Computing Relationships and Relatedness Between Contextually Diverse Entities.

A thesis presented
by

Karl Grieser

Submitted in total fulfillment of the requirements for the degree of Doctor of Philosophy

November 2011

The Department of Information Systems
The University of Melbourne
Computing Relationships and Relatedness Between Contextually Diverse Entities.

Abstract

When presented with a pair of entities such as a ball and a bat, a person may make the connection that both of these entities are involved in sport (e.g., the sports baseball or cricket, based on the individual’s background), that the composition of the two entities is similar (e.g., a wooden ball and a wooden stick), or if the person is especially creative, a fancy dress ball where someone has come dressed as a bat. All of these connections are equally valid, but depending on the context the person is familiar with (e.g., sport, wooden objects, fancy dress), a particular connection may be more apparent to that person. From a computational perspective, identifying these relationships and calculating the level of relatedness of entity pairs requires consideration of all ways in which the entities are able to interact with one another. Existing approaches to identifying the relatedness of entities and the semantic relationships that exist between them fail to take into account the multiple diverse ways in which these entities may interact, and hence do not explore all potential ways in which entities may be related. In this thesis, I use the collaborative encyclopedia Wikipedia as the basis for the formulation of a measure of semantic relatedness that takes into account the contextual diversity of entities (called the Related Article Conceptual Overlap, or RACO, method), and describe several methods of relationship extraction that utilise
the taxonomic structure of Wikipedia to identify pieces of text that describe relations between contextually diverse entities. I also describe the construction of a dataset of museum exhibit relatedness judgements used to evaluate the performance of RACO. I demonstrate that RACO outperforms state-of-the-art measures of semantic relatedness over a collection of contextually diverse entities (museum exhibits), and that the taxonomic structure of Wikipedia provides a basis for identifying valid relationships between contextually diverse entities. As this work is presented in regard to the domain of Cultural Heritage and using Wikipedia as a basis for representation, I additionally describe the process for adapting the principle of conceptual overlap for calculating semantic relatedness and the relationship extraction methods based on taxonomic links to alternate contextually diverse domains, and for use with other representational resources.
# Contents

Title Page .......................... i
Abstract ................................ iii
Table of Contents ....................... v
List of Figures ........................ ix
List of Tables ........................... xi
Acknowledgments ....................... xii
Declaration ............................. xiii
Preface ................................. xiv

1 Introduction .......................... 1
  1.1 Utilising an Information Rich Domain .................. 4
  1.2 Research Objectives ................................. 9
  1.3 Thesis Overview ................................. 11

2 Background ........................... 15
  2.1 Background Overview ......................... 16
  2.2 Domains ................................ 18
    2.2.1 Museum Domains .......................... 20
  2.3 Conceptual Representation ....................... 22
    2.3.1 Word Sense Disambiguation and Polysemy ......... 24
    2.3.2 WordNet ................................ 26
    2.3.3 Wikipedia .............................. 27
    2.3.4 Cultural Heritage Ontologies ................. 32
  2.4 Conceptual Combination ....................... 35
    2.4.1 Lowest Common Subsumer ..................... 39
    2.4.2 Computational Measurement of Semantic Relatedness .. 42
  2.5 Semantic Relations ............................ 49
    2.5.1 Relation Extraction ........................ 53
  2.6 User Modelling and Prediction ................. 56
  2.7 Evaluation in Language Technology .............. 60
    2.7.1 Comparison to Gold Standard Data Sets ........ 60
# Contents

2.7.2 Manual Annotation of Results ........................................ 63  
2.7.3 Mechanical Turk ......................................................... 66  
2.8 Summary ......................................................................... 67  

## 3 Computational Resources and Domain Requirements 70  
3.1 Melbourne Museum ............................................................ 72  
3.1.1 Museum Collection ......................................................... 73  
3.1.2 Museum Catalogue ........................................................ 75  
3.1.3 Museum Visitors ............................................................. 77  
3.2 Wikipedia ......................................................................... 77  
3.2.1 Wikipedia XML Dumps .................................................. 78  
3.2.2 Wikipedia Category Hierarchy ......................................... 80  
3.3 Evaluation ....................................................................... 83  
3.3.1 Annotator Selection ......................................................... 83  
3.3.2 Data Collection and Online Surveys ................................. 84  
3.3.3 Mechanical Turk ............................................................. 85  
3.4 Summary ......................................................................... 86  

## 4 Semantic Relatedness 88  
4.1 Semantic Relatedness of Museum Exhibits .............................. 90  
4.1.1 Physical Distance .......................................................... 93  
4.1.2 Document Similarity ...................................................... 94  
4.1.3 Hierarchical Relatedness ................................................ 96  
4.1.3.1 Path-Based Relatedness ............................................ 96  
4.1.3.2 Information Content-based Ontological Similarity ......... 97  
4.2 Evaluation Data collection ................................................... 98  
4.2.1 Experiment Design ........................................................ 99  
4.2.2 Results ...................................................................... 101  
4.3 Experiments .................................................................... 105  
4.3.1 Results .................................................................... 106  
4.3.2 Analysis .................................................................... 108  
4.4 Evaluating RACO Over Existing Data ................................. 113  
4.5 Application of RACO to Topic Labelling ............................... 117  
4.5.1 RACO Thresholding ....................................................... 120  
4.5.2 Results .................................................................... 122  
4.6 Summary ....................................................................... 129  

## 5 Semantic Relationship Extraction 132  
5.1 Relationship Identification .................................................. 134  
5.1.1 Category Membership ..................................................... 135  
5.1.2 Article Text ................................................................ 137  
5.1.3 Article Links ................................................................ 140
Contents

Common Third Article ................................. 140
Sibling Reference ................................. 141
5.2 Relationship Extraction Task ......................... 143
  5.2.1 Extraction Discussion ......................... 144
5.3 Evaluation ........................................ 147
  5.3.1 Evaluation Design ......................... 149
  Refinement of the Annotation Interface ................. 152
  5.3.2 Mechanical Turk ............................... 154
  5.3.3 Museum Staff Annotation ......................... 157
5.4 Comparison of Method Performance ..................... 159
5.5 Summary .......................................... 165

6 Domain Diversity and Resource Portability ..................... 169
  6.1 Alternate Semantic Resources ......................... 170
    6.1.1 Domain-Specific Wikis ......................... 171
    6.1.2 Digital Encyclopedias ......................... 173
    6.1.3 Cyc ........................................ 176
    6.1.4 Resource Augmentation and Wikification ............. 180
  6.2 Alternate Domains of Application ..................... 183
    6.2.1 Film ........................................ 185
    Benefit to Domain .................................. 185
    Domain Specific Representation and Resources ............. 186
    Applications and Computational Use ..................... 189
    Adaptation of Methods ................................ 190
    6.2.2 Tourism ...................................... 194
    Benefit to Domain .................................. 195
    Domain Specific Representation and Resources ............. 196
    Application and Computational Use ..................... 199
    Adaptation of Methods ................................ 201
  6.3 Summary .......................................... 205

7 Conclusions ......................................... 209
  7.1 Summary and Research Contributions ..................... 209
  7.2 Future Work ....................................... 213
  7.3 Summary .......................................... 216

A Museum Exhibit Definitions .................................. 217

B Annotation Interfaces ................................ 220
  B.1 Museum Exhibit Relatedness Survey ..................... 220
  B.2 Relationship Evaluation Pilot Annotation 2 ............. 222
  B.3 Mechanical Turk Relationship Evaluation 1 ............. 222
B.4 Mechanical Turk Relationship Evaluation 2 . . . . . . . . . . . . . . . 224
B.5 Museum Staff Relationship Evaluation . . . . . . . . . . . . . . . . . . . . . 225
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>A section of Wikipedia’s category hierarchy</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>The Least Common Subsumer (LCS) of dog varieties.</td>
<td>40</td>
</tr>
<tr>
<td>3.1</td>
<td>Several exhibits at Melbourne Museum</td>
<td>74</td>
</tr>
<tr>
<td>3.2</td>
<td>The CSIRAC exhibit web page from the Melbourne Museum website.</td>
<td>76</td>
</tr>
<tr>
<td>3.3</td>
<td>A piece of raw text with markup from the Wikipedia Porcupine article.</td>
<td>79</td>
</tr>
<tr>
<td>3.4</td>
<td>Portion of the Wikipedia category hierarchy surrounding the several Planet articles.</td>
<td>81</td>
</tr>
<tr>
<td>4.1</td>
<td>An example of an exhibit pair used in the data collection exercise.</td>
<td>100</td>
</tr>
<tr>
<td>4.2</td>
<td>Dendrogram of the clustering of exhibits based on the average pairwise relatedness ratings.</td>
<td>104</td>
</tr>
<tr>
<td>4.3</td>
<td>Dendrogram of the clustering of exhibits based on the minimum pairwise relatedness ratings.</td>
<td>105</td>
</tr>
<tr>
<td>4.4</td>
<td>Linear regression plots of candidate label scores.</td>
<td>127</td>
</tr>
<tr>
<td>5.1</td>
<td>Common ancestry membership</td>
<td>137</td>
</tr>
<tr>
<td>5.2</td>
<td>Relationship identification using a reference to a common third article.</td>
<td>141</td>
</tr>
<tr>
<td>5.3</td>
<td>Identifying a relationship between two articles by identifying a relation to a sibling article.</td>
<td>142</td>
</tr>
<tr>
<td>5.4</td>
<td>The second iteration of the annotation interface.</td>
<td>150</td>
</tr>
<tr>
<td>5.5</td>
<td>The final design of the HIT distributed to Mechanical Turkers.</td>
<td>155</td>
</tr>
<tr>
<td>5.6</td>
<td>Validity percentage of relationship extraction methods determined by majority class of manually annotation of relationships, separated by annotation source.</td>
<td>168</td>
</tr>
<tr>
<td>6.1</td>
<td>The IMDB web page for the film <em>Jurassic Park</em></td>
<td>188</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparison of the Copenhagen article in Wikipedia (left) and Wiki-travel (right).</td>
<td>198</td>
</tr>
<tr>
<td>B.1</td>
<td>The set of profiling questions given to museum members on starting the exhibit relatedness survey.</td>
<td>221</td>
</tr>
</tbody>
</table>
B.2 Instructions given to museum members taking part in the exhibit relatedness survey. .................................................. 221
B.3 Instructions given to University of Melbourne postgraduates performing the second annotation task. .......................... 223
B.4 The Instructions given for the first iteration of the Mechanical Turk Relationship Evaluation Task. .......................... 224
B.5 The Instructions given for the second iteration of the Mechanical Turk Relationship Evaluation Task. ......................... 224
B.6 The presentation format used to present relationships to Turkers in the second Mechanical Turk annotation interface. ........ 225
B.7 Instructions given to Melbourne Museum staff members performing the relationship evaluation task. .......................... 226
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Average relatedness score of exhibit pairs given to all survey participants.</td>
<td>102</td>
</tr>
<tr>
<td>4.2</td>
<td>Overall Pearson correlation ($\rho$) and statistical significance (p-value) between the gold-standard relatedness scores and the various exhibit relatedness estimation methods (the highest correlation is highlighted)</td>
<td>107</td>
</tr>
<tr>
<td>4.3</td>
<td>Pearson correlations between human judgements of relatedness and computational measures, partitioned into discretised distance bands.</td>
<td>108</td>
</tr>
<tr>
<td>4.4</td>
<td>Spearman correlation between RACO, WLVM and existing word similarity datasets.</td>
<td>115</td>
</tr>
<tr>
<td>4.5</td>
<td>Inter-annotator agreement of evaluated topic labels calculated using Fleiss’ Kappa.</td>
<td>123</td>
</tr>
<tr>
<td>4.6</td>
<td>Average evaluation score for label candidates, separated by data set.</td>
<td>124</td>
</tr>
<tr>
<td>4.7</td>
<td>Correlation between annotator evaluations and RACO scores, calculated using Spearman correlation.</td>
<td>125</td>
</tr>
<tr>
<td>5.1</td>
<td>Article pairs used in testing, number of relationships extracted by respective methods and corresponding RACO scores. The exhibit titles correspond to the exhibit definitions in Appendix A.</td>
<td>145</td>
</tr>
<tr>
<td>5.2</td>
<td>Fleiss’ Kappa statistic for different partitions of the first round of annotation.</td>
<td>151</td>
</tr>
<tr>
<td>5.3</td>
<td>Fleiss’ Kappa statistic for each extraction method, separated by annotation source.</td>
<td>157</td>
</tr>
<tr>
<td>5.4</td>
<td>Evaluation counts for relations extracted by each extraction method, and as a total, separated by annotation group. Each entry is a comma separated set of $validity,invalidity,undecided$. In the percentage column and row, these three values are presented as an average percentage of the extracted relationships’ validity.</td>
<td>160</td>
</tr>
<tr>
<td>A.1</td>
<td>The exhibits codes, exhibit names, and respective gallery memberships of exhibit entities used in experiments in this thesis.</td>
<td>218</td>
</tr>
<tr>
<td>A.2</td>
<td>The Wikipedia article titles used to represent exhibits used in experiments in this thesis.</td>
<td>219</td>
</tr>
</tbody>
</table>
Acknowledgments

First and foremost, I want to thank my supervisors: Tim and Liz. Their advice and guidance during my candidature has been indispensable. Throughout my candidature they have been model supervisors and have provided me with the solid foundation from which I have been able to build my skills as a researcher.

From an early age, my parents tried to teach me the value of knowledge, and while it took time for me to actually listen to what they were trying to tell me, I did come to eventually understand what they meant. Mum and Dad, thank you so much for raising me to be the person that I am, and for always having faith in me.

The many people in the University of Melbourne LT-Group and the denizens of 6.05, past and present (in no particular order): Jeremy, Lars, Olivia, Andy, Willy, Marco, Patrick, Mel, Bec, Tara, Richard, Sumukh, David, Jey Han, Florian, Bo, Li, Clint, Jim, Steven, and Ned. Thank you for being around to talk over research conundrums and programming problems, or just being willing to grab a coffee (during work) or beer (after or during work) with me.

Thank you to Fabian Bohnert for documenting the physical layout of Melbourne Museum so meticulously and allowing me access to the valuable visitor trace data that he so painstakingly collected. A very special thanks to Carolyn Meehan of Museums Victoria for giving me access to Melbourne Museum data, and helping me with so many iterations of interfaces for visitor surveys and annotation exercises.

Thank you all.
Declaration

This is to certify that:

(i) this thesis comprises only my original work towards the PhD except where indicated in the Preface,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) this thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.
Preface

Sections of the following chapters of this thesis have been published in a different form in the following articles:

Chapter 4

- Using collaboratively constructed document collections to simulate real-world object comparisons.
  Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg.

- Using ontological and document similarity to estimate museum exhibit relatedness.
  Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg.
Chapter 1

Introduction

The adage that a computer is able to perform complex calculations far beyond the capacity of the human brain but does not know how to tie a shoelace is often used to demonstrate that computers are not truly intelligent, but are merely machines that are only able to carry out a set of predefined instructions. This inability to infer connections between previously acquired knowledge and discover new information independently is described by many researchers as a major hurdle on the road to true artificial intelligence (McCarthy 1977; Minsky 1986).

