigation on multiple factors important for extracting accurate and robust hypernym knowledge from BERT, including pattern structures, various data conditions, and incorporation of external sources. Our comprehensive analysis highlights the efficacy of anchored definitional patterns, especially in addressing challenges such as rare hyponyms and abstract hypernyms. Furthermore, anchored prompts demonstrate a significant increase in the reliability of retrieved hypernyms under paraphrased prompts.

We demonstrated that more improvements are observed by incorporating large-scale word associations as an external source to improve the quality of anchors. We effectively mined and scored anchor candidates from a large-scale word association network (SWOW), either filtering anchors retrieved from BERT or serving as a supplementary scoring mechanism. Our results show that the two anchor sources, namely BERT and SWOW, obtain comparable improvements when used alone, while the combination of both yields the best performance. In line with our earlier experiments on commonsense question answering (Chapter 3), this again suggests knowledge encoded in pre-trained language models is complementary to human associations and confirms our hypothesis that word associations can help construct effective prompts for hypernym extraction. Our findings can direct future work on prompt design to extract robust and consistent hypernym knowledge.

In the next chapter, we will conclude this thesis and point out several promising directions in the line of understanding large-scale word associations and utilising it to advance NLP tasks.

Chapter 6

Conclusions

Commonsense knowledge is crucial for imbuing AI systems with a commonsense reasoning ability. However, the acquisition of this knowledge has been a long-standing challenge in NLP due to its implicit and vast nature (Davis and Marcus, 2015).

In this thesis, we turned our attention to the technique of 'free word association', a tool prevalent in cognitive psychology for studying human memory. We explored word associations as a novel means of acquiring commonsense knowledge. These associations, reflecting spontaneous connections among words, stem from the intuitive thinking system of humans (Kahneman, 2012). They naturally emerge in the human minds when presented with cue concepts, thereby effortlessly revealing the implicit connections among concepts. Large-scale word associations like SWOW (De Deyne et al., 2019) offer valuable insights into human understanding of concepts, grounded in experience. However, the significance of this resource has not been fully recognised in NLP. This is partly due to the scarcity of datasets and understanding behind these associations. Consequently, their utility could be unknown from a different filed. This thesis uncovers the underlying relationships in word associations, exploring their use as a commonsense knowledge graph to improve commonsense question answering and knowledge extraction. Furthermore, the thesis highlights the importance of cross-disciplinary research between cognitive psychology and NLP, emphasising the potential for mutual advancement in both fields.

In this chapter, we revisit the research questions proposed in Section 1.1 with findings and implications (Section 6.1), and present the future directions (Section 6.2).

6.1 Research Question Revisited

Question 1: *What relational knowledge is encoded in word associations?*

We investigated the rationales and relationships underlying word associations in Chapter 3. We introduced a two-stage collection framework to collect a large-scale word association explanation dataset (WAX), comprised of sentences that explain why two words are associated, and relation labels derived from these explanations to reveal high-level structures. Our analysis of relations involved both top-down and bottom-up approaches. In the top-down analysis, we examined human-labelled relation labels, uncovering a predominance of semantic relations in word associations. In the bottom-up analysis, we clustered the provided explanations, highlighting the diverse and complex reasons for associations, and found significant overlap with the relation ontology identified in our top-down analysis. This is important as our study demonstrates that reasons in word associations are explainable and can be explicitly labelled with semantic relations. Notably, these relations can be mapped to relations in existing commonsense knowledge graph `ConceptNet`, providing evidence that word associations encode commonsense knowledge.

We introduce the task of word association relation classification, using our relation ontology as a ground truth. This task involved fine-tuning pre-trained language models to predict types of relations between associated word pairs. To test our hypothesis that explanations supply ample context signal for models to recognise relations, we compared the classification of relations with and without explanations. We found that supplying explanations significantly improved model performance, underlining the effectiveness of explanations in disambiguating relations. In our analysis of performance across various relations, we discerned that while models adeptly predicted some relation types, such as taxonomic (synonyms and antonyms), they struggled with context-dependent relations and

eventuality relations. This is likely due to the models' inadequacy in capturing the complexity of these relations, a direction for future improvement.

Our WAX dataset represents the first large-scale resource that contextualises word associations within human memory. Beyond this, our study demonstrates the feasibility of automating relation labelling for word associations with NLP models, contributing significantly to both cognitive psychology and NLP. This offers a valuable resource and opens new avenues for deeper understanding of human cognition and for probing the capabilities and limitations of NLP models.

Question 2: *Can large-scale word associations improve performance on downstream commonsense reasoning tasks? How does the knowledge they encode differ from the existing largest commonsense knowledge graph* `ConceptNet`*?*

We explored the potential of large-scale word associations (`SWOW`) as a commonsense knowledge resource, comparing them with `ConceptNet` from multiple perspectives in Chapter 4. Our analysis highlighted differences in their respective graph structures and knowledge content. Using a human-curated dataset of explicit situational knowledge, we examined the situational commonsense knowledge embedded within both graphs and found that `SWOW` encodes this knowledge more directly. We incorporated `SWOW` into four knowledge-augmented models for commonsense question answering and compared the resulting improvements to those achieved when using `ConceptNet` as an external commonsense knowledge graph. Our results demonstrated that despite the lack of explicitly labelled relations, `SWOW` achieved similar improvements to `ConceptNet` across three datasets, thereby confirming our hypothesis that word associations encode valuable commonsense knowledge.

This study is the first to successfully show the practical use of large-scale word associations for commonsense tasks. It introduces a novel method of incorporating commonsense knowledge from human 'free word associations' into complex modelling. This is particularly beneficial for both cognitive psychology and NLP communities, as it offers a unique intersection of understanding of knowledge encoded in human associations and how they can improve model's commonsense reasoning ability.

Question 3: *How to better understand and robustly extract implicit relational knowledge encoded in pre-trained language models? Can word associations contribute to the knowledge extraction?*

As PLMs continue to advance, understanding and extracting their embedded knowledge becomes crucial, especially in contrast to existing resources. This study specifically investigates (a) effective approaches for extracting hypernyms from BERT, and (b) the role of large-scale word associations in enhancing our ability to extract hypernyms. We integrated a pattern-based approach from corpus mining with the recent prompt-based approach to examine various patterns under different data conditions, including abstractness, frequency, and robustness. Our findings revealed that "anchoring patterns", which utilize the input word along with accompanying sibling words (anchors), significantly improved performance in challenging scenarios such as abstract and rare concepts. Furthermore, we proposed using large-scale word associations to acquire anchors and compared them against automatically mined anchors from BERT and WordNet. We found that combining anchors from word associations with automatically mined anchors yielded the best performance on hypernym extraction in terms of accuracy and robustness.

This study connects traditional pattern-based approaches with modern prompt-based methods, showing their combined use enhances knowledge extraction effectiveness and robustness in PLMs. Moreover, our findings on word associations enhancing hypernym extraction highlight the complementary nature of this knowledge to BERT, implying divergences in concept representation between human behavioural data and PLMs. This identifies knowledge gaps in PLMs and suggests directions for their improvement.

6.2 Future Directions

6.2.1 Labelling Word Association Relations at Scale

In Chapter 3, we found that relationships encoded in word associations are diverse and context-dependent. We show that using explanations to ground the relation between cue-association pairs can improve models' relation classification ability. However, it is costly

to acquire a large number of human-written explanations for all cue-association pairs in SWOW, causing the difficulty of labelling the entire SWOW based on explanations. One potential solution is to use large language models, such as ChatGPT (OpenAI, 2022) or GPT4 (OpenAI, 2023), to automatically generate explanations. A recent work (Wan et al., 2023) also demonstrates that using GPT3 (Brown et al., 2020) to generate explanations between concepts can improve the relation classification performance in general domain. To capture the diverse relationships under various context, hundreds of explanations for each cue-association pair can be generated. WAX can serve as a tool to evaluate whether models form the associations in a similar or different reasons as humans. This would allow us to capture the relation distribution for each cue-associations and identify the most probable multiple relations between each pair of cue-associations.