A solution to the problem of an incomplete understanding of how diverse concepts interact is to utilise existing sources of knowledge to discover additional relationships and conceptual interactions. Information rich resources such as newspapers, encyclopedias and the world wide web provide a knowledge base that describe conceptual entities and the interactions they have with one another. Relationships and concepts described in bodies of text are evident to a human reader as they possess the knowledge of what the occurrence of concepts or relationships looks like. For example, in the sentence a dog is a canine, a human reader is able to identify that a dog and a
Canine are concepts and that they are related through the is-a relation. In order to identify this piece of knowledge, a computer must possess definitions of the conceptual entities dog and canine and the is-a relation. In this example, the relationship between the two concepts is explicitly stated, however, bodies of text that contain an implied relationship between two entities require further resources describing the content of the concepts involved and the possible interactions between them. For example, in the sentence The weather was hot, so I went to the beach, the human reader may infer that beaches are frequented on hot days so that one may swim to cool down, however this association of weather and beaches is not evident to a machine without a knowledge base wherein this relation is stated. In order to identify this relationship, a computer would require knowledge that describes the fact that a beach is a location, that hot weather makes people want to cool down, that swimming is an activity that allows a person to cool down, and that it is possible to swim at the beach.

Further difficulty lies in the identifying the specific relation that may exist between a given two entities. A sentence such as in the beach-weather example contains context to specify the relationship type that is being described between the two entities. The same two entities occurring in a different context may indicate a different relation. For example, in the context of holiday locations and the association of weather and beach, the relation between these two entities might be one of destination desirability: the desirability of the beach as a holiday destination is dependent on the weather being hot. This relation may be completely different when the two entities occur in an alternate context, such as a discussion about conservation: extreme weather conditions at a beach may cause erosion. Thus the identification of relations between
entities draws on knowledge of the content of each entity and of the context in which the entities appear.

The identification of relations between entities that appear in multiple different contexts remains an open problem. Previous approaches identify relationships by recognising portions of text that fit a recognised pattern (e.g., the pattern NOUN is a NOUN to recognise is-a relationships such as dog is a canine), see Auger and Barroère (2008) for a review of pattern based methods. A pattern-based approach does not allow for the identification of relationships that have not been previously described, resulting in the omission of unforeseen relationships. This is partially a problem of representation: entities possessing multiple relationships to other entities and, depending on the entity description used, all possible relationships between all entities relationships may not be described using the same format. Furthermore, if a rigid format is used to describe entities and the relationships between them, diverse relationships which are not able to be described using this rigid format may be left out of the representation altogether.

In this thesis I demonstrate that Wikipedia, a semi-structured, broad-coverage knowledge base, can be used as a foundation for identifying relationships between entities occurring in diverse contexts, and can also be used as a basis for calculating the semantic relatedness of these entities. In doing so, I present a newly formulated measure of determining how related a pair of entities is, i.e., a measure of semantic relatedness, that takes into account the contextual diversity of these entities. To identify specific relationships between entities, I present multiple methods of relationship extraction that utilise the taxonomic structure of Wikipedia’s semantic network to identify portions of text that contain relationships. I also describe the construction of a new evaluation data set for the Cultural Heritage domain using an established
methodology, and describe the development of a method for the manual evaluation of extracted relationships from semi-structured text.

1.1 Utilising an Information Rich Domain

Throughout this thesis I use the phrase *contextual diversity* to describe the quality of an object, artifact or concept (hereafter referred to as an *entity*) possessing relations to multiple other entities, but for differing reasons. When considered alongside another entity, a subset of the entity’s content (the definition of the entity) may be identified as being more relevant to the content of the other entity. I use the term *context* to refer to the overlapping subset of content that is common to multiple entities. In this regard, a contextually diverse entity is an entity which possesses relations to multiple contexts, and relations to multiple entities that also refer to those contexts.

For example, consider an exhibit that may appear as a museum exhibit: a bark canoe. This entity can be defined in and of itself, but depending on the context in which it is placed, a particular aspect of the bark canoe’s definition may become prominent. When considering the context of botany (or, *placed in* the context of botany), the type of tree the bark has come from, the location this tree grows, its properties in resisting water, and the surrounding ecosystem become prominent. When placed in context of the indigenous tribe that constructed the canoe, properties such as the tools with which it was constructed, spiritual meaning of the tree from which it came, and it use in providing food for the tribe become prominent. When placed in the context of other watercraft, qualities relating to its design, location of use (e.g., ocean or stream), and purpose (e.g., fishing or transport) become prominent. Some of these
properties may even be relevant in multiple contexts, e.g., the waterproof properties of the tree are relevant to a botanical context and to a ship-building context.

The bark canoe exists as a single artifact in a museum collection, and is presented as an exhibit in this form; in this respect, the single exhibit of the bark canoe is a single entity. In a collection of museum exhibits the bark canoe may be portrayed in terms of any one of the above mentioned contexts by a museum curator by placing it with other exhibits relevant to that context. When described solely with relation to a certain context, museum visitors may be unaware of the contextual diversity of the exhibit, and an element of understanding of the interaction of the exhibit with other collections of exhibits is lost. From a computational perspective, this can be viewed as a problem of representation: how can a group of entities be described in a manner that allows them to possess multiple diverse relations with one another while still being able to utilise a common description format to describe these relations. Indeed, if relations are diverse enough there may be no finite set of relations to which all inter-entity relations subscribe.

Given a large collection of contextually diverse entities, identifying which of these entities are related (across all contexts), the strength of relatedness between them, and the myriad semantic relations that exist within the collection becomes difficult due to there not existing a set of relation types to which all inter-entity relations apply. The automatic identification of semantic relations between contextually diverse entities remains an open problem, particularly when the diversity of relations makes rule-based relation identification methods unable to sufficiently capture all relations within a given collection, and is the purpose of this thesis: to identify a measure of semantic relatedness that takes into account the contextual diversity of entities, and
to formulate methods of relation extraction designed to identify relations between contexts in which the entity occurs.

In order to demonstrate the diversity of context that requires the use of novel measures of semantic relatedness and novel approaches to identifying the semantic relations that exist between them, a collection of entities is required as a basis for comparison. Existing studies that examine the semantic relatedness of entities utilise standardised data sets of entity pairs. These collections of entity pairs are widely used in research, but are not comprised of contextually diverse entities. Chiefly, these sets have been used in the development of measures of semantic relatedness over a variety of semantic networks, however these measures are designed to work with semantic networks with homogeneous conceptual structure and a defined set of relations, and perform poorly over semantic networks other than those they were developed over (Ponzetto and Strübe 2007b). The research presented in this thesis is intended to examine the relatedness of contextually diverse entities and their interactions.

In reference to the desired extraction of relationships between contextually diverse entities (e.g., beach and weather), a collection of entities represented using a format that is able to describe this diversity is desirable, i.e., a collection of entities represented using a format that describes the content of each entity with reference to the multiple contexts in which it appears. A desirable domain of application with which to experiment requires resources to represent the collection of entities and their context within that domain (i.e., the description of how the entity relates to other entities in its domain). For this reason, I focus on the measurement of semantic relatedness and the identification of semantic relationships (collectively referred to hereafter as conceptual combination) between contextually diverse entities within an information rich domain.
The desirability of a domain to be used for experimentation also depends on the benefit to be gained from applying extracted relationships or comparing entities (through calculating semantic relatedness) to existing tasks within the domain. For the purposes of this thesis, the domain used in experimentation and development is that of Cultural Heritage: the domain of museums, historical artifacts and related entities. A discussion of alternate information-rich domains with contextually diverse content is presented in Chapter 6.

The Cultural Heritage (CH) domain is comprised of contextually diverse entities, representing physical objects, periods of time, that possess multiple relationships to one another. This domain also has two groups of people that stand to benefit from the analysis of conceptual combination of Cultural Heritage entities for different reasons: museum visitors in understanding the connection between collections of CH entities, and museum curators in organising CH entities. The Cultural Heritage domain is chosen due to the availability of resources, and an existing relationship with a Cultural Heritage institution: Melbourne Museum. This relationship has enabled access to resources such as a pool of exhibits to use as a core set of contextually diverse entities, information relating to visitor behaviour in the museum, and existing groups of people familiar with the content at the institution. This final aspect is particularly important during the evaluation phases of experiments in this thesis: a pool of annotators familiar with the collection being analysed is desirable when evaluating extracted relationships between museum exhibits and the level of relatedness of museum exhibits.

This thesis investigates the manner in which relationships between contextually diverse entities may be identified, and their level of relatedness measured. The identification of the strength of relatedness between pairs of museum exhibits could be
used as a component in the construction of a content based recommender system, potentially for use in prediction of visitor paths. In addition to the scoring of exhibit relatedness, the extraction of the relations between exhibits can identify specific reasons for exhibit relatedness. Exhibit-exhibit relations could be used as a component in the construction of personalised summaries of visitors’ tours and in the explanation of exhibit recommendations.

Existing Cultural Heritage resources use specialised description formats to describe the content of entities and the relations between them. There exist multiple resources that catalogue conceptual entities and the diverse relationships that exist between them. These range from the highly structured representation of single entities, defined as a logical combination of more general classes, to large bodies of plain text that describe the content of a collection of museum exhibits. I strike a compromise between these two extremes: I utilise a collaboratively constructed hierarchical semantic network of encyclopedic entities that are described using a combination of plain text and explicit links to other entities within the semantic network. Specifically, I use the collaborative encyclopedia, Wikipedia, to represent these conceptually diverse entities. A more complete discussion of the rationale behind this choice, is given in Chapter 2, while further discussion of the use of alternate semantic networks for the methods constructed in this thesis is performed in Chapter 6. An analysis of the portability of the methods constructed in this thesis to other domains is also performed in Chapter 6.
1.2 Research Objectives

This thesis is primarily concerned with the measurement of strength of relatedness and identification of relationships between contextually diverse entities. The use of a knowledge base that represents its content using a semi-structured format (an explicit set of links and properties that must be adhered to, combined with unstructured information, e.g., bodies of text) to describe the content of an entity, allows the knowledge base to describe potential relationships between two entities (e.g., through the use of a plain text description to describe the interaction of two entities). However, as there exists no common format for a relationship’s description, the automated recognition of relationships within an entity’s content becomes more difficult: both structured and unstructured content must be examined to identify relationships that may exist between entities.

The primary objectives of this thesis are:

1. To identify a method for calculating the strength of relatedness between entities in a collection of contextually diverse entities.

2. To construct methods of semantic relationship extraction to identify relationships between entities contained in a semantic network of contextually diverse entities.

There already exist measures of semantic relatedness designed to function over semantic networks of contextually diverse entities, however, these measures do not take into account qualities such as the taxonomic organisation of entities. My proposed measure of semantic relatedness utilises the taxonomic organisation of entities within the collaborative encyclopedia, Wikipedia. Wikipedia and its network structure is described further in Section 2.3.3 and Section 3.2, while the motivation behind using
taxonomic organisation to identify the conceptual overlap of entities is described in Chapter 4. A direct comparison to an alternate measure of semantic relatedness constructed for use with Wikipedia that does not utilise the taxonomic overlap of entities is also performed in Chapter 4.

In formulating methods of relationship identification between contextually diverse entities, a novel approach is required due to the untyped links between entities present in Wikipedia (the semantic network used in this thesis). Previous approaches to extracting semantic relations in unstructured text have relied on pattern matching to identify relations and relation arguments. This approach is infeasible in Wikipedia as all links between entities are untyped due to the semi-structured representation used. The research presented in this thesis demonstrates that taxonomic links between entities, such as common ancestry and common links to additional entities, can be used as the basis of identifying semantic relations between entities.

These research objectives are approached in sequence. The use of Wikipedia to represent a collection of contextually diverse entities is described in Chapter 2 and Chapter 3, the construction of a measurement of semantic relatedness for use over this network is described in Chapter 4, and finally the use of the taxonomic structure of Wikipedia to identify semantic relations between entities is described in Chapter 5. In each of these chapters, constructed methods are evaluated using human annotators: Chapter 4 uses an existing methodology to evaluate calculated semantic relatedness scores against gold-standard human evaluations of the same data; and Chapter 5 required the formulation of a new method of evaluation to take into account the complex nature of semantic relationships without a common form. The motivation behind each of these evaluation approaches is described in Section 2.7, Section 3.3,
and in their respective chapters. A description of the content of each of these chapters is presented below.

1.3 Thesis Overview

The work presented in this thesis draws on areas of research from cognitive science, language technology, computing and Cultural Heritage. Chapter 2 places the research presented in this thesis in context of these research areas, an overview of the interaction of this existing research is presented in Section 2.1. An overview of domains in conceptual representation is described in Section 2.2, examining the difference in content of different collections of entities. A broader examination of the foundations of conceptual representation and the differences in computational semantic networks is presented in Section 2.3. This includes a discussion of the advantages and disadvantages of using a large-scale, semantic network to represent contextually diverse entities, as well as the problem of Word Sense Disambiguation, which inspires the use of measures of semantic relatedness over collections of museum exhibits. The notion of conceptual combination is further expanded upon in Section 2.4, which contains a description of established measures of semantic relatedness over multiple semantic networks. Section 2.5 examines different approaches to extracting semantic relations from unstructured text, and the difficulties in identifying untyped relations in unstructured text. Section 2.6 contains a discussion of the use of semantic relations and semantic relatedness in User Modelling applications. Finally, the evaluation methodology for the computational methods proposed in this thesis is described in Section 2.7.

Chapter 3 describes the specifics of the resources used in the experiments con-
duced in this thesis. This begins with a description of Melbourne Museum in Section 3.1, including its layout, exhibit catalogue and museum website. Section 3.2 contains a description of how Wikipedia was adapted to a computational semantic network, with specific mention of how problems such as network cycles, identification of Least Common Subsumer, and reproducibility of experimental results were addressed. Section 3.3 concludes with a discussion of the specific resources used in the evaluation framework for experiments conducted in Chapter 4 and Chapter 5. Included in this section is a description of the Amazon Mechanical Turk evaluation framework.

Chapter 4 presents the creation and evaluation of a measure for calculating the semantic relatedness of contextually diverse entities in a hierarchical semantic network. I present a novel measure of semantic relatedness (named the Related Article Category Overlap, or RACO, method) designed to work with contextually diverse entities contained in a semantic networks in Section 4.1. This section also describes the motivation for the creation of a this new measure with reference to existing measures of semantic relatedness, and contains a summary of existing measures of semantic relatedness that are used as comparison in experiments presented later in the chapter. The Cultural Heritage domain used in this thesis presents a novel domain for experimentation, but there exists no Cultural Heritage evaluation data set for use with calculations of semantic relatedness. Section 4.2 explains how an evaluation data set was constructed using a collection of museum exhibits from the Melbourne Museum catalogue, and relatedness judgements provided by Melbourne Museum visitors. The performance of the RACO measure of semantic relatedness is compared to existing measures of semantic relatedness over the aforementioned museum visitor judgement data set in Section 4.3. As a demonstration of the portability of this measure to
a different domain, Section 4.4 demonstrates the use of the \textit{RACO} measure over an existing data set, and compare it with existing state of the art measures of semantic relatedness.

Chapter 5 demonstrates the use of taxonomic links between entities in a semantic network as a basis for extracting specific semantic relations between entities. Section 5.1 provides a motivation for examining the taxonomic organisation of articles as a basis for identifying semantic relations between entities, and presents a group of relationship extraction methods that utilise the taxonomic structure in a variety of ways to identify unstructured strings of text that contain a valid relation between the two entities. The application of these relation identification methods in extracting relationships between entities representing museum exhibits present at Melbourne Museum is presented in Section 5.2. Evaluating these extracted relationships is a multi-layered task, and multiple approaches to the problem of evaluation were taken: the use of different groups of evaluators and a discussion of the precise nature of the evaluation task itself is presented in detail in Section 5.3. Finally, a comparison of the methods of relationship identification is performed in Section 5.4, including a comprehensive discussion of the nature of the evaluation task and how the results can be interpreted with regard to the amorphous nature of the task.

Chapter 6 examines the portability of the methods proposed in this thesis. The application domain used in the development and testing of measures of semantic relatedness and relation extraction is that of Cultural Heritage, and much of this thesis refers to this domain being desirable due to its diversity and existing resource base. However, the methods developed in this thesis are not intended to be restricted solely to the Cultural Heritage domain. Section 6.1 describes a number of alternate resources that can be used interchangeably with the methods devised in this thesis,
such as alternate encyclopedic networks to Wikipedia. This also includes a discussion of augmenting existing document collections to create domain specific entities that are able to interface with existing semantic networks. This is followed by a discussion of two alternate domains of contextually diverse entities that are able to take advantage of the methods presented in this thesis, and the level of adaptation required for the transposition of these methods. These alternate domains are presented in Section 6.2.

Finally, I conclude this thesis in Chapter 7 with a summary of how the methods created in this thesis address my original research objectives and contribute to the existing literature on semantic relatedness and relation extraction.
Chapter 2

Background

The goal of identifying complex interactions between contextually diverse entities requires the adaptation of resources and methodologies from a number of computing fields. This task can be approached by decomposing the overall task into a number of sub-problems. The research objectives described in Section 1.2 go some way to identifying the major sub-problems contributing to the solution to the central problem of this thesis, and these sub-problems can be again described in terms of combinations of existing research areas. Specifically, the research areas of semantic representation, psychological combination, semantic comparison, relation extraction, and their application to the domain of Cultural Heritage are examined throughout this thesis. While each of these areas have been the subject of extensive scientific research over many decades, their use in combination has been relatively limited, particularly with relation to the Cultural Heritage domain.
Chapter 2: Background

2.1 Background Overview

The notion of conceptual combination (i.e., semantic relatedness and relationship identification) has been thoroughly explored in relation to cognitive science (Wisniewski 1996; Estes 2003b; Komatsu 1992), however, the use of semantic networks to represent conceptual knowledge introduces a complicating factor: that of diverse conceptual representation of concepts (Thagard 1996; Weyer and Borning 1985). The representation used to describe a concept will differ from network to network. This can be due to the domain of application of the semantic network (e.g., the representation of a protein in a bio-medical ontology is different to the representation of that same entity in an encyclopedic semantic network), or the desired structure of the network (e.g., a strict is-a hierarchy versus a broader categorisation scheme). Semantic networks may represent all entities in the network using a common conceptual structure, or each entity may use a less structured representation in order to ensure no information regarding the entity is omitted. These two approaches both have advantages and drawbacks: a single structure used for each entity allows entities to be compared based on the differing or overlapping content of attributes, but entities which cannot be completely described using the single conceptual structure will have information omitted. Representing each entity using a structure that accurately encompasses its content without redundancy or omission of non-applicable fields, allows a semantic network to accurately represent its content. However, this differing entity structure makes direct comparison of entities (using attributive comparison) difficult. A full discussion of conceptual representation using semantic networks and ontologies takes place in Section 2.3.

Representing relationships between concepts using a structured semantic network also encounters this problem: using a defined set of relationships limits the ability
to represent all possible connections between concepts (e.g., unforeseen relationships, or relationships unique to a pair of concepts). The heterogeneity of conceptual and relational representation in collections of knowledge with diverse conceptual structure can result in duplication of content (Hakimpour and Geppert 2001) or concepts being incorrectly or incompletely described (Hakimpour and Geppert 2001; Kashyap and Sheth 1996). Less structured representations of knowledge (e.g., wikis) allow for diverse representations of conceptual structure and the relationships that exist between these concepts, as their construction is an evolving process (Mathes 2004). Multiple approaches have been made towards the identification of semantic relationships between entities in Wikipedia, however the majority of these approaches rely on a defined set of relationships to find conceptual combinations between entities (e.g., Nastase and Strübe (2008)). An analysis of conceptual combination and the representation of semantic relations with respect to diverse collections of knowledge is performed in Section 2.4.

The diversity of content present in large, cross-domain collections of knowledge presents further advantages in representing the contextual diversity of entities and the relationships between them. There exist multiple approaches to identifying relationships between conceptual entities. These approaches utilise a variety of techniques in the identification of relationships such as the recognition of textual patterns and occurrence statistics. The use of a semi-structured knowledge base, as is done in this thesis, introduces additional elements that can be used as the basis of relationship identification, however, complicating factors (such as non-uniform relationship representation) are introduced. Section 2.5 describes existing approaches to relation extraction, how they apply to diversely structured semantic networks, and how I
address current problems with the identification of relationships between entities in heterogeneous semantic networks.

In order to demonstrate the application of semantic relatedness and semantic relations between entities with diverse conceptual structure, I briefly describe the fields of User Modelling and Recommender Systems in Section 2.6. These fields of research are referred to throughout this thesis due to the relevance of the calculation of relatedness between entities for use as the basis of content-based recommender systems, and for the use of extracted relationships in describing the conceptual links between entities for the purpose of building trust in recommender systems. I also describe how the research presented in this thesis contributes to the field of Recommender systems.

Finally, I describe the existing techniques used in tasks of relatedness calculation and relation extraction. Specifically, the use of gold standard data-sets and appropriate measures used to determine correlation with gold standard, and the use of manual annotation techniques in evaluating results of natural language processing tasks, relevant to the methods in this thesis. The discussion of evaluation techniques and their appropriateness to the methodology of the work in this thesis takes place in Section 2.7.

2.2 Domains

When considering the conceptual diversity of a collection of entities, the domain of which the entities are members can have an effect on what components of the entities are considered. For example, when considering a collection of art works, it may be advantageous to be able to identify relationships based on similarities on composition, medium and artist. If this same collection of art works is placed in the
context of a larger collection of entities (e.g., a collection of geological specimens) the
relationships between these entities may be more diverse and hence require a different
conceptual representation to sufficiently capture the relationships between entities
(e.g., the identification of natural pigments used in for painting). A collection of
entities relating to a central theme or purpose is referred to as a domain. A definition
of the term domain is provided by Magnini et al. (2002): a set of words between which
there are strong semantic relations. I extend this definition to a broader conceptual
context by defining a domain as a collection of concepts that are semantically related,
or occur in similar contexts.

The advantage in only considering a restricted domain of entities is the content
of the entities and the relationships between them may be described using a struc-
ture that is entirely relevant to the entities within the domain, without regard for
misrepresenting entities outside that domain (as they are not included). This allows
entities to be described using only properties relevant to the domain (and hence the
entities), and a closed set of relationships relevant to the domain (Guizzardi 2006).
A limitation of this limited consideration of the content of an entity and its collection
of relationships occurs when considering entities outside this domain: the comparison
of an extra-domain entity to an in-domain entity requires a reconsideration of the
content of the in-domain entity, as the extra-domain entity may possess an entirely
different conceptual frame. For example, the domain of medical knowledge contains
many sub-domains, including patient records, medicines and their effects, and dis-

eses. Entities within these sub-domains possess differing conceptual frames from
those of other sub-domains, but possess relationships to entities from cross-domain
collection of entities (e.g., a patient may be afflicted by a disease, or a disease may
be treated by a medicine). Two major approaches to resolving conflicting conceptual
representation of entities where combination is required is to have a secondary system of inter-domain relationships (e.g., biomedical: Humphreys and Lindberg (1993); Cultural Heritage: Lagoze and Hunter (2001)), and the use of broad conceptual frames that account for a large range of conceptual diversity and inter-entity relationships (e.g., Doerr (2003)). The specific advantages and disadvantages of each of these approaches is further discussed in Section 2.4. The following section examines domain diversity and conceptual diversity among within the Cultural Heritage domain.

2.2.1 Museum Domains

The domain of Cultural Heritage refers to the body of knowledge surrounding human culture and history. This can include historical records such as those kept in archives, repositories of knowledge such as libraries and collections of physical objects such as the contents of a natural history or art museum. This domain also possesses multiple sub-domains of Cultural Heritage entities. For example, the conceptual content of a collection of art works and a collection of geological specimens differs, and the inter-entity relationships that apply to these particular sub-domains also differ (e.g., similar artistic composition versus geological formation).

The incorporation of multiple Cultural Heritage sub-domains into a broader collection of Cultural Heritage entities is necessary when considering the interaction of cross-collection interaction and comparison. The direct comparison of entities from differing CH sub-domains requires reconsidering the content of entities and the types of relationships which pertain to them. Cultural Heritage collections that contain entities from multiple sub-domains require greater consideration of applicable relationships between entities. This thesis makes use of the sub-domain of Cultural Heritage surrounding museum exhibits and their collections.
Digital museum catalogues (described further in Section 2.3.4) contain many millions of entries relating to exhibits which may be on display at a museum, or in storage. These catalogues provide a system for storing the knowledge associated with exhibits in a collection, so that the content of an entire collection may be accessed and examined as desired. A museum catalogue can be comprised of entities from a single sub-domain, such as art works, and thus utilise a description format that contains conceptual structure and relationships only relevant to the content of the catalogue (e.g., only considering qualities such as artist, medium, etc., in an art museum). With broader exhibition collections, a cross-domain consideration of the properties of exhibits is necessary to identify commonality and relationships between entities (Doerr 2003). For example, the catalogue of a natural history museum may contain art works, fossils, technological exhibits and living exhibits (e.g., an ant colony). The range of potential relationships between these entities is greater than that of a single domain (Sowa 1992). A cross-domain museum exhibit catalogue must contain broad conceptual frames to sufficiently diverse entities, and a large enough set of relationships to sufficiently describe the interaction between entities.

In this thesis, I utilise a cross-domain collection of entities for the purpose of discovering novel relationships between entities within a broad collection of Cultural Heritage entities. The diversity necessary to describe the content of entities and the relationships that exist between them requires the use of a representational framework that does not restrict the description of the domain. The examination of the approach to conceptual representation of diverse collections of entities is presented in the following section.
2.3 Conceptual Representation

Semantic networks are used as a representation of human cognitive organisation of knowledge (Sowa 1992). The placing of cognitive entities into conceptual frames allows for organisation and comparison (Thagard 1996; Beckwith et al. 1990). A common example used is that of a university student using a conceptual frame to represent each subject in which they have enrolled. Each subject will have common details such as a course code, title, lecturer, tutor and timetable (Thagard 1996). Semantic networks expand on this notion of conceptual frames, by representing the relationships that exist between each concept as links in a network.

*Semantic Networks* typically refer to a computational representation of a domain of knowledge. This representation is comprised of a set of nodes corresponding to concepts within the domain, and these nodes are placed into a graph structure that identifies the interactions, or relationships, between the concepts (Sowa 1992).

As stated at the beginning of this chapter, semantic networks are constructed for the purpose of representing a structured, machine-readable collection of knowledge. This collection of entities can be restricted to a single domain such as art collections (e.g., Iconclass: Aroyo et al. (2007)) and biomedical knowledge (e.g., UMLS: Humphreys and Lindberg (1993)), or can encompass a broad range of knowledge, encompassing multiple domains knowledge (e.g., WordNet or Wikipedia).

An organised representation of knowledge provides a platform for the cataloguing of knowledge, and the automated analysis of the structure of this knowledge. By utilising network or entity construction, semantic networks allow for systematic and repeatable methods for conceptual comparison (Lehmann 1992).

The use of each network varies based on its content. For example, a lexical ontology such as WordNet is used in tasks such as Word Sense Disambiguation (described
in Section 2.3.1) to identify the semantic relatedness of words. A Cultural Heritage network such as Iconclass (Wang et al. 2008) which catalogues art entities (such as paintings, artists, sculpture, etc.) can be used to construct a collection of art exhibits revolving around a certain theme or interest.

The classical method of conceptual representation takes the position that a conceptual entity is defined as a collection of attributes that describe all aspects of its nature (Komatsu 1992). This notion of single entities defined as a set of attributes, and organised in an inheritance hierarchy (e.g., the Linnaean Taxonomy of biological classification), has been superseded by more comprehensive systems of conceptual representation that describe a multi-layered network of relations and diverse representation (Komatsu 1992). Depending on the domain of the semantic network, a single entity frame and a defined set of relationships can be a sufficient representation. For example, WordNet (Fellbaum 1998) contains a single representation for each entity (a synset), and a finite set of relationships present between these entities. This representation provides a simple definition for each entity, describing what each concept represents conceptually, but lacks any description of how it interacts with other entities. As the presence of complex interactions between entities is one of the elements required for the experiments performed in Chapter 4 and Chapter 5, simple definitions of entities organised using a small, finite set of relations unsuitable for use in this thesis.

More diverse domains of knowledge require representations that allow for different conceptual structures. Museum collections contain exhibits with vastly different conceptual structures. For example, a biological specimen, a painting and a steam car may have some attributes in common: the discoverer of a biological specimen may be considered analogous to the painting’s artist, or the steam car’s builder,
but biological classifications have no relevance whatsoever to the painting or steam car. Diverse collections of entities require representations that are able to capture this diversity. Ontology models such as the ABC Ontology (Lagoze and Hunter 2001), a Cultural Heritage representation framework, allow interoperability between multiple diverse catalogues, and semantic network models such as the CIDOC Conceptual Reference Model (Doerr 2003) provide a facility for describing diverse collections and the links between these entities.

The diversity of structure of concepts present in encyclopedic resources (e.g., Encyclopedia Britannica or Wikipedia) necessitates that a non-uniform structure be used to represent conceptual entities and the content associated with them (Weyer and Borning 1985). In the remainder of this section, I describe two semantic networks used in this thesis: WordNet (Section 2.3.2) and Wikipedia (Section 2.3.3). WordNet is a lexical semantic network which has been explored extensively with regards to conceptual combination, and Wikipedia is an encyclopedic semantic network that utilises a semi-structured conceptual representation.

### 2.3.1 Word Sense Disambiguation and Polysemy

One subtask of Computational Linguistics from which this thesis draws extensively, is the task of Word Sense Disambiguation or WSD. Word Sense Disambiguation is the task of identifying the specific sense of a word being used in a piece of text. For example, in the sentence *The man swam to the bank* the sense of *bank* is most likely the sense referring to the edge of a river or body of water, rather than the financial institution. Computational methods of WSD aim to automatically identify the sense of a word as it appears in text by using knowledge relating to the word’s context.