Another promising direction for future research is the further development of the relation ontology. For the lack of large-scale dataset to recognise the relations among words, the relation ontology is typically defined by researchers (Fitzpatrick, 2006, Fitzpatrick and Thwaites, 2020) manually, including our study in Chapter 3. To ease the process of relation labelling for human annotators in our current work, we adopted a highly condensed relation ontology consisting of 16 relations. This was achieved by initially combining relations from multiple relevant ontologies, and then merging similar relations and association directions. This decision was informed by observations from our preliminary experiments and the subjective judgments of our researchers. Our relation ontology forms a basis for future research, however, the determination of an optimal relation ontology for word associations remains an open question. With the viability of human-generated explanations in WAX, a promising opportunity arises for enhancing the ontology development process. An intriguing avenue for future exploration could involve utilizing these explanations to automatically learn the relation ontology, thereby streamlining the process.

6.2.2 Identifying the Boundary Between Human and Model Associations

In this thesis, we focus on understanding human word associations and their integration into models. Another research line focuses on automatic extraction of word associations from various sources, including large-corpora (Griffiths et al., 2007b, Lin et al., 2019b, Hu et al., 2020) and pre-trained language models (A. Rodriguez and Merlo, 2020, Yao et al., 2022). However, a significant gap still exists between human and model associations.

Recent advancements in large language models have demonstrated emerging capabilities (Wei et al., 2022). These models exhibit remarkable on human cognition tasks such as the theory of mind (OpenAI, 2023). Can they simulate human-like cognitive associations? This is where large-scale human word associations, like SWOW, can play a pivotal role. Serving as evaluation tools, they can be used to examine the ability of these models to generate human-like associations. Specifically, they can aid in evaluating models' capability in terms of various aspects, such as the association strength, relation types, and nuanced meanings that a single word can invoke in different contexts. For example, how does a model associate words in an abstract versus concrete scenario? Or how does it handle frequent versus rare words? Can those generated associations capture perceptual or affective information that human captured based on perceptions from multi-modalities? By comparing the outputs from these evaluations with human cognition, we can further identify and quantify the disparities between human and text-based machine cognition, providing insightful directions for future research.

6.2.3 Cross-lingual and Multilingual Word Associations

In this thesis, we focus on the English word associations. However, it is worth noting that SWOW is a multilingual project, covering 18 languages, each collected independently from native speakers.

One natural extension of our study is to explore non-English SWOW datasets. Applying the WAX collection and evaluation framework from Chapter 3, which focuses on unveiling

the latent association relation structures, to other languages, could potentially illuminate both shared and unique association patterns across languages. Evaluating the impact of these non-English word associations on downstream tasks, akin to our English SWOW study in Chapter 4 and 5, could further enrich our comprehension of cross-lingual semantics and their utility in diverse NLP tasks.

Another noteworthy avenue for future research is the integration of multilingual SWOW. Current multilingual knowledge graphs, especially in the commonsense domain, are quite limited. Some multilingual knowledge graphs such as BabelNet (Navigli and Ponzetto, 2012) and ConceptNet 5.5 (Speer et al., 2017) exist, but these resources are predominantly English-centric, causing under-representation of other languages, whose source are often acquired based on text corpora. In contrast, SWOW collects associations directly from native speakers across all its languages, thereby capturing knowledge beyond what traditional text corpora offer for non-English languages. Integrating these diverse associations into a unified knowledge graph could enable a more effective knowledge transfer across languages, proving invaluable for tasks such as machine translation, cross-lingual information retrieval, and language learning.

However, integrating multilingual word associations into a unified knowledge graph involves navigating several complex challenges. The volume of data varies greatly across languages, leading to potential representation bias, i.e., high-resource languages such as English will be over-represented and low-resource languages will be under-represented. Languages are deeply intertwined with culture, possessing unique semantic variances of associations that direct translations may not always capture the full meaning (Lim et al., 2022). Especially those concepts involving nuances, idiomatic expressions, and cultural references, will cause difficulty when during translation. Aligning different associations for the same concept due to cultural, historical, or social reasons adds another layer of complexity. The process of harmonizing these diverse datasets into a unified knowledge graph demands innovative solutions, including effective cross-lingual mappings, cultural context integration, and robust handling of semantic variances across languages.

The ‘free word association’ game, simple in concept yet profound in impact, opens doors to understanding memories, cultural biases, subconscious links, and cognitive patterns. With the development of both cognitive psychology and NLP models, we are now at an exciting point in exploring the world of associations in both humans and models more deeply.

References

M. A. Rodriguez and P. Merlo. 2020. Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 376–385.

L. von Ahn, M. Kedia, and M. Blum. 2006. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 75–78.

J. Aitchison. 1994. *Words in the mind : an introduction to the mental lexicon*, 2nd edition. Wiley-Blackwell.

A. Akbik, D. Blythe, and R. Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

B. AlKhamissi, M. Li, A. Celikyilmaz, M. T. Diab, and M. Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

X. Amatriain. 2023. Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*.

J. Anacleto, H. Lieberman, M. Tsutsumi, V. Neris, A. Carvalho, J. Espinosa, M. Godoi, and S. Zem-Mascarenhas. 2006. Can common sense uncover cultural differences in computer applications? In *Artificial Intelligence in Theory and Practice: IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream*, pages 1–10.

J. R. Anderson. 1995. *Learning and memory: An integrated approach.* John Wiley & Sons.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. DBpedia: A nucleus for a web of open data. In *ISWC'07/ASWC'07: Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, pages 722–735.

J. L. Ba, J. R. Kiros, and G. E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

L. Baldini Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

M. E. Bales and S. B. Johnson. 2006. Graph theoretic modeling of large-scale semantic networks. *Journal of biomedical informatics*, 39(4):451–464.

D. A. Balota and J. H. Coane. 2008. Semantic memory. In *Learning and Memory: A Comprehensive Reference*, pages 511–534. Elsevier.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

R. Bansal, M. Aggarwal, S. Bhatia, J. Kaur, and B. Krishnamurthy. 2022. CoSe-Co: Text conditioned generative CommonSense contextualizer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1128–1143.

Y. Bar-Hillel. 1960. The present status of automatic translation of languages. In Franz L. Alt, editor, *Advances in Computers*, volume 1, pages 91–163. Elsevier.

M. Baroni, R. Bernardi, N.-Q. Do, and C.-c. Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32.

M. Baroni and A. Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.

M. Baroni, B. Murphy, E. Barbu, and M. Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34 2:222–54.

L. W. Barsalou. 1983. Ad hoc categories. *Memory & Cognition*, 11(3):211–227.

W. F. Battig and W. E. Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, 80:1.

L. Bauer, Y. Wang, and M. Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230.

G. Bernier-Colborne and C. Barrière. 2018. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 725–731.

M. Bevilacqua, R. Blloshmi, and R. Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 35(14):12564–12573.

P. Bhargava and V. Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, volume 36, pages 12317–12325.

S. A. Bhaskar, M. Köper, S. Schulte Im Walde, and D. Frassinelli. 2017. Exploring multi-modal Text+Image models to distinguish between abstract and concrete nouns. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He. 2023. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

M. Bolognesi, R. Pilgram, and R. V. den Heerik. 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 49:1984–2001.

R. Bommasani, K. Davis, and C. Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

F. Bond and R. Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

F. Bond and K. Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

G. Bordea, P. Buitelaar, S. Faralli, and R. Navigli. 2015. SemEval-2015 task 17: Taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910.