A word may contain multiple meanings, which may simply be homonyms of one
another and completely unrelated in meaning (e.g., a river bank and a financial bank), or their meanings may be closely related. When multiple senses of the same word have related meaning these senses are termed to be polysemes. For example, bank (the financial institution), and bank (a coin bank, e.g., a piggy bank) are polysemes, as are wood (the substance of which a tree is made) and wood (the geographical area containing trees). This is because their meanings are similar: a river bank and a financial bank are not polysemous as their meanings are unrelated, but are homonyms as they are simply spelled the same.

In this thesis, I adapt the principle of word sense disambiguation to identifying museum exhibit relatedness. Specifically, WSD techniques are used to determine the connections that exist between museum exhibits. This adaptation is described further in Chapter 4. The foundation of this principle is that museum exhibits can be represented as conceptual entities. This is demonstrated by the use of Cultural Heritage ontologies and databases to represent museum exhibit collections (see Section 2.3.4). As museum exhibits are information rich entities, they contain many potential connections to other concepts. Identifying the reason for two exhibits being related to one another can be seen as analogous to mutual disambiguation of two words. For example, an exhibit describing immigration and an exhibit describing gold mining may be related to one another as the discovery of gold in many areas of the world has lead to increased immigration to that area, i.e., a gold rush. In this respect, an exhibit is a conceptual entity that is comprised of multiple facets, analogous to the representation of a word as being composed of multiple different senses (synsets in the case of WordNet).

The use of WSD methods to calculate relatedness between a corpus of museum
exhibits is an approach that has not been explored prior to the research presented in this thesis.

2.3.2 WordNet

WordNet (Fellbaum 1998) is a lexical semantic network that is comprised of a hierarchy of synsets. A synset is a set of word senses that share a common meaning. For example, the synset describing an automobile is comprised of the words *car, auto, automobile, machine, motorcar*, and defined as “a motor vehicle with four wheels; usually propelled by an internal combustion engine”.¹ WordNet organises synsets using an inheritance hierarchy of *hypernym* or *is-a* relations (e.g., a *dog is-a canine*).

In WordNet, a word is represented by a collection of word senses, with each word sense being a member of a synset. For example, *car* contains multiple word senses, one of which is the aforementioned *automobile* sense. Other senses of car include the sense referring to *railway car, elevator car* and *cable car*.

In addition to the hypernymy hierarchy, synsets are linked to one another using relations such as meronymy and holonymy. WordNet also contains verb, adjective and adverb senses, which are linked using separate relationships (e.g., *pertainymy* and *antonymy* in the adverb section of WordNet). In this thesis, only the noun class of WordNet is examined.

The use of hierarchical inheritance creates a tree structure with very specific terms at the leaves of the tree, and general terms near the root of the hierarchy.² The hyper-

¹This definition is the definition used in WordNet-3.0. All examples presented in this thesis are done using the content and structure defined in WordNet-3.0.

²Strictly speaking, the WordNet hierarchy is not a tree as it contains multiple-inheritance. However, it’s hierarchical structure, the use of a single root node, and terminal or leaf nodes are all properties it shares in common with trees. In this case, WordNet can be described as a semantic network with tree-like properties, or a Directed Acyclic Graph (DAG), but for the sake of simplicity I use the term tree to describe its structure.
nym hierarchy tree structure contains 25 base concepts as its most general concepts: synsets such as *entity* are the most generalised forms of conceptual representation, and thus all synsets in WordNet are subtypes of these 25 base concepts. This hierarchical structure creates a system of conceptual inheritance between children and parent nodes (Miller 1990). For example, the immediate hyponym of *collie* is *dog*. By virtue of being a *dog*, the *collie* also inherits all properties of *dog*’s ancestors: *canine*, *carnivore*, *placental mammal* and so on. The interaction between hypernymy and inheritance is further expanded upon in Section 2.4.1.

In WordNet, a single source for the conceptual structure of the English language has been established (Beckwith *et al.* 1990; Miller *et al.* 1990). As a large scale, comprehensive lexical semantic network, WordNet is used for many diverse purposes involving human languages, from machine translation (e.g., in the PANGLOSS system: Knight and Luk (1994)) and lexical parsing (e.g., as used in Shi and Mihalcea (2005)) to text classification (e.g., Scott and Matwin (1998)). The application of WordNet most relevant to this thesis is that of word sense disambiguation (WSD) and, more specifically, semantic relatedness. The process of determining a numeric measurement of relatedness of word senses has extensively built upon the relations within WordNet and its lexical hierarchy (Budanitsky and Hirst 2005). Semantic relatedness as a product of conceptual comparison, including its relation to WordNet, is explored in Section 2.4.

### 2.3.3 Wikipedia

The majority of semantic networks comprise information about a single domain, but there exist a few that cover many domains. Large scale digital semantic networks have been suggested to be a valuable resource as far back as 1945. Indeed,
Bush (1945)’s dream of *Wholly new forms of encyclopedias [...]*, ready-made with a *mesh of associative trails running through them*, has largely been realised with the construction of electronic encyclopedias (Weyer and Borning 1985) and the digitisation of existing encyclopedias such as Encarta and Encyclopedia Britannica. More recently, user-constructed encyclopedias such as Wikipedia have become popular, allowing the layperson to have a direct hand in editing the content and structure of the encyclopedia.

Encyclopedic semantic networks are appropriate for the methods presented in this thesis as they represent a broad set of knowledge, not restricted to a specific domain, and they are able to represent entities without a common conceptual frame. An ontology of entities with a uniform entity structure (all entities are described using an identical structure) allows the entities to be easily compared by examining their attribute-value overlap. However, entities that cannot be defined in terms of a rigid set of attributes cannot be appropriately placed into such a network, or easily compared to any entity within the semantic network (Komatsu 1992).

Wikipedia is an online digital encyclopedia created using collaborative editing. Anyone may edit Wikipedia, however the Wikipedia community constantly reviews additions and changes to articles to maintain the quality of the encyclopedia as a whole. Wikipedia editing guidelines stipulate that articles must be without bias and adhere to a neutral point of view style of writing (Wikipedia 2011f), and any statements must cite reliable references (Wikipedia 2011c). Wikipedia initially lacked an organisational structure, but introduced a system of category membership in order to organise increasing numbers of articles. Due to the multiple and diverse editorship the structure and content of Wikipedia is constantly evolving, resulting in a unique snapshot of the encyclopedia at any given time. Wikipedia maintains the entire
edit history of all articles, and periodically creates dumps of the entire content of Wikipedia (described further in Section 3.2.1).

Wikipedia has become the go-to standard for semantic qualities of exploring encyclopedic information (Zesch et al. 2007b; Medelyan et al. 2009). It is a digital encyclopedia based on the wiki framework, a framework built around collaborative editorship. This is described variously as a collaborative ontology and as a Folksonomy (Mathes 2004). This is one aspect of Wikipedia that provides an advantage over professionally curated semantic networks such as WordNet: the large editorship of Wikipedia has created a knowledge base that covers a greater range of content than many other publicly accessible knowledge bases (see Section 6.1 for a comparison between Wikipedia and other knowledge bases).

Collaborative editing can be a boon and a bane to a comprehensive network of knowledge (Kittur et al. 2007): multiple editors can work together in a continual process of peer review to reach consensus on the best representation of a concept (Kittur and Kraut 2008). Conversely, lax reviewing of others’ work can lead to false statements being read by large numbers of people and being regarded as the truth (Kittur and Kraut 2008). In one case, false information found on Wikipedia lead to widespread reporting of the involvement of a former journalist in the Kennedy Assassination (Seigenthaler 2005). Multiple studies have been performed with regard to Wikipedia’s accuracy and resilience to malicious editing (e.g., Giles (2005), Viégas et al. (2007)). Processes such as restricted editorship of contentious articles, increased monitoring on articles about living people and current events, and communities of editors focussing on specific topics or projects in Wikipedia, have greatly reduced the prevalence of erroneous data in Wikipedia (Viégas et al. 2007).

A concept in Wikipedia is called an article, at its most basic level it can be
considered an encyclopedic entry described by a body of text. This basic textual description of the concept is enhanced through the use of inter-article links and a system of category membership. Wikipedia stipulates that each article must be a member of at least one category, and in turn each category must belong to at least one, more generalised, category (Wikipedia 2011b). For example, the Wikipedia article on *Art*\(^3\) contains many thousands of words of text, hundreds of links to other articles, and is a member of the categories *Aesthetics*, *Arts* and *Visual arts*.

As each category is required to be a member of another more general category, this creates a hierarchical category graph, similar to the hierarchical organisation of WordNet. This inheritance hierarchy is a branching structure, with categories being subsumed by increasingly less specific categories. The category membership of all Wikipedia articles can be traced back to the two top level categories: *Fundamental* and *Main topic classifications*. An example of a section of this hierarchy is shown in Figure 2.1. The adaptation of the Wikipedia category hierarchical into a computational semantic network is described in detail in Section 3.2.

---

\(^3\)As contained in the September 9th, 2009 dump used in this thesis.
As Wikipedia articles are required to be part of one or more categories, and categories themselves are subsumed by more general categories, the category network in Wikipedia forms a hierarchical structure of categories with articles as its leaf nodes. There is also a requirement that there can be no cycles in this category subsumption, however, as is explained in Chapter 4, this is not strictly adhered to.

This hierarchy has been used as a parallel to other semantic networks in terms of functionality and content (Sarjant et al. 2009). With respect to this thesis, the most relevant of these comparisons is to WordNet.

Wikipedia handles polysemy differently to WordNet: the financial institution sense of bank and the physical building sense of bank are represented as separate word senses in WordNet, but are both represented in the same Wikipedia article. Due to the closeness of meaning (and overlap in content), these entities are described in a single article in Wikipedia, rather than as separate senses as in WordNet. Whereas a word in WordNet may have multiple senses that possess fine-grained differences, Wikipedia treats closely related entities (with overlapping content) as a single article. In cases when a word can potentially have multiple meanings (i.e., homonymy), Wikipedia will collect all potential meanings of the word into a disambiguation article. This article lists all potential meanings of a term, and their respective articles. Articles with similar titles must be unique enough to be distinguished from other articles with a similar topic; if two articles share too much content, or the same topic (as with polysemy), Wikipedia guidelines stipulate that they must be merged (Wikipedia 2011e).

As Wikipedia has a larger editor base, and significantly more information per concept than WordNet, it has recently been explored as an alternative semantic network (Auer and Lehmann 2007; Milne and Witten 2008a; Ponzetto and Strübe 2007b;
Wubben 2008; Zesch et al. 2007b). Part of the attraction of Wikipedia is the multitude of relationships between concepts. Links between articles can be part of a hierarchy (Ponzetto and Strübe 2007b; Zesch et al. 2007a), presented as attributive qualities (Auer and Lehmann 2007), or as in-line links in the body of the article (Milne and Witten 2008a; Wubben 2008).

Wikipedia’s diversity of representation, detailed descriptions of entities, hierarchical classification system, explicit and implicit relations between articles, as well as its sheer size make it a desirable resource for experiments of semantic relatedness and relation extraction over diverse collections of entities. The application and use of Wikipedia in calculating semantic relatedness is covered in Section 2.4, and its use in semantic relation extraction is covered in Section 2.5. The adaptation of Wikipedia as a computational resource in this research is described in Section 3.2.

2.3.4 Cultural Heritage Ontologies

Semantic networks such as the CIDOC Conceptual Reference Model (CIDOC-CRM) or the Iconclass art cataloguing system allow curators to stitch together a vast network of artifacts into segments of knowledge that represent the complex interactions of timelines, events and themes revolving around those artifacts (Doerr 2003). A key feature of these frameworks is their ability to represent collections from multiple physical locations that share a common thread while being displayed in separate collections in different museums.

The advantage of cataloguing a homogeneous collection of objects, such as books or art works, is that the objects fit within a common conceptual frame. For example, all paintings will have artist, medium, style and date attributes. This allows for easy comparison of entities by examining the attribute overlap (a process of attributive
comparison, as described by Estes (2003a)). Artifacts within natural history or technology museums lack the ability to be reliably described within a common frame. For example, one may be able to identify many attributive similarities between a nautilus shell and a fossilised trilobite (e.g., both are extinct prehistoric animals), but there are far fewer similarities between the structure of a steam engine and the nautilus shell (although one may consider the property that both the steam engine and the nautilus have a hard outer shell a relating factor, this is tricky to identify when using a simple attributive representation to describe each entity).

This inability to represent entities with diverse physical structure is a drawback of multiple museum catalogue technologies. The KE-EMu ontology is a template that uses a collection of attributes to describe museum exhibit entities, and a customised set of relations to relate exhibits to one another. The KE-EMu ontology can be adapted to a particular collection, and defines each artifact in a collection as a set of attributes that describe the artifact (Tedd and Large 2005). This approach works well for biological or natural science collections (e.g., as used in Tennyson and Bartle (2008)), but for diverse collections rigid definitions of artifact structure can mean that many artifacts are poorly defined (Keene 1998; Eklund et al. 2009).

The CIDOC Conceptual Reference Model is an ISO standard of semantic representation of Cultural Heritage entities such as artifacts, Cultural Heritage institutions and geographic locations (Doerr 2003). Entities are defined by their relationship to defined data types (e.g., events: an exhibit describing a historic battle; or dates: a painting was created in a certain year). When an entity requires additional description not provided for by existing data types, network curators can create new subclasses of abstract data types. The existing data types are comprehensive enough to allow entities such as historical figures, timelines and historical events to be represented,
and any relationships that may tie them together to be specified. This formalism allows for a highly detailed representation of Cultural Heritage entities, but requires a new editor to spend a lot of time becoming acquainted with the description language, making extensibility an issue.

Many museums are moving away from using these highly structured attributive definitions of artifacts and exhibits, preferring to use author and user annotations to layer additional information and meaning to existing descriptions of exhibits (Chan 2007; Eklund et al. 2009; Trant and Wyman 2006). In this respect many online museum collections have taken on the aspect of a Folksonomy, a community edited collection of linked information (Mathes 2004). This is largely due to an emerging trend of museums opening up their collections to users through online collection catalogues.

The comprehensive representation of diverse collections of museum exhibits has been approached from multiple angles: specialising the domain of representation in order to create an accurate entity structure and organisation (e.g., Iconclass), or using a generalised entity structure and set of relations (e.g., the CIDOC-CRM, KE-EMu ontologies, and the ABC ontology model). These representation suffer from the inability to accurately represent the intricacies of relations between concepts (in the case of broad coverage semantic networks, due to the need to maintain a generalised structure that can represent heterogeneous concepts), or from an over-specific conceptual representation (in the case of domain specific semantic networks). Unstructured representations of museum exhibit entities have the potential to describe much more diverse qualities of exhibits and their relations with other exhibit entities, but the lack of structure makes identifying these relations a difficult task. In this thesis I use an unstructured semantic network (Wikipedia) to represent a collection of museum ex-
hibit entities, and address the problems of entity comparison and relation extraction with this unstructured representation.

### 2.4 Conceptual Combination

Conceptual comparison is the act of comparing two conceptual entities to one another for the purpose of identifying relations or similarities (Thagard 1996; Wisniewski 1996). The particular aspect of conceptual combination being examined in this thesis is the specific combination of noun-noun concept pairs. The process of conceptual combination of noun-noun pairs is the cognitive process that occurs when a person regards two conceptual entities (represented by nouns), and attempts to identify how the two nouns relate to one another (Estes and Glucksberg 2000; Estes 2003a). The combination of two nouns can have the effect of the creation of a new concept (e.g., *ostrich burger*), or the comparison of the two nouns (e.g., *dog cat*), however, knowing which of these two combinations will occur depends on the particular cognitive process of the person performing the conceptual combination (Estes 2003b).

There have been several examinations of the cognitive process that takes place when comparing conceptual entities with common as well as diverse conceptual structure (e.g., Bassok and Medin (1997), Estes (2003b) and Wisniewski (1996), Wisniewski and Bassok (1999)). Entities with a common conceptual structure or function (e.g., *car* and *bicycle*) are compared via a process of *comparison*, while the conceptual combination of entities with highly different structures (e.g., *forest* and *hunter*) is termed as a process of *integration* (Bassok and Medin 1997; Wisniewski 1996). Building on work in Wisniewski and Bassok (1999), Estes (2003b) found that when
entities with a common conceptual structure were compared, annotators were likely to identify points of difference rather than commonality (e.g., a car has four wheels whereas a bike only has two), whereas entities with different conceptual structures were combined, annotators were more likely to use an integrative process, picking up on common threads that related them despite their differences (e.g., a hunter hunts in a forest).

This appears to be a clearcut definition, however, there lies a great deal of ambiguity in identifying concepts that possess a common conceptual structure. For example, a layperson may regard a *car* and a *bicycle* as possessing a similar conceptual structure, as they are both forms of wheeled transport, but an automotive engineer may place *bicycles* and *cars* into completely different conceptual frames.

In tasks of computational linguistics where lexical combination is considered, the conceptual combination of words is measured through semantic relatedness. Semantic relatedness is a quantification of how closely related two terms are: how often the two terms refer to, or occur in, the same context, whether they pertain to a common theme, or how close in meaning they are (Budanitsky and Hirst 2005). The calculation of the relatedness is on a continuous scale, for example, on a scale from 0 to 1 (0 being unrelated, and 1 being identical) the terms *tiger* and *cat* would obtain a relatedness score close to 1 as they are closely related as they are biologically related, and a tiger is a sub-type of cat, whereas the terms *cord* and *noon* would obtains a score near 0 as they do not share a related meaning, and are unlikely to occur in the same context.

Semantic similarity is a special sub-case of semantic relatedness, only examining the closeness of meaning of terms. This is again measured on a continuous scale, ranging from synonymous in meaning to completely unrelated in meaning (Budanitsky and Hirst 2005). Examples of highly similar word pairs include *dog* and *canine,*
automobile and car, and wizard and magician (Rubenstein and Goodenough 1965). Semantic relatedness is less tangible as it requires knowledge of how the concepts interact. For example, the terms beach and swimming are not semantically similar (the action of swimming is very different in meaning from the location, beach), but are semantically related as swimming is an activity that commonly takes place at a beach. In this respect, semantic relatedness takes into consideration how concepts interact with one another: concepts that occur in similar contexts and possess relationships with one another, without necessarily being related in meaning, are considered semantically related. In the previous example, the terms beach and swimming may be considered semantically related.

Multiple approaches to determining semantic relatedness and similarity exist. These range from the analysis of the English language as it is used (e.g., the statistical co-occurrence of words in a piece of text (Resnik 1995; Jiang and Conrath 1997)), to the analysis of the conceptual organisation of word senses in a semantic network such as WordNet (Rada et al. 1989; Palmer and Wu 1995; Hirst and St-Onge 1998; Leacock et al. 1998). The frequency of term co-occurrence in a body of text can be used to determine how closely related two words are: if two words frequently occur in the same context, they have a high likelihood of being related to one another (Rijsbergen 1977). These co-occurrence statistics can be gathered from a simple textual corpus such as SemCor (Miller et al. 1993), or a segmented corpus annotated with individual word senses such as the Brown Corpus (Francis et al. 1982). At the other end of the spectrum is to examine the structure of the language itself using a semantic network of concepts (Beckwith et al. 1990). Much of the work regarding relatedness using semantic networks has been performed using WordNet as the semantic network, however, there also exist instances using the computational knowledge base
of Cyc (e.g., Curtis et al. (2006)), and more recently Wikipedia (e.g., Gabrilovich and Markovitch (2007), Ponzetto and Strübe (2007c), Wubben (2008)). There also exist measures of semantic relatedness which utilise both approaches, weighting the nodes within these semantic networks with their relative probability of occurrence in a textual corpus (e.g., Resnik (1995)).

Semantic networks that utilise a simple conceptual definition for entities and organise entities using an inheritance hierarchy, such as WordNet, provide a means for determining the semantic similarity of words, as the hierarchical closeness indicates closeness of meaning (Beckwith et al. 1990). However, using a restricted set of relationships, WordNet does not contain the facility to describe complex interactions with other entities, and in this respect is unable to completely take advantage of the interactions between entities for the purpose of calculating semantic relatedness.

Wikipedia provides another basis for computing the semantic relatedness of concepts: the diversity of links between articles have been used to calculate the semantic relatedness of Wikipedia articles. Links can be used to represent the simple attributive qualities of an article (as defined by a template), or they may appear in the body text of an article, describing a complex interaction between two articles. As Wikipedia articles are unstructured, the use of these links varies widely from study to study. These links have been used to determine the relative popularity of an article (Milne and Witten 2008a), the validity of article content (Wu and Weld 2008), or the semantic distance between articles by traversing inter-article links (Wubben 2008). Alternatively, they have been matched to existing patterns to extract semantic relations (Auer and Lehmann 2007).

Wikipedia is a valuable resource for analysing the conceptual combination between conceptual entities, using a structured organisation of categorisation and able to de-
scribe complex interactions between entities. The unstructured nature of Wikipedia articles makes extracting elements such as semantic relations between articles unsuited to pattern matching approaches: while there will may exist semantic relations that adhere to predefined patterns (e.g., such as those identified by Auer and Lehmann (2007) and Ruiz-Casado et al. (2006)), there exist many relations that cannot be identified using this methodology. Section 2.4.2 describes how Wikipedia has been adapted in multiple studies to computing the semantic relatedness of entities. The extraction of specific semantic relations from Wikipedia is further described in Section 2.5.

2.4.1 Lowest Common Subsumer

The node representing the domestic dog in the WordNet hierarchy is many levels below its most general ancestor (the entity node), but it still possesses many child nodes, e.g., the nodes puppy, hunting dog and Welsh corgi. In this case, dog, domestic dog, canis familiaris is the common ancestor of these three nodes (as well as all other children, grand children and so on). The children of hunting dog (e.g., courser, hound and terrier) also possess dog as a common ancestor, however, hunting dog is the more specific ancestor, as it is deeper in the hierarchy (i.e., further from the most general node, entity), and is termed to be the Lowest Common Ancestor, or Lowest Common Subsumer (LCS).

One method of selecting the LCS from a group of subsumers is to select the deepest subsumer in the hierarchy, i.e., the most distant from the root of the hierarchy. Distance is often defined as the count of the number of intermediate nodes between a starting node and the unique root (Rada et al. 1989) (this count includes the unique root and the starting node). This requires a unique root node that subsumes all
Chapter 2: Background

Figure 2.2: The Least Common Subsumer (LCS) of dog varieties as defined by WordNet. In this case, the LCS of \textit{guard dog} and \textit{terrier} is \textit{dog, domestic dog, canis familiaris}, but the LCS of \textit{hound dog} and \textit{terrier} is \textit{hunting dog}.

nodes in the hierarchy. In hierarchies where there exists no unique root node (e.g., WordNet), an artificial node that contains all existing top level nodes as its children is inserted as the root (e.g., Resnik (1995), Lin (1998)). In WordNet, the hypernym paths of the words \textit{cat} and \textit{dog} are:

- **Cat**: \textit{domestic cat} $\rightarrow$ \textit{cat} $\rightarrow$ \textit{feline} $\rightarrow$ \textit{carnivore} $\rightarrow$ \textit{placental mammal} $\rightarrow$ \textit{mammal} $\rightarrow$ \textit{vertebrate} $\rightarrow$ \textit{chordate} $\rightarrow$ \textit{animal} $\rightarrow$ \textit{organism} $\rightarrow$ \textit{living thing} $\rightarrow$ \textit{whole} $\rightarrow$ \textit{object} $\rightarrow$ \textit{physical entity} $\rightarrow$ \textit{entity}.

- **Dog**: \textit{domestic dog} $\rightarrow$ \textit{canine} $\rightarrow$ \textit{carnivore} $\rightarrow$ \textit{placental mammal} $\rightarrow$ \textit{mammal} $\rightarrow$ \textit{vertebrate} $\rightarrow$ \textit{chordate} $\rightarrow$ \textit{animal} $\rightarrow$ \textit{organism} $\rightarrow$ \textit{living thing} $\rightarrow$ \textit{whole} $\rightarrow$ \textit{object} $\rightarrow$ \textit{physical entity} $\rightarrow$ \textit{entity}.

In this case, the nodes \textit{carnivore}, \textit{placental mammal}, \textit{mammal}, \textit{vertebrate}, \textit{chordate}, \textit{animal}, \textit{organism}, \textit{living thing}, \textit{whole}, \textit{object}, \textit{physical entity}, and \textit{entity} all subsume the nodes \textit{domestic cat} and \textit{dog}. However, the LCS of these two nodes is \textit{carnivore} as it is furthest from the root node (in this case, the \textit{entity} node).

Formally, the least common subsumer of two nodes ($c_1$ and $c_2$) in a hierarchy is the node $s$, subject to the requirement that $s$ is a subsumer of the two nodes ($S(c_1, C_2)$).
and that it does not subsume any other common subsumer of the two nodes:

\[
lcs(c_1, c_2) = s, \text{ s.t., } s \in S(c_1, c_2) \cap \neg \exists s': \text{hyper}(s, s') \quad (2.1)
\]

The manner in which the LCS is determined varies from network to network. This can be due to multiple inheritance causing multiple paths from a node to the root. One method of selecting a single LCS from a group of subsumers is to select the subsumer “closest” to the pair of nodes being tested (e.g., as done in Snow \textit{et al.} (2006)). For example, the node \textit{dog} can be traced via the hypernym hierarchy to the \textit{entity} node through two paths. In addition to the path described above, it can also be traced to the root node (\textit{entity} in this case) via the path: \textit{dog} $\rightarrow$ \textit{canine} $\rightarrow$ \textit{carnivore} $\rightarrow$ \textit{placental mammal} $\rightarrow$ \textit{mammal} $\rightarrow$ \textit{vertebrate} $\rightarrow$ \textit{chordate} $\rightarrow$ \textit{animal} $\rightarrow$ \ldots $\rightarrow$ \textit{entity}.

The node \textit{domestic cat} is also subsumed by the nodes \textit{domestic animal} (an immediate parent node) and \textit{carnivore} (\textit{domestic cat} $\rightarrow$ \textit{cat} $\rightarrow$ \textit{feline} $\rightarrow$ \textit{carnivore}). In this case, \textit{domestic animal} would be the LCS of \textit{dog} and \textit{domestic cat} as it is an immediate parent of both nodes, whereas \textit{carnivore} is 2 nodes distant from \textit{dog} and 3 nodes distant from \textit{domestic cat}.

The LCS is used in multiple measures of semantic relatedness and similarity used over the WordNet hierarchy. The experiments conducted in this thesis utilise the former of these two methods: the LCS is termed to be the farthest subsumer from the root. This is in part to to identify the most specific node that subsumes both entities, but also to combat the situation where a short path to the root is accessible from both entities, thus causing the root to be identified as the LCS. In practice, the shortcut-to-root problem in the Wikipedia category hierarchy greatly effected the calculation of the LCS (further discussed in Section 4.3.2).
2.4.2 Computational Measurement of Semantic Relatedness

Early experiments with computational calculations of semantic relatedness utilised existing lexical networks such as Roget’s Thesaurus (e.g., Osgood et al. (1957)) and machine readable dictionaries (e.g., Lesk (1986)). Existing semantic networks (e.g., Roget’s Thesaurus, WordNet), computational knowledge bases (e.g., Cyc, further described in Section 6.1.3, and textual corpora (e.g., The Brown Corpus: Francis et al. (1982) and SemCor: Miller et al. (1993)) provide a basis for identifying the semantic relatedness or similarity of individual semantic concepts.

An early approach to measuring the levels of relatedness between semantic entities was identified by Osgood et al. (1957). By grading each word in terms of its EVALUATION (good-bad), POTENCY (potent-impotent), and ACTIVITY (active-passive), Osgood et al. (1957) employed a basic geometric distance metric to identify the semantic relatedness of words. This required that all words being evaluated be manually annotated in terms of the three axes of EVALUATION, POTENCY and ACTIVITY, and thus was an extremely time expensive task.

The construction of Machine Readable Dictionaries removed the need for manual annotation of individual word senses, and allowed the construction of more complex measures of semantic relatedness. For example, Lesk (1986) used the glosses of individual word senses to determine how often two word senses used common words in their definitions: word senses that contained many overlapping terms (e.g., *pine*: “kind of evergreen tree with needle-shaped leaves...” and *cone* “fruit of certain evergreen trees...”) had a higher semantic relatedness than word senses that had very little term overlap (e.g., *pine*: “waste away through sorrow and illness” and the above sense of *cone*). For example, in the glosses of the word senses *pine* and *cone*, the words *evergreen* and *tree* occur in both, indicating that they have a degree of relatedness (Lesk
1986). This is a simplistic examination of the overlap in content of terms, but has been applied over other semantic networks (WordNet) with some success (Baldwin et al. 2008; Banerjee and Pedersen 2003; Budanitsky and Hirst 2005).

WordNet is an important lexical resource that reflects the structure of the English language, but it lacks the multitude of link types found in more diverse semantic networks that facilitate calculations of the broader measure of semantic relatedness. Hirst and St-Onge (1998) explored the additional link structure available in the verb class of WordNet (using relationships such as meronymy, holonymy and antonymy, that exist among the verb senses of words in WordNet) to provide a link based method of semantic relatedness. However, due to the diverse links between synsets, which make sense for taxonomically close concepts, spurious connections can be made between concepts by following convoluted link paths and identifying unrelated items as semantically similar (Budanitsky and Hirst 2005). Hirst and St-Onge (1998) classified links between synsets in terms of “upward,” “downward” and “horizontal” with respect to the hypernymy hierarchy. The semantic relatedness of a pair of synsets was weighted by the tortuosity of the path between them: if a path was highly tortuous, the relatedness score was reduced. The more tortuous a path, the less likely that it was to be an accurate score of relatedness.