G. Bordea, E. Lefever, and P. Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, pages 2787–2795.

C. Borg, M. Rosner, and G. Pace. 2009. Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 26–32.

A. Bosselut, R. Le Bras, and Y. Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, volume 35, pages 4923–4931.

A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

G. H. Bower. 2000. A Brief History of Memory Research. In *The Oxford Handbook of Memory*, pages 3–32. Oxford University Press.

J. Breen. 2004. JMdict: a Japanese-multilingual dictionary. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 65–72.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

M. Brysbaert, A. B. Warriner, and V. Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

F. Buades-Sitjar, C. Planchuelo, and A. Duñabeitia. 2021. Valence, arousal and concreteness mediate word association. *Psicothema*, 33 4:602–609.

Á. Cabana, C. Zugarramurdi, J. C. Valle-Lisboa, and S. De Deyne. 2023. The "Small World of Words" free association norms for Rioplatense Spanish. *Behavior Research Methods*, pages 1–18.

J. Camacho-Collados. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. *arXiv preprint arXiv:1703.04178*.

J. Camacho-Collados, C. Delli Bovi, L. Espinosa-Anke, S. Oramas, T. Pasini, E. Santus, V. Shwartz, R. Navigli, and H. Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724.

E. Cambria, Y. Song, H. Wang, and A. Hussain. 2011. Isanette: A common and common sense knowledge base for opinion mining. *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 315–322.

H. d. M. Caseli, B. A. Sugiyama, and J. C. A. Silva. 2010. Using common sense to generate culturally contextualized machine translation. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 24–31.

N. Chambers and D. Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. T. Koehn, and T. Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *15th Annual Conference of the International Speech Communication Association*, pages 2635–2639.

J. Chen, J. Chen, and Z. Yu. 2019. Incorporating structured commonsense knowledge in story completion. In *Proceedings of Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, volume 33, pages 6244–6251.

Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417.

Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

T. Chklovski and P. Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 33–40.

H. H. Clark. 1970. Word associations and linguistic theory. *J. Lyons (Ed.), New horizons in linguistics*, 3:271–286.

L. Clouatre, P. Trempe, A. Zouaq, and S. Chandar. 2021. MLMLM: Link prediction with mean likelihood masked language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4321–4331.

A. M. Collins and E. F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407 – 428.

A. M. Collins and M. R. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247.

P. Cramer. 1968. *Word association*, first edition. Academic Press.

G. S. Cree and K. McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163–201.

M. Cremer, D. Dingshoff, M. de Beer, and R. Schoonen. 2011. Do word associations assess word knowledge? a comparison of l1 and l2, child and adult word associations. *International Journal of Bilingualism*, 15(2):187–204.

A. M. Dai and Q. V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3079–3087.

E. Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys*, 56(4).

E. Davis and G. Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

J. Davison, J. Feldman, and A. Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178.

D. Daza, M. Cochez, and P. Groth. 2021. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, pages 798–808.

S. De Deyne, Y. N. Kenett, D. Anaki, and M. Faust. 2016a. Large-scale network representations of semantics in the mental lexicon. In *Big data in cognitive science*, pages 183–189. Psychology Press.

S. De Deyne, D. Navarro, and G. Storms. 2013a. Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

S. De Deyne, D. J. Navarro, A. Perfors, and G. Storms. 2016b. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology. General*, 145 9:1228–54.

S. De Deyne, D. J. Navarro, and G. Storms. 2013b. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45:480–498.

S. De Deyne, D. J. Navarro, G. Collell, and A. Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922.

S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

S. De Deyne, A. Perfors, and D. J. Navarro. 2016c. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

S. De Deyne, A. Perfors, and D. J. Navarro. 2016d. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

S. De Deyne and G. Storms. 2008a. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

S. De Deyne and G. Storms. 2008b. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40:198–205.

F. De Saussure et al. 1916. Nature of the linguistic sign. *Course in general linguistics*, 1:65–70.

J. Deese. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.

J. Deese. 1966. *The structure of associations in language and thought*. The Johns Hopkins University Press.

B. Devereux, L. K. Tyler, J. Geertzen, and B. Randall. 2014. The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods*, 46:1119 – 1127.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

P. Dognin, I. Melnyk, I. Padhi, C. Nogueira dos Santos, and P. Das. 2020. DualTKB: A Dual Learning Bridge between Text and Knowledge Base. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8605–8616.

Y. Du, Y. Wu, and M. Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6133–6143.

H. Dubossarsky, S. De Deyne, and T. T. Hills. 2017. Quantifying the structure of free association networks across the life span. *Developmental psychology*, 53(8):1560.

N. Durrani, H. Sajjad, F. Dalvi, and Y. Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880.

N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285.

Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg. 2021a. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Y. Elazar, H. Zhang, Y. Goldberg, and D. Roth. 2021b. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500.

D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(19):625–660.

A. Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

E. A. Feigenbaum. 1984. Knowledge engineering. the applied side of artificial intelligence. *Annals of the New York Academy of Sciences*, 426:91–107.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

C. Fellbaum. 2010. Wordnet. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer.

C. Fellbaum, M. Alkhalifa, W. Black, S. Elkateb, A. Pease, H. Rodriguez, and P. Vossen. 2006. Introducing the Arabic WordNet project. In *Proceedings of the 3rd Global Wordnet Conference*.

Y. Feng, X. Chen, B. Y. Lin, P. Wang, J. Yan, and X. Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

J. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

T. Fitzpatrick. 2006. Habits and rabbits: word associations and the L2 lexicon. *Eurosla Yearbook*, 6:121–145.

T. Fitzpatrick and P. A. Thwaites. 2020. Word association research and the L2 lexicon. *Language Teaching*, 53:237–274.

D. Frassinelli, D. Naumann, J. Utt, and S. Schulte m Walde. 2017. Contextual characteristics of concrete and abstract words. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

L. Frermann, I. Titov, and M. Pinkal. 2014. A hierarchical bayesian model for unsupervised induction of script knowledge. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–57.

P. Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

C. Gardent, A. Shimorina, S. Narayan, and L. Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

P. Garrard, M. A. Ralph, J. R. Hodges, and K. Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, 18(2):125–74.

M. Gimenes and B. New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48:963–972.

C. Girju. 2002. *Text mining for semantic relations*. ProQuest Information and Learning.

R. Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83.

R. Girju, A. Badulescu, and D. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 80–87.

R. Girju, A. Badulescu, and D. Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18.

A. Gladkova, A. Drozd, and S. Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

J. Goñi, G. Arrondo, J. Sepulcre, I. n. Martincorena, N. Vélez de Mendizábal, B. Corominas-Murtra, B. Bejarano, S. Ardanza-Trevijano, H. Peraita, D. P. Wall, and P. Villoslada. 2011. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive Processing*, 12(2):183–196.

A. Gordon, Z. Kozareva, and M. Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398.

J. Gordon and B. V. Durme. 2013. Reporting bias and knowledge acquisition. In *In Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

J. Gordon, B. V. Durme, and L. K. Schubert. 2010. Learning from the web: Extracting general world knowledge from noisy text. In *Proceedings of the AAAI 2010 Workshop*

on Collaboratively-built Knowledge Sources and Artificial Intelligence (WikiAI 2010), volume WS-10-02, page 10–15.

M. Gósy and M. Kovács. 2002. The mental lexicon: Results of some word association experiments. *Acta Linguistica Hungarica*, 49:179–224.