WordNet-based measures of semantic similarity utilising the hierarchical distance between nodes use the path length via the hypernymy hierarchy to determine the semantic distance between nodes (Rada et al. 1989). This calculation of path distance via the hypernym hierarchy forms the basis for many of the methods described in this section.

As there exist multiple senses of each word, identifying the semantic similarity between two words means considering all synsets that comprise both words. The
standard way to deal with this is to measure the similarity of all senses of the first word \((S_1)\) compared to all senses of the second word \((S_2)\), resulting in \(|S_1| \cdot |S_2|\) comparisons. The senses which have the highest semantic similarity are selected as the word senses. All of the following measures are measures between individual word senses.

The measure of semantic similarity devised by Leacock et al. (1998) finds the shortest path between the two synsets \((sp(c_1, c_2))\) using hypernym and synonym relationships. This path length is then scaled by the maximum depth of WordNet \((D)\), and the log likelihood taken:

\[
s_{\text{lich}}(c_1, c_2) = -\log \frac{sp(c_1, c_2)}{2 \cdot D}
\]  

(2.2)

For items that are close together in the hierarchy, the similarity score will be higher, indicating a high semantic similarity. However, this measure does not take into account the relative depth of the initial synsets, and synsets become much less specific towards the base concepts of WordNet. This inability to distinguish between general and specific concepts is a drawback of this measure.

The Wu-Palmer (Palmer and Wu 1995) measure addresses the issue of relative depth (short paths between items deeper in the hierarchy are more meaningful than short paths between items towards the top of the hierarchy) by scaling the depth of the two nodes \((\text{depth}_{c_1} \text{ and } \text{depth}_{c_2})\) by the depth of their LCS \((\text{depth}(\text{lcs}_{c_1, c_2}))\).

\[
s_{\text{wup}}(c_1, c_2) = \frac{2 \cdot \text{depth}(\text{lcs}_{c_1, c_2})}{\text{depth}_{c_1} + \text{depth}_{c_2} + 2 \cdot \text{depth}(\text{lcs}_{c_1, c_2})}
\]  

(2.3)

These methods are effective for any taxonomy provided that the each link between the nodes in the hierarchy are of equal distance. In WordNet it can be the case that
the semantic distance from one word to its direct hypernym is greater than that of another word elsewhere in the hierarchy to its respective direct hypernym. The hypernym paths for *domestic dog* and *domestic cat* given in Section 2.4.1 demonstrate the difference that exists between some elements that exist at a similar conceptual level. For example, when *carnivore* is used as the LCS, the distance from *carnivore* to *domestic cat* is greater than the distance from *carnivore* to *domestic dog* (due to the interceding *cat* node) even though cats and dogs are commonly considered to be on the same conceptual level. The unit length of hypernymy relationships in WordNet would indicate that *domestic cat* is a more specific instance of *carnivore* than *domestic dog*. A solution to this limitation of WordNet is to weight the edges of the hypernym hierarchy to reflect the semantic distance between synsets (Resnik 1995).

The frequency of use of word senses in pieces of text has been used as a basis to populate the WordNet hypernymy hierarchy with edge weights. Resnik (1995) used The Brown Corpus to identify the frequency of the nouns in WordNet, and then used these frequencies to weight the edges between nodes. Whenever a noun was encountered in the text, the number of occurrences of that word sense (e.g., *cat*), as well as the occurrences of all of its subsumers (e.g., *feline*, *carnivore*, *mammal*, etc.) would be incremented. This is due to the property that an occurrence of a *cat* is also an occurrence of a *feline* (and all subsequent subsumers). The probability of each concept was then calculated as the ratio of times that concept was counted and the total number of nouns observed in the text (*total_nouns*): \( \hat{p}(c) = \frac{\text{freq}(c)}{\text{total_nouns}} \). The semantic distance between a concept and its hypernym can be calculated by taking the difference between the two nodes’ probabilities.

Resnik (1995) found that the most effective measure of comparison using this
methodology was to adapt the information theoretic concept of Information Content as a way of weighting the edges between nodes in a hierarchy. Information Content (IC) is a measure of the log likelihood of a concept as it appears in a corpus: $IC(c) = -\log p(c)$. Less frequently occurring concepts obtain a higher IC, and more abstract concepts obtain a lower IC. This is due to the property of subsumption: general (or more abstract) concepts subsume more specific concepts. When the count of a term occurring in a corpus is incremented, all of its subsumers are also incremented.

Resnik (1995) utilised the principle of Information Content to produce a measure of semantic similarity that identified the subsumer with the greatest Information Content from the set of all concepts that subsumed the two initial concepts ($S(c_1, c_2)$) being compared:

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [IC(c)] \quad (2.4)$$

In a hierarchy with a unique root node, concepts are hierarchically subsumed by a series of super-concepts, ending with the root node. In this case, the LCS used is the deepest subsuming node in the hierarchy (see Section 2.4.1). Lin (1998) expanded on the Information Theoretic approach presented by Resnik (1995) by scaling the Information Content of each synset by the information content of their LCS:

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(lcs_{c_1, c_2})}{IC(c_1) + IC(c_2)} \quad (2.5)$$

The resulting score is the semantic distance between the two concepts and their LCS, with larger numbers representing a large semantic distance (and hence low semantic similarity), and small numbers representing high semantic similarity.
A similar approach is taken by Jiang and Conrath (1997):

\[
sim_{jcn}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \cdot IC(lcs_{c_1, c_2})}
\]  

In this case, Jiang and Conrath used SemCor to populate the WordNet synsets with their respective IC. Note, however, that the closer two concepts are in the hierarchy, the greater their similarity. Indeed, if a word is compared with itself, the similarity is infinite, as opposed to Lin’s measure which produces a score between 0 and 1.

Lesk (1986)’s method of glossary term overlap has also been used over WordNet synset glossaries to determine the relatedness of terms (Banerjee and Pedersen 2003). This approach of using definitions and textual content to determine the overlap in content is especially relevant to the problem of diverse conceptual representation as it allows the conceptual structure of a collection of entities to be diverse enough to sufficiently represent a concept, rather than subscribe to a rigid template that may not suit its conceptual structure, or may impart additional relationships that are not represented in a standardised conceptual frame.

In separate studies, Zesch et al. (2007a) and Strübe and Ponzetto (2006) demonstrate that the use of existing measures of semantic relatedness developed over a semantic network with homogeneous representation (WordNet) are able to be adapted to a network with a differing network structure. However, this transposition results in a drop in performance when used over the same evaluation data sets. For this reason, measures of semantic relatedness designed to take advantage of the diversity in structure of individual entities, wealth of content, and organisation of heterogeneous semantic networks is desirable (such as those described by Milne and Witten (2008a) and Gabrilovich and Markovitch (2007)).
Wikipedia has also been used as a basis for the calculation of semantic relatedness and semantic similarity. Its large amount of textual content and inter-article links make it valuable as a resource for calculating semantic relatedness. Strübe and Ponzetto (2006) transposed measures of semantic relatedness and similarity developed for the WordNet hierarchy for use over Wikipedia’s category hierarchy, demonstrating a noticeable drop in performance for the same data sets. Aside from this transposition, Wikipedia’s inter-article links have also provided a platform for the construction of computational measures of semantic relatedness explicitly for use over the Wikipedia semantic network.

Wubben (2008) found the path distance between Wikipedia articles using the links in the text of each article. Wubben identified all paths between articles, and selected the shortest path as the semantic distance between the articles. Two versions of this shortest path method were tested, the first using the directed link structure of Wikipedia, and only following the links in the direction that the appear within Wikipedia. The second method disregarded the direction of links. It was found that using a graph composed from undirected links provided a higher correlation with gold standard relatedness scores.

The information content based method first proposed by Resnik (1995) was given a new twist by Milne and Witten (2008a). As Wikipedia is self referential (articles link to related articles), no external source text is needed to find the referred-to-edness of a concept. Milne and Witten used an adapted information content measure that weighted the total number of links to an article from all other Wikipedia articles with the number of links between the two articles being compared ($\text{links}(c_1, c_2)$):

$$
rel_{mil}(c_1, c_2) = |\text{links}(c_1, c_2)| \times \log \sum_{a \in W} \frac{|W|}{|\text{links}(c_1, a)|}
$$

(2.7)
where \( a \) is an article in \( W \), Wikipedia. This method is known as the Wikipedia Link Vector Model (WLVM).

Gabrilovich and Markovitch (2007) created a method of semantic analysis called Explicit Semantic Analysis (ESA) derived from Latent Semantic Analysis (LSA) that takes into account mutual links and references between articles. ESA was designed as a document categorisation tool that could identify related topics to the initial query. This provided a summary of the closest related items to an initial article. As this method was scalable, the initial query could have been a single document, a single word, or an entire document collection.

The semi-structured and constantly evolving content and structure of Wikipedia allow it to be a powerful tool for the representation of diverse knowledge (Mathes 2004). However, the semi-structured nature of the content makes the problem of identifying relatedness distinct from structured hierarchies such as WordNet. The use of Wikipedia for conceptual representation and computational measurement of relatedness is an emerging field, and despite much work done towards this goal in recent years, there remain many aspects of Wikipedia to be analysed. This thesis proposes new methods for identifying the relatedness of Wikipedia articles.

### 2.5 Semantic Relations

The measurement of semantic relatedness or similarity produces a numeric score indicative of the strength of relatedness or similarity, however, it can often be advantageous to know the basis for the relatedness.

When considering semantic entities such as concepts in an ontology or real world entities (e.g., companies, locations and people), the connections between these entities
are termed as semantic relations. Semantic relations can be used as a representation of conceptual abstraction and inheritance (e.g., an is-a relation) (Brachman 1983), or composition (e.g., has-part, is-part-of) (Winston et al. 1987). Beyond these basic semantic relations, more complex interactions between concepts have been explored.

There exist two separate viewpoints on the nature of semantic relations: that all relations are able to be classified in terms of a finite set of relations, and that there exists no finite, bounded set of relations. I take the latter of these two positions: that there exist too many relations and relation frames to realistically account for, and that the relations present between Wikipedia articles are described using an unstructured approach that does not allow them to be completely identified when using rule-based approaches. This is not to say that there does not exist an argument for using pattern or rule based approaches to identifying relations between entities in Wikipedia, but that there also exist relations that do not adhere to standardised representations.

Based on the application or experimental domain, the number of relation classifications can vary from a small set of relations (e.g., Girju et al. (2007), Rosario and Hearst (2004)) to large catalogues of relations that describe the use and appearance of each relation and its arguments (Baker et al. 1998). For example, the relation between ship and silver in the sentence *The silver ship usually carried silver bullion bars, but sometimes the cargo was gold or platinum*, is a CONTENT-CONTAINER relation (with silver being the content, and ship the container). Depending on the task or experimental domain (e.g., the classification of semantic relations between nominals: Girju et al. (2007)) other sets of relations can include relations such as CAUSE-EFFECT, PRODUCT-PRODUCER, and ENTITY-DESTINATION. The silver ship example is an example of the problem of identifying the relations that exist between pairs of nouns (termed noun compounds).
Others argue that there can be no finite set of relations to which all relations subscribe (Banko et al. 2007; Wu and Weld 2010). One reason for this is the constant evolution of language: new verbs are constantly being created, and thus new ways of entities interacting with one another. The use of a predefined set of relations to which all relations subscribe presupposes that this limited set of relations will be broad enough in scope to classify any future relation that may be created. For example, the verb *grok* (a type of understanding of a genre or knowledge domain that also has nuanced connotations about the level of understanding of the grokker), is difficult to classify with respect to Girju et al. (2007) conventions, and in shoehorning it into a pre-existing category, much of the information regarding its specialised use is ignored.

There exist multiple approaches in identifying semantic relations between entities from document collections, including identifying the interactions between entities as described in bodies of text (e.g., identifying subject-verb-object phrases), and matching text to predefined patterns of relations (e.g., *X is based in Y* for a location relation). Pattern matching approaches to identifying semantic relations between concepts are insufficient to identify all potential relations between concepts due to nonstandard representation of relations between concepts in semantic encyclopedias such as Wikipedia (Krötzsch et al. 2006).

The extraction of information from unstructured text is also approached in the field of Open Information Extraction (OpenIE): the extraction of relations between entities from large bodies of unstructured text using machine learning methods (e.g., Yates et al. (2007), Wu and Weld (2010), Banko et al. (2007)). OpenIE differs from traditional approaches to Information Extraction in that it does not need a predefined set of relations between entities from which to learn patterns of relation types (Banko and Etzioni 2008). OpenIE has been used to identify factual qualities associated
with entities (e.g., identifying textual information associated with predefined entity frames: Soderland et al. (2004), or Wikipedia infoboxes: Wu and Weld (2010)), and to identify unstructured textual strings where relations are described (e.g., Banko and Etzioni (2008)).

OpenIE takes the approach that there exists information to be extracted from text but it is does not exist in a format that makes it easily identifiable. This necessitates that the structure of relations must be previously identified using machine learning or rule based systems and then, using these lexical patterns, further relations between other entities may be found in unseen text. This differs from the identification of relations between noun compounds in that the entities between which relations are to be extracted are unknown, and only the structure of the relations themselves are known.

I take a similar stance to the above approaches taken to OpenIE: that there exists information to be extracted from text but it is does not exist in a format that makes it easily identifiable. While the methods and systems presented in the above studies use lexical analysis and pattern based approaches to identify unstructured strings of text, the methods I present in Chapter 5 examine the taxonomic structure of Wikipedia in identifying the existence of relations in Wikipedia articles.

The work presented in Chapter 5 of this thesis focuses on the identification of semantic relations using semantic networks and document collections, and the classification of these relations is beyond the scope of this thesis. Much research on semantic relation extraction use the terms semantic relation and semantic relationship interchangeably (e.g., Storey (1993)). I make the distinction that a semantic relation is a conceptual combination for which the combination type (or classification) is known, whereas a semantic relationship is the unclassified existence of a relation between
two conceptual entities. i.e., the identification that a relation exists, rather than the classification of relation type.

2.5.1 Relation Extraction

Relation extraction is the task of identifying a relation existing between two entities. The example of the noun compound silver ship in the previous section is one such pair of entities, and the relation identified in that case was one of Content-Container. Without the context supplied in the sentence The silver ship usually carried silver bullion bars, but sometimes the cargo was gold or platinum, the relation could be potentially one of composition, i.e., that the ship is made of silver. Natural language text describing the interaction between entities is often used in identifying which specific relation is being invoked in a given instance. A standard approach to the identification of relations between entities in text is to examine the parsed lexical structure of a sentence, or piece of text, and identifying the operative verb being used to describe the interaction (e.g., Auer and Lehmann (2007)), or by using a pattern matching approach based on a predefined set of relations that are likely to appear in text (e.g., Hearst (1998)).

The use of pattern matching approaches and semantic frames to identify relations in semi-structured document sources (e.g., Wikipedia) has been approached by Auer and Lehmann (2007), utilising structured portions of Wikipedia articles (article infoboxes) to identify these relations. Multiple other approaches have used the identification of relational patterns in text to identify semantic relations (e.g., Ruiz-Casado et al. (2006), Wu and Weld (2008), Sarjant et al. (2009), Yap and Baldwin (2009)), however, the vast majority of relations present in Wikipedia articles occur in the semi-structured body text of each article (Milne and Witten 2008a) where
the natural language text and user defined article links do not subscribe to a formal structure (Wang et al. 2007).

OpenIE systems such as TextRunner (Banko et al. 2007), KnowItAll (Soderland et al. 2004) and Kylin (Wu and Weld 2007) have been constructed to identify relations and entity properties. These systems use machine learning techniques to identify patterns in unstructured text that indicate the occurrence of a factual statement about an entity or the relation between two entities. TextRunner analyses the lexical structure of pieces of text to identify Named Entities and possible relations between them, and uses these as a basis for identifying instances of these relations in other pieces of text, and for identifying synonymous entities. Wu and Weld (2008) use Wikipedia as a basis to learn the patterns that identify factual statements about an entity, specifically, how unstructured text statements in an article match with the attribute-value pairs present in the same article’s infobox.

These approaches either identify relation frames (i.e., the occurrence of verbs and relevant arguments) appearing in natural language text (e.g., Banko and Etzioni (2008), Wu and Weld (2010)) or use predefined patterns (e.g., Soderland et al. (2004)) to identify the structure of relations as they appear in text and extrapolate these patterns to identify similar relations between unseen entities.

The networked structure of Wikipedia has also been used as a source for defining semantic relations between articles. Auer and Lehmann (2007) used the attributive properties of article infoboxes to construct RDF triples describing entities and relations between them. This was done using basic analysis of the text surrounding mentions of linked-to articles to identify the specific nature of the relationship. Aside from the use of template based infoboxes, Wikipedia features such as the category network have been used as a method of identifying classified relation types. Nastase
and Strübe (2008) parsed Wikipedia category titles to determine the nature of the relation that existed between articles. However, this approach still fails to access the relationships present in the body text. Sarjant et al. (2009) also used the category hierarchy and links in the article preamble to identify child-parent relations between articles. This method uses inter-article links at a very basic level, but again as a method for identifying an already classified relation type.

Extracting relationships from existing semantic networks and document collections necessitates that the structure of a relationship be known in advance (e.g., relationships of the form \( X \) is a sub-field of \( Y \)). This supervised approach to relationship extraction has the potential to miss many less frequently used relationship types (e.g., \( X \) once owned a dog named \( Y \), or \( X \) can be used as a substitute for \( Y \) in automotive repair). Enumeration of every potential relationship type in semantic networks or document collections that have no consistent frame for a relationship is infeasible. The research presented in this thesis demonstrates a solution to identifying valid relationships between entities in a broad coverage, semi-structured document collection: Wikipedia.

More recently, there has been a drive towards the unsupervised extraction of semantic relations over Wikipedia. Authors such as Yan et al. (2009) advocate a clustering approach based on the dependencies between concepts in Wikipedia in order to identify collections of similar semantic relations. Wang et al. (2007) also advocate the use of ontology-based identification of semantic relations rather than the use of pattern and classification-based approaches to relation identification. The simple identification of the occurrence of a reference to an article can provide a basis for identifying a semantic relationship. For example, if an article \( p \) links to a related article \( p' \), it is probable that a semantic relation exists between these two articles.
The linking guidelines for Wikipedia editors stipulate that an article should only be linked to if it is relevant to the central topic of the article it is being linked from (Wikipedia 2011d), however there will invariably exist multiple levels of relatedness, and dimensions of relatedness, between relevant linked articles. For example, the Wikipedia article Art links to the articles Jackson Pollock and The Ming Dynasty as they are both relevant to the subject of art, but a clear statement about which of these two topics is more related to the topic of Art is difficult. Indeed, if another article was chosen at random from Wikipedia (e.g., Electrostatic nuclear accelerator), how does one determine if there exists a relationship to Art? The diversity of Wikipedia is also its Achilles’ heel in this case: large amounts of descriptive content can describe relations to many topics, but these relations may be convoluted or of spurious quality and undesirable. Separating relations of high quality from undesirable relations is further examined in Chapter 5.

The research presented in this thesis (particularly Chapter 5) presents a solution to the identification of relationships between pairs of articles using network based relations, as well as links between articles and sentence level textual analysis.

2.6 User Modelling and Prediction

Semantic relations and semantic relatedness are particularly valuable in the field of user modelling and recommendation. User modelling is the process of anticipating user behaviour patterns or future behaviour based on previous behaviour or content based similarities in existing data. These predictions can be made based on the past actions of other users (collaborative filtering (Zukerman and Albrecht 2001)), or by identifying similarities or patterns in content encountered previously by the
user (content-based models) (Schafer et al. 2007). These user models can be used to produce recommender systems (Resnick and Varian 1997) or for personalising content to a user (Fink and Kobsa 2002).

Often the content associated with an item can be as simple as a portion of text (e.g., Grieser et al. (2006)), however, multiple content-based recommender systems have used semantic networks to identify interactions between items represented by an entity within the network (e.g., The Netflix Prize (Bennett and Lanning 2007), Digital Libraries (Ferran et al. 2005), and Personalised TV Guides (Smyth and Cotter 2000)). In addition to utilising ontologies and other semantic networks to determine the relatedness of items in recommender systems, specific relationships between items can be used to enhance the recommendations presented to users.

Direct knowledge of how two objects relate to each other, rather than just knowledge of how closely related they are, allows recommender systems to describe the conclusion it reached to the system’s user, making the process more transparent. In this, explicit knowledge of the relations that exist between two entities in a recommender system can be used to demonstrate to an end user the specific quality that relates two items, and can remove an element of scepticism that a user may have in the system (Massa and Bhattacharjee 2004).

Content based or collaborative filtering recommender systems such as those on Amazon or YouTube are able to posit that an item is related to other items that the user has previously examined, but unable to explicitly identify why they are related. In the case of collaborative filtering systems, this quality is not as identifiable as in content based systems. This is largely due to the decisions that users make being subject to their own interests and influencing factors such as advertising or what was on TV last night. Other recommender systems that heavily utilise content in their
recommendation process may make use of a user model to identify the particular qualities of a recommendation that interest the user. For example, the music recommendation website Pandora.com\textsuperscript{4} allows users to create a simulated radio station that plays music based on songs that they enjoy. A user selects songs to use as radio station seeds, and Pandora then identifies the qualities of these songs as entered in the Music Genome Project (Westgren 2011), and selects songs that possess similar qualities to these seed songs to play on the user’s new radio station. Pandora exhibits transparency of recommendation as it allows a user to determine why a song has been selected to play on their radio station. The user will then be told what elements of the song overlap with the desired qualities of the radio station. The identification of the qualities that each song possesses is time consuming as it requires expert annotators to listen to each song and describe elements of its composition in terms of specific phrases (e.g., \textit{big bass sound} or \textit{strong female vocals}).

Last.fm is another music recommender system that allows users to cultivate a collection of music based around their interests, but in contrast to Pandora allows greater user input into the process. Last.fm removes the time consuming and subjective task of expert manual annotation by \textit{crowdsourcing} the annotation process, allowing users to tag songs with words they believe describe the songs. This allows users to reinforce existing tag lists and examine the qualities of each song, adding to the understanding of the user as to why songs may be related to one another. The transparency of recommendation present in both of these systems allows users to query the recommendations provided by the system. This increases user confidence in the system as they are able to clarify directly identify why an item has been recommended to them (Herlocker \textit{et al.} 2000). Conversely, it also allows a user to

\textsuperscript{4}Pandora: http://www.pandora.com
examine the reasoning behind recommendations it doesn’t agree with, resulting in the user placing more trust in a system in future (O’Donovan and Smyth 2005). The explanation of a recommendation also aids the end user in identifying why items are connected to one another, thus enabling them to gain a greater understanding of the interactions of collections of items (Massa and Bhattacharjee 2004).

In these cases of content-based recommender systems, whether they are constructed by experts in the domain (e.g., Pandora) or by a large non-expert community (e.g., Last.fm), a large user knowledge base is needed from which to extract relationships. In both of these cases all information on each song or entity is constrained in terms of attributes that the entity has been labelled with, and relationships are identified based solely on common attributes between entities.

Recommender systems can be greatly enhanced through the use of measures of semantic relatedness for calculating the relatedness of items in a content-based recommender system: using a collection of documents describing items in a recommender system, the relatedness scores of item pairs can be used as the basis of the recommender system. This is particularly attractive given a large collection of items when manual annotation of the relatedness of all item pairs is a time-expensive task. Recommender systems also benefit from the application of semantic relations: the description of the specific reasons for items being related to one another allows transparency and increases user trust in the recommender system (Herlocker et al. 2000; O’Donovan and Smyth 2005).
2.7 Evaluation in Language Technology

Due to the varying nature of language technology tasks, many different approaches are taken to the evaluation of methodological performance of a system or technique. Tasks such as machine translation, event detection and word sense disambiguation all require very different approaches to evaluation. Two approaches to evaluating systems are to compare a system to an established gold standard set of results, and to manually annotate the results produced by a system. These two approaches to evaluation are pertinent to the experiments conducted in Chapter 4 and Chapter 5. An examination of their application to the evaluation of the methods presented in this thesis is done in Section 3.3, but an overview of each of these evaluation methodologies and previous use is presented in this section.

2.7.1 Comparison to Gold Standard Data Sets

Calculations of semantic relatedness as performed by the metrics in Section 2.4 produce a numeric score of relatedness between word pairs. To examine the effectiveness of these measures, a gold standard set of relatedness scores is needed for comparison. However, the measurement of human judgements of semantic relatedness is terms of a nominal scale is a difficult process. Conceptual combination is an amorphous task, and the notion of representing a human judgement of a level of relatedness with and one person’s judgement of how related two concepts are may differ markedly to that of another person. The construction of a set of numeric gold standard of human judgements of semantic relatedness has been approached by numerous authors.
Rubenstein and Goodenough (1965)’s approach to the creation of a semantic relatedness gold standard has become a template for the establishment of semantic relatedness datasets. Rubenstein and Goodenough identified a set of 65 word pairs that were designed to range from highly synonymous to completely unrelated. These pairs were given to two separate groups of 15 and 38 undergraduate students who were asked to rank the word pairs in terms of synonymy, and then asked to give each of these word pairs a score between 0 and 4 based on their “similarity of meaning”.

This set of relatedness scores was further refined to a smaller subset of 30 word pairs by Miller and Charles (1991): 10 highly synonymous pairs (rated between 3 and 4), 10 mid-range pairs (rated between 1 and 3) and 10 pairs of low synonymy (rated between 0 and 1). The 30 pairs were re-evaluated using the same experimental design as Rubenstein and Goodenough (1965).

These two datasets have become the de facto standard for the evaluation of computational measures of semantic similarity (Budanitsky and Hirst 2005).

Recently, a larger set of 353 word pairs was developed by Finkelstein et al. (2002). This set was designed to test word pairs which had a high likelihood of relatedness (e.g., hotel & reservation) as well as pairs with a high likelihood of synonymy (e.g., furnace & stove). In this collection of 353 word pairs, Miller and Charles’s 30 word pairs were included. The word pairs were this time evaluated in terms of their relatedness, rather than synonymy and similarity as done by Miller and Charles (1991) and Rubenstein and Goodenough (1965). The 353 word sets were partitioned into two separate sets of size 153 word pairs and 200 word pairs; these partitions were then evaluated by 13 and 16 annotators respectively. The evaluation scale for this experiment was to provide a numerical score between 0 (“totally unrelated words”) and 10 (“very much related or identical words”) based on the relatedness of the words.
For this evaluation, the annotators were not required to place the word pairs in order of perceived relatedness before being given a score of relatedness. Agirre et al. (2009) split this 353 word pair data set into two separate partitions, one containing pairs judged to test relatedness, and the other to test similarity.

The performance of a computational measure of semantic relatedness is to produce a score for each of the pairs in one of the above lists, and then test if the scores produced by the measure correlate with the relevant gold standard data set. The correlation between two lists of scores (one generated by a measure, the other a gold standard set) is measured using one of two correlation coefficients: Pearson correlation coefficient, and the Spearman rank-order correlation.

Simply stated, the Pearson correlation coefficient ($\rho$) is calculated by dividing the covariance of two sets of values ($\text{cov}(X,Y)$) by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$  \hspace{1cm} (2.8)

Covariance is a measure of how often a change in one set of data is reflected in the other (i.e., an indication of how often the two sets of data show the same variations). This formula produces a $\rho$ value of between $-1$ and $+1$, with a correlation of $\rho = -1$ indicating a perfect negative correlation (a change in one set of data is reflected oppositely in the other set of data) and a correlation of $\rho = +1$ indicating a perfect positive correlation (a change in one set of data is reflected identically in the other set).

The calculation used to calculate the Spearman correlation coefficient is a simpler one: both sets of data ($X$ and $Y$) are first sorted in ascending order, after which the
difference in these ranks is taken:

\[ d_i = \text{rank}(X_i) - \text{rank}(Y_i) \]  

(2.9)

The sum of the squares of differences in rank is then weighted with regard to the length of the data sets \((n)\) and subtracted from 1:

\[ \rho = 1 - \frac{6 \times \sum d_i^2}{n(n^2 - 1)} \]  

(2.10)

This calculation of correlation places a greater importance on the correct ranking of data points rather than solely identical changes in magnitude. For a complete discussion of the formulation of the Pearson and Spearman correlation coefficients, see Kendall et al. (1994).

### 2.7.2 Manual Annotation of Results

A second approach used in this thesis is the manual annotation of results produced by a system. This is common for the evaluation of non-numeric data, as statistical analysis of results cannot be performed without a gold standard. In many cases, a lack of a gold standard set of results is due to the prohibitive cost of constructing a comprehensive standard. The lack of a gold standard set of data may also be due to the uniqueness of a task, and the understanding of a human annotator is required to identify nuances in the results.

Even in established tasks of natural language technology, the creation of large-scale corpora for evaluation is still an issue, particularly for languages other than English: Fürstenau and Lapata (2009) used a small set of sentences manually annotated with semantic role labels (see Section 2.5) as a seed corpus to identify similar labels oc-
curring in a larger corpus. Indeed, even tasks which use an established gold-standard and a statistical analysis of comparison require the manual annotation of a domain gold-standard (e.g., the creation of semantic relatedness word pair sets as described in Section 2.7.1). The explicit use of manual annotation of system results has been used for tasks such as sentiment analysis (Wilson et al. 2005) and the classification of epistemically modalised statements (determining the perceived level of truth in a statement) (Rubin 2007) due to the lack of an existing gold standard against which to compare.

If there exists little agreement between annotators, then the resulting analysis of an annotated set of results has little meaning: if annotators rarely agree on the classification of a data point, it indicates the classification is a soft classification at best. This can be due to many factors: the line between two classes may be too loosely defined for an annotator to confidently classify a data point; the evaluation task itself may be poorly defined (e.g., the correct questions are not being asked of the annotators, or ill-defined instructions may have been used); or the results produced by the system may be of too poor quality to clearly classify according to evaluation guidelines. To evaluate the effectiveness of the evaluation, the overall agreement between annotators (inter-annotator agreement) must be calculated.