R. Green, C. A. Bean, and S. H. Myaeng. 2002. *The Semantics of Relationships: An Interdisciplinary Perspective*. Kluwer Academic Publishers.

H. P. Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

T. L. Griffiths, M. Steyvers, and A. Firl. 2007a. Google and the mind: Predicting fluency with pagerank. *Psychological Science*, 18(12):1069–1076.

T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. 2007b. Topics in semantic representation. *Psychological Review*, 114 2:211–44.

L. Guerrero, A. Claret, W. Verbeke, G. Enderli, S. Zakowska-Biemans, F. Vanhonacker, S. Issanchou, M. Sajdakowska, B. S. Granli, L. Scalvedi, et al. 2010. Perception of traditional food products in six european regions using free word association. *Food quality and preference*, 21(2):225–233.

D. Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, J. Zhou, and M. Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758.

M. Hanna and D. Mareček. 2021. Analyzing BERT's knowledge of hypernymy via prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282.

S. Hao, B. Tan, K. Tang, B. Ni, X. Shao, H. Zhang, E. Xing, and Z. Hu. 2023. BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5000–5015.

Z. S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, pages 539–545.

M. A. Hearst. 1998. Automated discovery of wordnet relations. *WordNet: an electronic lexical database*, 2.

I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99.

I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

D. Hendrycks and K. Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

J. D. Hollan. 1975. Features and semantic memory: Set-theoretic or network model? *Psychological Review*, 82(2):154–155.

P. Hosseini, D. A. Broniatowski, and M. Diab. 2022. Knowledge-augmented language models for cause-effect relation classification. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 43–48.

E. Hovy, Z. Kozareva, and E. Riloff. 2009. Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 948–957.

J. Howard and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Z. Hu, J. Luo, C. Zhang, and W. Li. 2020. A natural language process-based framework for automatic association word extraction. *IEEE Access*, 8:1986–1997.

P.-L. Huguet Cabot and R. Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

K. A. Hutchison, D. A. Balota, J. H. Neely, M. J. Cortese, E. R. Cohen-Shikora, C.-S. Tse, M. J. Yap, J. J. Bengson, D. Niemeyer, and E. Buchanan. 2013. The semantic priming project. *Behavior research methods*, 45:1099–1114.

J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi. 2021. (Comet-) ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 6384–6392.

F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. A. Szekely. 2021. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347.

D. Jain and L. Espinosa Anke. 2022. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 151–156.

S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. Yu. 2022. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33:494–514.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

M. N. Jones and D. J. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, 114(1):1.

M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

T. Joyce. 2005. Constructing a large-scale database of japanese word associations. *Glottometrics*, 10:82–99.

D. Jurafsky and J. H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd edition. Prentice Hall PTR.

J. Juraska, P. Karagiannis, K. Bowden, and M. Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162.

D. Kahneman. 2012. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

M. Kaneko and D. Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

P. Ke, H. Ji, Y. Ran, X. Cui, L. Wang, L. Song, X. Zhu, and M. Huang. 2021. JointGT: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538.

C. Kelly. 2014. Automatic extraction of property norm-like data from large text corpora. *Cognitive science*, 38 4:638–82.

G. H. Kent and A. J. Rosanoff. 1910. *A Study Of Association In Insanity*, volume LXVII. The American Journal of Insanity.

D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907.

B. Khazaeenezhad and A. Alibabaee. 2013. Investigating the role of L2 language proficiency in word association behavior of L2 learners: A case of Iranian EFL learners. *Theory and Practice in Language Studies*, 3(1):108.

G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. 1973. An associative thesaurus of english and its computer analysis. *The Computer and Literary Studies*, pages 153–165.

N. Kitaev and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2676–2686.

P. Kolers. 1963. Interlingual word associations. *Journal of Verbal Learning and Verbal Behavior*, 2:291–300.

A. Korshuk. 2005. Learning more about cultures through free word association data. *Journal of Intercultural Communication*, 5(1):1–11.

L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet. 2009. Directional distributional similarity for lexical expansion. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 69–72.

Z. Kozareva and E. Hovy. 2010. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1482–1491.

G. Kremer and M. Baroni. 2011. A set of semantic norms for german and italian. *Behavior Research Methods*, 43:97–109.

A. A. Kumar. 2021. Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28:40–80.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations*. OpenReview.net.

T. K. Landauer. 1986. How much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive science*, 10:477–493.

T. K. Landauer and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

F. Lehmann. 1992. *Semantic networks in artificial intelligence*. Elsevier Science Inc.

D. B. Lenat, M. Prakash, and M. Shepherd. 1985. Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, 6:65.

H. J. Levesque, E. Davis, and L. Morgenstern. 2011. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561.

O. Levy, S. Remus, C. Biemann, and I. Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

X. Li, A. Taheri, L. Tu, and K. Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455.

Z. W. Lim, H. Stuart, S. De Deyne, T. Regier, E. Vylomova, T. Cohn, and C. Kemp. 2022. A computational approach to identifying cultural keywords across languages. *PsyArXiv*.

B. Y. Lin, X. Chen, J. Chen, and X. Ren. 2019a. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839.

B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840.

C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304.

R. Lin and H. T. Ng. 2022. Does BERT know that the IS-a relation is transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99.

S.-Y. Lin, H.-C. Chen, T.-H. Chang, W.-E. Lee, and Y.-T. Sung. 2019b. Clad: A corpus-derived chinese lexical association database. *Behavior Research Methods*, 51:2310 – 2336.

Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2181–2187.

C. Liu, T. Cohn, and L. Frermann. 2021. Commonsense knowledge in word associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 481–495.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the 8th International Conference on Learning Representations*. OpenReview.net.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

S. Lv, D. Guo, J. Xu, D. Tang, N. Duan, M. Gong, L. Shou, D. Jiang, G. Cao, and S. Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8449–8456.

J. Lyons. 1977. *Semantics: Volume 2*. ACLS Humanities E-Book. Cambridge University Press.

Y. Ma, H. Peng, and E. Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, volume 32, pages 5876–5883.

C. Malaviya, C. Bhagavatula, A. Bosselut, and Y. Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 2925–2933.

A. B. Markman and J. R. Rein. 2013. The Nature of Mental Concepts. In *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, volume 30, pages 6297–6308.

J. P. McCrae, A. Rademaker, F. Bond, E. Rudnicka, and C. Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252.

J. P. McCrae, A. Rademaker, E. Rudnicka, and F. Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19.

K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

K. McRae and M. N. Jones. 2013. Semantic Memory. In *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

K. McRae, S. Khalkhali, and M. Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. *The adolescent brain: Learning, reasoning, and decision making*, pages 39–66.

K. McRae, V. D. S. R, and M. S. Seidenberg. 1997. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99.

O. Melamud, J. Goldberger, and I. Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

T. Mihaylov and A. Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832.

T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*. OpenReview.net.

G. A. Miller. 1985. Wordnet: A dictionary browser. In *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary, Information in Data*, pages 25–28. University of Waterloo.

G. A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

G. A. Miller and C. Fellbaum. 1991. Semantic networks of english. *Cognition*, 41(1-3):197–229.

B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

F. Moghimifar, L. Qu, T. Y. Zhuo, G. Haffari, and M. Baktashmotlagh. 2021. Neural-symbolic commonsense reasoner with relation predictors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 797–802.

S. Mollin. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics & Linguistic Theory*, 5(2).

A. S. Morais, H. Olsson, and L. J. Schooler. 2013. Mapping the structure of semantic memory. *Cognitive science*, 37(1):125–145.

H. Moss and L. Older. 1996. *Birkbeck word association norms*. Psychology Press.

H. E. Moss, R. K. Ostrin, L. K. Tyler, and W. D. Marslen-Wilson. 1995. Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, memory, and cognition*, 21(4):863.

N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

A. Mousavi, X. Zhan, H. Bai, P. Shi, T. Rekatsinas, B. Han, Y. Li, J. Pound, J. Susskind, N. Schluter, I. Ilyas, and N. Jaitly. 2023. Construction of paired knowledge graph-text datasets informed by cyclic evaluation. *arXiv preprint arXiv:2309.11669*.

D. Moussallem, M. Wauer, and A.-C. N. Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19.

S. Muresan and J. Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 231–234.

G. Murphy. 2004. *The big book of concepts*. MIT press.

S. Namei. 2004. Bilingual lexical development: A Persian–Swedish word association study. *International Journal of Applied Linguistics*, 14(3):363–388.

D. Naumann, D. Frassinelli, and S. Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 76–85.

R. Navigli. 2022. 22 ontologies. In *The Oxford Handbook of Computational Linguistics*, pages 518–546. Oxford University Press.

R. Navigli and S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

R. Navigli, P. Velardi, and J. M. Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3716–3722.

T. Nayak and H. T. Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 8528–8535.

S. Necşulescu, S. Mendes, D. Jurgens, N. Bel, and R. Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 182–192.

A. Neelakantan, B. Roth, and A. McCallum. 2015. Compositional vector space models for knowledge base completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 156–166.

D. Nelson, C. McEvoy, and T. A. Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36:402–407.

D. L. Nelson and T. A. Schreiber. 1992. Word concreteness and word structure as independent determinants of recall. *Journal of memory and language*, 31(2):237–260.

J. von Neumann. 1958. *The Computer and the Brain*. Yale University Press.

T.-P. Nguyen, S. Razniewski, J. Romero, and G. Weikum. 2023. Refined commonsense knowledge from large-scale web contents. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8431–8447.

S. Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

H. B. Nissen and B. Henriksen. 2006. Word class influence on word association test results 1. *International Journal of Applied Linguistics*, 16(3):389–408.

P. Norvig. 1987. *A Unified Theory of Inference for Text Understanding*. Ph.D. thesis, University of California, Berkeley.

OpenAI. 2022. OpenAI: Introducing ChatGPT https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

N. Ordan and S. Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

M. Orita. 2002. Word associations of japanese efl learners and native speakers: Shifts in response type distribution and the associative development of individual words. *ARELE: Annual Review of English Language Education in Japan*, 13:111–120.

C. E. Osgood, T. A. Sebeok, J. W. Gardner, J. B. Carroll, L. D. Newmark, S. M. Ervin, S. Saporta, J. H. Greenberg, D. E. Walker, J. J. Jenkins, et al. 1954. Psycholinguistics: a survey of theory and research problems. *The Journal of Abnormal and Social Psychology*, 49(4p2):i.

S. Ostermann, A. Modi, M. Roth, S. Thater, and M. Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3567–3574.

S. Ostermann, M. Roth, and M. Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117.

S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120.

P. Pantel and D. Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 321–328.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

M. Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 137–145. Association for Computing Machinery.

B. Peng, E. Chersoni, Y.-Y. Hsu, and C.-R. Huang. 2022. Discovering financial hypernyms by prompting masked language models. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 10–16. European Language Resources Association.

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54.

F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473.

A. Piermattéo, J. Tavani, and G. Monaco. 2018. Improving the study of social representations through word associations: Validation of semantic contextualization. *Field Methods*, 30:329–344.

S. Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Y. Pinter, R. Guthrie, and J. Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112.

A.-L. Pitel, F. Eustache, and H. Beaunieux. 2014. Component processes of memory in alcoholism: pattern of compromise and neural substrates. In Edith V. Sullivan and Adolf Pfefferbaum, editors, *Alcohol and the Nervous System*, volume 125, pages 211–225. Elsevier.

B. Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

J. Pujara, E. Augustine, and L. Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1751–1756.

G. Qin and J. Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.

X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

M. R. Quillian. 1966. *Semantic memory*. Ph.D. thesis, Carnegie Institute of Technology.

M. R. Quillian. 1967. Word concepts: a theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5):410–430.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog, 1:8*.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

N. F. Rajani, B. McCann, C. Xiong, and R. Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

H. Rashkin, A. Bosselut, M. Sap, K. Knight, and Y. Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2289–2299.

A. Ravichander, E. Hovy, K. Suleman, A. Trischler, and J. C. K. Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.

J. Read. 1993. The development of a new measure of l2 vocabulary knowledge. *Language Testing*, 10:355 – 371.

G. Recchia and M. N. Jones. 2012. The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, 6:315.

B. V. Rensbergen, S. D. Deyne, and G. Storms. 2016. Estimating affective word covariates using word association data. *Behavior Research Methods*, 48:1644–1652.

M. Ribeiro, S. Singh, and C. Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101.

H. L. Roediger, F. M. Zaromb, and W. Lin. 2017. A typology of memory terms. In John H. Byrne, editor, *Learning and Memory: A Comprehensive Reference (Second Edition)*, second edition edition, pages 7–19. Academic Press.

S. Roller and K. Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172.

S. Roller, D. Kiela, and M. Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–363.

M. R. Rosenzweig. 1961. Comparisons among word-association responses in English, French, German, and Italian. *The American Journal of Psychology*, 74(3):347–360.

T. Safavi and D. Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067.

A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. 2017. A simple neural network module for relational reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4974–4983.

A. Santos, S. E. Chaigneau, W. K. Simmons, and L. W. Barsalou. 2011. Property generation reflects word association and situated simulation. *Language and Cognition*, 3(1):83–119.

E. Santus, A. Lenci, Q. Lu, and S. S. im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42.

E. Santus, F. Yung, A. Lenci, and C.-R. Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69.

M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, volume 33, pages 3027–3035.

M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

T. Schick and H. Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 8766–8774.

M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pages 593–607.

E. Segev. 2021. *Semantic network analysis in social sciences*. Routledge.

M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations*. OpenReview.net.

C. Shani and J. Vreeken. 2023. Towards concept-aware large language models. *arXiv preprint arXiv:2311.01866*.

P. Shi and J. Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv preprint arXiv:1904.05255*.

K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

V. Shwartz and Y. Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870. International Committee on Computational Linguistics.

V. Shwartz, Y. Goldberg, and I. Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398.

V. Shwartz, E. Santus, and D. Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75.

V. Silva, S. Handschuh, and A. Freitas. 2018. Recognizing and justifying text entailment through distributional navigation on definition graphs. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 32(1).

P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237.

A. Sinopalnikova. 2004. Word association thesaurus as a resource for building wordnet. In *Proceedings of the Second International WordNet Conference, GWC 2004*, pages 199–205.

E. E. Smith, E. J. Shoben, and L. J. Rips. 1974. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3):214–241.

R. Snow, D. Jurafsky, and A. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17, pages 1297–1304.

R. Snow, D. Jurafsky, and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808.

R. Speer and C. Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, pages 161–176.

R. Speer, J. Chin, and C. Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, AAAI'17, pages 4444–4451.

R. Speer and C. Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686.

M. Steyversa and J. B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

S. Storks, Q. Gao, and J. Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.

L. B. Szalay and J. E. Deese. 1978. *Subjective Meaning and Culture: An Assessment Through Word Associations*. Lawrence Erlbaum.

J. Szymanski and W. Duch. 2007. Semantic memory architecture for knowledge acquisition and management. In *6th International Conference on Information and Management Sciences (IMS2007)*, pages 342–348. California Polytechnic State University.

M. Taboada and W. C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

K. Takeoka, K. Akimoto, and M. Oyamada. 2021. Low-resource taxonomy enrichment with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2747–2758.

A. Talmor, J. Herzig, N. Lourie, and J. Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

N. Tandon, G. de Melo, F. Suchanek, and G. Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 523–532. Association for Computing Machinery.

N. Tandon, G. de Melo, and G. Weikum. 2017. WebChild 2.0 : Fine-grained commonsense knowledge distillation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, System Demonstrations*, pages 115–120.

A. Thawani, B. Srivastava, and A. Singh. 2019. SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 43–51.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

E. Tulving. 1972. Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York.

B. D. Van Durme. 2009. *Extracting implicit knowledge from text*. University of Rochester.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations*. OpenReview.net.

D. P. Vinson and G. Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40:183–190.

D. Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1063–1064.

E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682.

Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, and S. Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of*

the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355.

H. Wang, G. Lu, J. Yin, and K. Qin. 2021a. Relation extraction: A brief survey on deep neural network based methods. In *2021 The 4th International Conference on Software Engineering and Information Management*, pages 220–228.

P. Wang, N. Peng, F. Ilievski, P. Szekely, and X. Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140.

Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29:2724–2743.

W. M. Wang, Z. Li, Z. Tian, J. Wang, and M. N. Cheng. 2018b. Extracting and summarizing affective features and responses from online product descriptions and reviews: A Kansei text mining approach. *Engineering Applications of Artificial Intelligence*, 73:149–162.

X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei, and M. J. Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 7208–7215.

X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Z. Wang, J. Zhang, J. Feng, and Z. Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, volume 28, page 1112–1119.

A. Webson and E. Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, pages 2300–2344.

J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259.

J. Weeds and D. Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 81–88.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

N. Weir, A. Poliak, and B. V. Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society*, CogSci.

S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658.

A. Williams, N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

L. T. Wilson. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

B. Wolter. 2001. Comparing the l1 and l2 mental lexicon: A depth of individual word knowledge model. *Studies in second language acquisition*, 23(1):41–69.

L. L. Wu and L. W. Barsalou. 2009. Perceptual simulation in conceptual combination: evidence from property generation. *Acta Psychologica*, 132 2:173–89.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

M. Xiao and C. Liu. 2016. Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263.

Y. Xu, C. Zhu, R. Xu, Y. Liu, M. Zeng, and X. Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207.

P. Yao, T. Renwick, and D. Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970.

M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, and J. Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 37309–37323.

H. Yokokawa, S. Yabuuchi, S. Kadota, Y. Nakanishi, and T. Noro. 2002. Lexical networks in L2 mental lexicon: Evidence from a word-association task for Japanese EFL learners. *Language education & technology*, 39:21–39.

X. Yu, Z. Xu, and L. Sun. 2011. On Chinese EFL learners' homonym processing in relation to their organization of L2 mental lexicon. *International Journal of English Linguistics*, 1(2):40.

Z. Yu, H. Wang, X. Lin, and M. Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1390–1397.

A. Zareva. 2007. Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second language research*, 23(2):123–153.

R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

H. Zhang, D. Khashabi, Y. Song, and D. Roth. 2021a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4004–4010.

H. Zhang, X. Liu, H. Pan, Y. Song, and C. W. Leung. 2020a. ASER: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, pages 201–211.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *8th International Conference on Learning Representations*. OpenReview.net.

X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec. 2021b. GreaseLM: Graph reasoning enhanced language models. In *Proceedings of the 9th International Conference on Learning Representations*. OpenReview.net.

Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

J. Zhao, Y. Kim, K. Zhang, A. Rush, and Y. LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5902–5911.

W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Y. Zhao, J. Zhang, Y. Zhou, and C. Zong. 2021. Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4039–4045.

J. Zhou, J. X. Huang, Q. V. Hu, and L. He. 2020a. Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. *Knowledge-Based Systems*, 205:106292.

X. Zhou, Y. Zhang, L. Cui, and D. Huang. 2020b. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Y. Zhu, J. Wan, Z. Zhou, L. Chen, L. Qiu, W. Zhang, X. Jiang, and Y. Yu. 2019. Triple-to-text: Converting rdf triples into high-quality natural languages via optimizing an inverse kl divergence. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 455–464.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

M. Zortea, B. Menegola, A. Villavicencio, and J. F. d. Salles. 2014. Graph analysis of semantic word association among children, adults, and the elderly. *Psicologia: Reflexão e Crítica*, 27:90–99.

Appendix A

WAX Annotation Guideline

In Chapter 3, we crowd-sourced the WAX dataset using Amazon Mechanical Turk. Here, we present the key instructions for our two-phase data collection (cf., Figure 3.2).[1] In Phase 1, we first elicit associations from participants, with instructions as shown in Figure A.1, and then ask the same participants to explain their associations in a short sentence, as shown in Figure A.2. In Phase 2, we selected reliable participants to label the high-level relation types for a sampled set of (cue, association, explanation) triples, with the relevant instructions displayed in Figure A.3..

[1]For the full instructions, please refer to our GitHub repository: https://github.com/ChunhuaLiu596/WAX/tree/main/annotation_guideline.

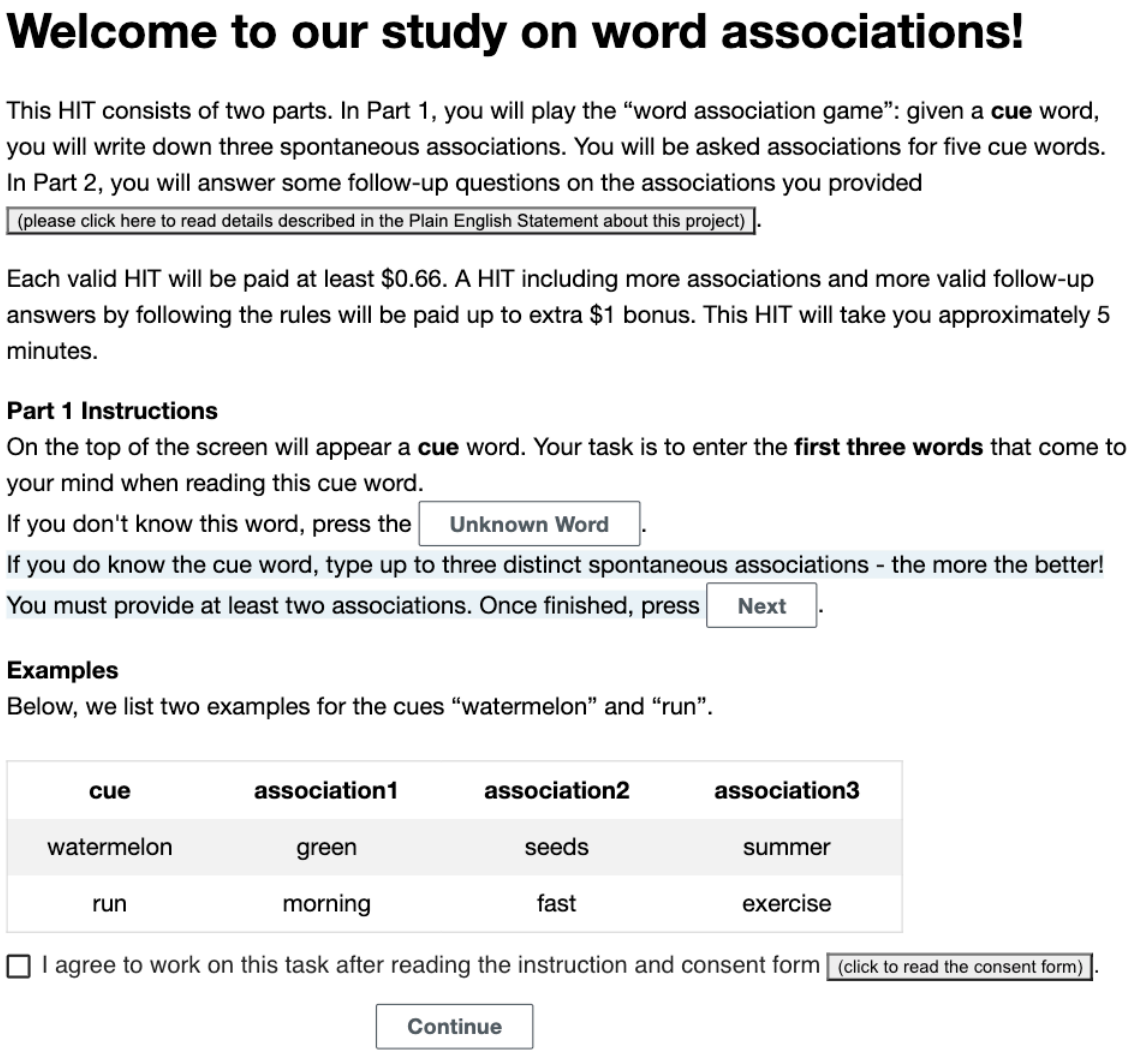


Figure A.1 WAX Annotation Interface for Phase 1: Word Association Generation. Each participant is given five cue words and instructed to generate up to three associations that first come to their mind.

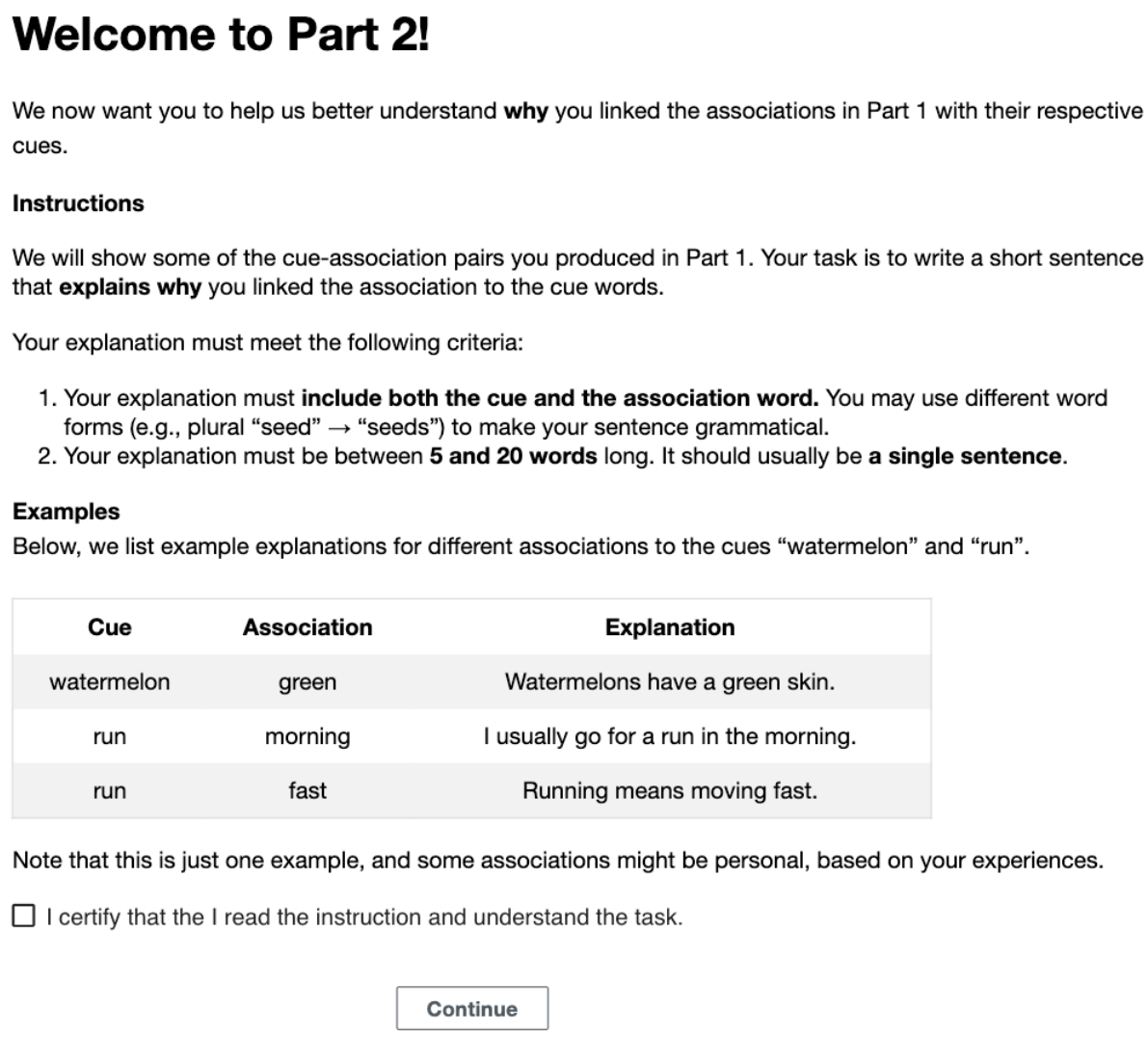
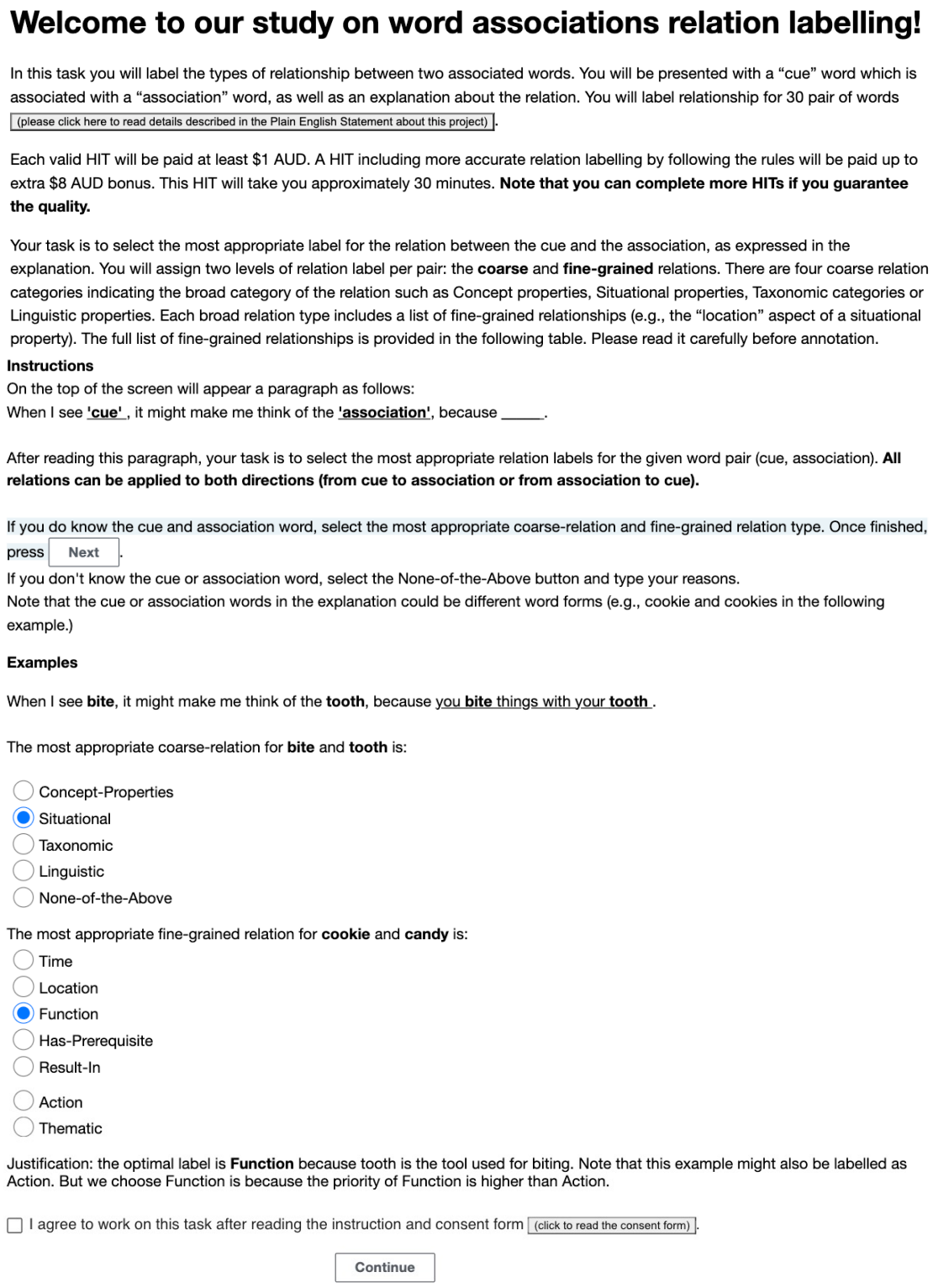


Figure A.2 WAX Annotation Interface for Phase 1: Word Association Explanation. Participants who generate the associations are provided with a task to give natural language sentence explanations, describing why they associate the given cue words with their associations.



Welcome to our study on word associations relation labelling!

In this task you will label the types of relationship between two associated words. You will be presented with a "cue" word which is associated with a "association" word, as well as an explanation about the relation. You will label relationship for 30 pair of words (please click here to read details described in the Plain English Statement about this project).

Each valid HIT will be paid at least $1 AUD. A HIT including more accurate relation labelling by following the rules will be paid up to extra $8 AUD bonus. This HIT will take you approximately 30 minutes. **Note that you can complete more HITs if you guarantee the quality.**

Your task is to select the most appropriate label for the relation between the cue and the association, as expressed in the explanation. You will assign two levels of relation label per pair: the **coarse** and **fine-grained** relations. There are four coarse relation categories indicating the broad category of the relation such as Concept properties, Situational properties, Taxonomic categories or Linguistic properties. Each broad relation type includes a list of fine-grained relationships (e.g., the "location" aspect of a situational property). The full list of fine-grained relationships is provided in the following table. Please read it carefully before annotation.

Instructions

On the top of the screen will appear a paragraph as follows:
When I see **'cue'** , it might make me think of the **'association'**, because _____.

After reading this paragraph, your task is to select the most appropriate relation labels for the given word pair (cue, association). **All relations can be applied to both directions (from cue to association or from association to cue).**

If you do know the cue and association word, select the most appropriate coarse-relation and fine-grained relation type. Once finished, press Next .
If you don't know the cue or association word, select the None-of-the-Above button and type your reasons.
Note that the cue or association words in the explanation could be different word forms (e.g., cookie and cookies in the following example.)

Examples

When I see **bite**, it might make me think of the **tooth**, because you **bite** things with your **tooth** .

The most appropriate coarse-relation for **bite** and **tooth** is:

- ○ Concept-Properties
- ● Situational
- ○ Taxonomic
- ○ Linguistic
- ○ None-of-the-Above

The most appropriate fine-grained relation for **cookie** and **candy** is:

- ○ Time
- ○ Location
- ● Function
- ○ Has-Prerequisite
- ○ Result-In
- ○ Action
- ○ Thematic

Justification: the optimal label is **Function** because tooth is the tool used for biting. Note that this example might also be labelled as Action. But we choose Function is because the priority of Function is higher than Action.

☐ I agree to work on this task after reading the instruction and consent form (click to read the consent form).

Continue

Figure A.3 WAX Annotation Interface for Phase 2: Labelling Relations. This phase involves labelling relations for each (cue, association, explanation) triple from Phase 1. The relation inventory is described in Section 3.3.2.

Appendix B

Relation Mappings between WAX and ConceptNet

In Figure 4.5 of Chapter 4, we compared the relations in the shared pairs of ConceptNet and SWOW (i.e., CN∩SW) with those in WAX. We mapped the relations in ConceptNet to those in WAX, and Table B.1 provides our mappings.

ConceptNet Relations	WAX Relations	ConceptNet Relations	WAX Relations
Antonym	Antonym	AtLocation	Location
CapableOf	Action	Causes	ResultIn
Entails	PartOf	HasA	PartOf
PartOf	PartOf	HasContext	Thematic
HasFirstSubevent	Action	HasLastSubevent	Action
HasPrerequisite	HasPrerequisite	HasProperty	HasProperty
HasSubevent	Action	IsA	CategoryExemplar
InstanceOf	CategoryExemplar	MadeOf	MadeOf
MannerOf	CategoryExemplar	ReceivesAction	Action
SimilarTo	Synonym	Synonym	Synonym
UsedFor	Function	MotivatedByGoal	EmotionEvaluation
CausesDesire	EmotionEvaluation	Desires	EmotionEvaluation
DistinctFrom	Members	DerivedFrom	Lexical
FormOf	Lexical	EtymologicallyRelatedTo	Lexical
EtymologicallyDerivedFrom	Lexical	CreatedBy	Others
NotHasproperty	Others	Definedas	Others
Capital	Others	LocatedNear	Others
NotDesires	Others	Genre	Others
Influencedby	Others	Genus	Others
Product	Others	Field	Others
Occupation	Others	Language	Others
NotCapableof	Others	Symbolof	Others

Table B.1 Relation mappings from ConceptNet to WAX relations.

Appendix C

Results of Reproducing KG-Augmented Models

For all KG-augmented models used in Chapter 4, we use the implementation from previous work (Wang et al., 2020).[1] Table C.1 compares our re-run results to the original numbers reported in the respective papers. All our reproduced scores are comparable to or better than reported numbers. All of our experiments are run on single GPU of NVIDIA V100 16G.

Model	CSQA		OBQA	
	Wang et al. (2020)	Our re-run	Wang et al. (2020)	Our re-run
w/o KG	68.69 (±0.56)	70.46 (±0.18)	64.80 (±2.37)	64.47 (±3.01)
+ GconAttn	69.88 (±0.47)	70.59 (±0.66)	64.75 (±1.48)	69.00 (±1.41)
+ RN	69.59 (±3.80)	72.79 (±0.63)	65.20 (±1.18)	65.30 (±0.99)
+ PG-Global	71.55 (±0.99)	71.69 (±0.29)	68.40 (±0.31)	67.73 (±0.61)
+ PG-Full	72.68 (±0.42)	72.20 (±0.08)	71.20 (±0.96)	67.40 (±0.28)

Table C.1 Comparisons of our re-production of various KG-augmented models with previous work. The RoBERTa-large encoder is used. We use the same provided code by Wang et al. (2020) for CSQA and OBQA. Note that text representation on OBQA is the the average pooling over the hidden states of the last layer of RoBERTa rather than 'CLS' token representation.

[1] https://github.com/INK-USC/MHGRN