A measure of the frequency of annotators’ agreement with one another is described by Fleiss (1971). Fleiss’ Kappa statistic is a measure of inter-annotator agreement, it applies in cases where there are multiple annotators placing data points into a fixed set of classes. It is an extension of Cohen’s Kappa statistic (Cohen 1960) which only measures the agreement between two annotators. These two statistical measures of agreement differ from measures such as the Pearson or Spearman correlation statistics.
in that they also take into account the possibility that annotators fell into agreement by chance.

Cohen’s Kappa statistic is calculated by:

\[
\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
\]  

(2.11)

where \(Pr(a)\) is the observed agreement between annotators, and \(Pr(e)\) is the probability that the annotators agreed by chance. The probability of chance agreement is taken from the sum of the annotators’ individual observed response to a class. For example, if two annotators (A and B) graded 50 relationships and annotator A graded 40 of those as valid, while annotator B graded 30 as valid, the probability of chance agreement is given by the sum of the agreement over both classes (valid and invalid). \(Pr(e) = 0.8 \times 0.6 + 0.2 \times 0.4 = 0.56\)

Fleiss’ Kappa statistic is given by the sum of all annotators’ agreement with all other annotators across all classes:

\[
\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}}
\]  

(2.12)

where \(\bar{P}\) is the sum of annotator agreement for all classes, and \(\bar{P}_e\) is the agreement by chance of all annotators across all classes.

In the evaluation processes described by Wilson et al. (2005) and Rubin (2007), the level of inter-annotator agreement differed greatly. Despite the variability of success depending on the definition of the evaluation task itself (Sim and Wright 2005), some authors have tried to provide an absolute scale upon which to base the success of evaluation. Landis and Koch (1977) created an interval scale labelled with
levels of agreement, ranging from poor agreement ($\kappa < 0$) to almost perfect agreement ($\kappa = 0.81 - 1.00$).

The evaluation task described by Wilson et al. achieved a large amount of agreement between annotators ($\kappa = 0.87$). Rubin’s evaluation task identified a much lower inter-annotator agreement ($\kappa = 0.41$), but declared this to be “fair agreement” based on the classification described by Landis and Koch. The level of agreement is highly dependent on factors such as the number of annotators, and the number of classes into which a data point can be placed (Sim and Wright 2005). Classes are treated as independent classifications in the calculation of Fleiss’ Kappa score. This creates instances where annotators disagreed only slightly (e.g., the difference between a 4 and a 5 on a nominal scale), but is treated as total disagreement in the calculation of Fleiss’ Kappa. Additionally, with a large number of annotators, the likelihood that they will all agree on the classification of a data point among a number of closely related classes is minimal, resulting in a low kappa score despite potentially only being minimally separated (Sim and Wright 2005).

2.7.3 Mechanical Turk

The selection of annotators is also a consideration when evaluating a natural language system. For example, multilingual annotators are required for accurate evaluation of machine translation systems, and an understanding of the English language and grammar is needed to evaluate extracted semantic roles. However, as the annotators are human beings, issues such as fatigue and disinterest in the task must be addressed. Many studies use members of the research group performing the study (e.g., Finkelstein et al. (2002), Agirre et al. (2009)), or undergraduates at the same educational institution (e.g., Rubenstein and Goodenough (1965)). Repeated use of
this sample population can result in over-familiarity with the task or a vested interest in the performance of the system, causing bias in the results. To remove the chance of bias and annotator fatigue it can be advantageous to draw from a more diverse population, or to partition the annotation task among multiple annotators.

The Amazon Mechanical Turk\(^5\) is a service that allows Human Intelligence Tasks (HITs) to be distributed among large groups of human annotators (colloquially known as *Turkers*). A HIT Requester is able to specify the number of times they want each HIT completed by (different) Turkers, the level of acceptable work a Turker has done before they are able to participate in a HIT, and the length of time they allow a Turker to take to complete a HIT. These options allow a Requester to have a high level of confidence about the quality of annotation. Snow *et al.* (2008) replicated many linguistic annotation tasks, including judgements of word pair relatedness similar to the experiments performed by Rubenstein and Goodenough (1965) and Finkelstein *et al.* (2002), and found that the annotations produced by Mechanical Turkers were comparable to the gold standard labels produced by domain experts. Mechanical Turk has also been successfully used for the evaluation of more complex linguistic tasks such as evaluation of translation quality (Callison-Burch 2009) and transcription of spoken language (Marge *et al.* 2010).

### 2.8 Summary

Conceptual representation and conceptual combination has been examined extensively in relation to cognitive science, and the diversity in approaches to conceptual representation has been identified as a hurdle in computational approaches to conceptual combination. Ontologies and semantic networks covering a single domain (e.g.,

\(^5\)The Amazon Mechanical Turk: [www.mturk.com](http://www.mturk.com)
Chapter 2: Background

WordNet, UMLS, Iconclass) utilise a single structural representation for all nodes in the network, with a finite set of relations sufficient for representing all interactions between entities. Semantic networks encompassing a broader range of concepts, such as digital encyclopedias, require a different approach to the representation of these concepts and the relations that exist between them.

Computational analysis of conceptual combination has been explored extensively, both in relation to the quantification of levels of relatedness between concepts, and the identification of specific relations between concepts. A large proportion of this work has been done by using WordNet synsets to represent associated conceptual entities. Recently, Wikipedia has been used as an alternate broad-coverage semantic network for calculating the semantic relatedness of concepts, and for identifying semantic relations between concepts.

The use of a semi-structured conceptual representation and the ability to represent any relation that may exist between these concepts makes Wikipedia a powerful resource for relationship discovery. Pattern based approaches to relation extraction from Wikipedia have proven successful, but miss the vast majority of relations as they do not fit any standard template or pattern. This can be due to them being described using unforeseen language (e.g., the relatively recent use of google as a verb), or being completely unique relations (e.g., Neil Armstrong being the first person to walk on the moon).

In this thesis I demonstrate that representing a diverse collection of objects with differing conceptual structures using a semi-structured document format (by using Wikipedia) allows for the discovery of relations without using a pattern based approach that would potentially miss unforeseen relations. The approach to this task is two-fold: the identification of the levels of relatedness between conceptual entities in
Wikipedia, and the specific identification of the relationships present between articles in Wikipedia. The detailed use of Wikipedia as a computational semantic network requires adaptation of existing computational resources. In addition to resources describing Wikipedia, the novel nature of the domain being used for experimentation requires evaluation resources to be created from scratch. The theoretical basis for the construction of these resources and the details of their construction is presented in the following chapter.
Chapter 3

Computational Resources and Domain Requirements

Existing approaches to the task of identifying semantic similarity and relatedness of word pairs make use of resources such as large textual corpora for statistical analysis (e.g., statistical co-occurrence of terms: Church and Hanks (1990)), or the use of rigidly defined and highly structured ontologies and semantic networks (e.g., WordNet and Iconclass) to determine the level of similarity or relatedness present between words. These resources provide a computational representation of a knowledge domain (either a cross-domain collection of knowledge, or a single domain), providing information on the organisation of knowledge (e.g., semantic networks) and how language is used (e.g., corpora). However, less constrained entities that possess diverse content and sets of relations (e.g., a collection of museum exhibits) require a representation that allows for diversity in entity description and relations.

In this research, the information rich domain of Cultural Heritage is utilised as a basis for experimentation. From a computational standpoint, the resources possessed
by this domain are desirable as it is: (1) comprised of a collection of contextually
rich entities, and (2) those entities are able to be represented computationally (see
Section 2.3.4). A problem inherent in using Wikipedia as the basis of representation
for contextually diverse entities is that it is constantly updated, potentially changing
experimental results over time. To ensure that the experiments performed in this
thesis are repeatable, a static version of Wikipedia must be used.

In addition to an established framework with which to represent these entities,
the evaluation of methods presented in this thesis require: a gold standard of relat-
edness judgements of pairs of conceptually diverse entities as determined by a group
of annotators familiar with the set of entities, and a system of manual evaluation
of the validity of relationships between these diverse entities. While the use of gold
standard collections of human judgements is an established practice in semantic re-
latedness literature (e.g., Rubenstein and Goodenough (1965), Miller and Charles
(1991), Finkelstein et al. (2002)), manual evaluation of semantic relationships has
not been previously approached. In an effort to demonstrate a rigorous approach to
this evaluation, I utilise multiple collections of human annotators to determine the
validity of extracted relationships. Both the groups of annotators used to evaluate
this data, and the technologies used to gather these annotations are described in this
chapter.

In this chapter I present the resources required for experiments of semantic re-
latedness and relationship extraction over a domain containing entities with diverse
conceptual structure. Much of the work in Chapter 4 was performed using a data
set comprised of exhibits from Melbourne Museum. Section 3.1 describes why Mel-
bourne Museum was chosen, and what specific resources were utilised. The collabo-
rative encyclopedia, Wikipedia, was used extensively in this thesis for computation.
Wikipedia’s use as a computational resource is described in Section 3.2. The resources required for the evaluation of the tasks presented in Chapter 4 and Chapter 5, are described in Section 3.3. A final summary of the resources described in this chapter is presented in Section 3.4.

### 3.1 Melbourne Museum

Numerous domains have the potential to benefit from the automatic identification of relationships between conceptual entities: recommender systems such as the Netflix film recommender system can utilise relations between films to identify why a user selects certain types of movies; the analysis of the relatedness of tourist destinations and activities can be used for the automatic creation of a holiday designed around a person’s interests. The applicability of these domains is explored thoroughly in Chapter 6. The domain used thesis is that of Cultural Heritage. This is for a number of reasons including the diversity of content in individual domain entities (museum exhibits), a pool of annotators and experts familiar with the content of the domain, and application for the use of the methods of relatedness calculation and relation extraction presented in this thesis.

The primary factor is that museum exhibits are information rich entities with a large amount of contextual diversity: the differing aspects of an entity that are relevant in different contexts (recall the bark canoe example given in Section 1.1). Melbourne Museum is a large museum with a diverse collection of exhibits, covering many genres (a description of the collection at Melbourne Museum is given in Section 3.1.1). This diversity of conceptual structure and the potential for multiple relations to exist between exhibits makes it a desirable domain for the development
of measures of semantic relatedness and relation extraction between contextually diverse entities. During the course of this thesis, I was generously granted access to the exhibit catalogue of Melbourne Museum for use in my experiments. Melbourne Museum also has a large visitor population that I was able to use for the data collection exercise described in Section 4.2, and the manual evaluation of relationship validity presented in Section 5.3.

3.1.1 Museum Collection

Melbourne Museum is Melbourne’s largest natural history museum and is located in a purpose built facility in Melbourne’s Carlton Gardens. In addition to containing a large natural history collection, it also contains exhibits and galleries relating to the environment, science and technology, Victorian history and indigenous culture. It has a diverse collection of exhibits including the taxidermised hide of Phar Lap, replicas of the largest gold nuggets found in the Ballarat Goldfields, collections of oil paintings by numerous local artists, and multiple assembled skeletons of prehistoric animals.

The museum is divided into six primary galleries, with a number of smaller collections of exhibits in the hallways and lobby area. The primary galleries are the Science and Life Gallery, Forest Gallery, Mind and Body Gallery, Te Pasifika Gallery, Bunjilaka Aboriginal Cultural Centre and the Melbourne Gallery. This was the state of the museum layout as of mid 2009. Melbourne Museum has since substantially altered the layout of exhibits within these galleries as well as relocated entire galleries. Museums maintain the need to restructure as new exhibits are added to collection and new exhibitions are created, thus the content and layout of a museum (in this case, Melbourne Museum) is constantly in
flux. Indeed, this is one of the issues that the work in this thesis addresses: pre-made paths such the Louvre’s Thematic Trails must be reconstructed when new content is added to a gallery, but the automatic identification of relationships between exhibits allows these paths to be recalculated once the new exhibit content is incorporated into the exhibit database.

These galleries are further divided into individual exhibitions, e.g., the Phar Lap exhibition in the Melbourne Gallery or the CSIRAC exhibition in the Science and Life Gallery. In addition to the permanent exhibitions, Melbourne Museum also hosts temporary collections such as exhibitions on Pompeii, The Titanic and
Tutankhamun. The exhibits used for data collection and experimentation in this thesis are only sampled from the permanent exhibitions at Melbourne Museum.

Forty exhibits were selected from across the entire museum collection for the experiments presented in this thesis. Exhibits from all galleries were included in this set, excepting the Bunjilaka Gallery due to cultural considerations in photographing the exhibits. These exhibits were chosen at random from the museum collection, with the requirement that they have an equivalent article that describes their content on Wikipedia. The full list of the exhibits used in the experiments conducted in this thesis, the Wikipedia articles which were used to describe their content, and the respective galleries of which they are members, are included in Appendix A.

3.1.2 Museum Catalogue

Melbourne Museum uses a specially constructed exhibit catalogue based on the KE-Emu cataloguing system (see Section 2.3.4). Exhibits are described in terms of a set of attribute values, and linked to associated multimedia resources (e.g., images, videos, etc.). As the collection at Melbourne Museum is diverse, it is not required that all fields be filled in. In some cases, only a dozen fields are entered out of several hundred. More recently, access to this catalogue has been placed on the Melbourne Museum website, and an API that queries this catalogue made freely available.\(^1\) This API has only been in place since late 2010, and I was unable to use it in the experiments in this thesis due to its initial non-existence.

The previous Melbourne Museum website utilised a simpler representation of exhibits: a single page would describe an entire exhibition area (e.g., The Marine Life or Phar Lap exhibitions), and a varying level of detail would be used to describe

the exhibits in this area. In some cases, each exhibit in the exhibition would have a
web page of its own, in other cases, all exhibits in the exhibition were described by a
single collective web page.

![CSIRAC exhibit web page from the Melbourne Museum website.](image)

Figure 3.2: The CSIRAC exhibit web page from the Melbourne Museum website.

The web pages from the original museum website that corresponded to the ex-
hibits selected for experimentation were pared down into plain text format, and used
as documents to represent their respective exhibits for some experiments described
in Section 4. A limitation of the museum website is that not all of the 40 selected
exhibits used in experimentation had a corresponding web page on the museum web-
site. This lack of coverage and its effects on the ability to compute exhibit relatedness
is addressed in Section 4.1.2.
3.1.3 Museum Visitors

A further benefit of the collaboration with Melbourne Museum is the access to large groups of people familiar with the collection at Melbourne Museum: museum staff and museum visitors. These two groups of people form a collection of domain experts that are able to evaluate results produced by constructed systems using a thorough knowledge of the museum and its layout. While other studies performed in conjunction with Melbourne Museum utilise first time visitors to the museum (namely Bohnert et al. (2008)), I focus on repeat visitors that are highly familiar with the collection at Melbourne Museum. This is done by only asking Melbourne Museum visitors who have signed up to the Museum’s membership program to participate in evaluation tasks. These repeat museum visitors were especially valuable in constructing a gold-standard data set for evaluating newly constructed measures of semantic relatedness (see Section 4.2). Melbourne Museum staff are used in the evaluation of extracted relationships presented in Section 5.3.

3.2 Wikipedia

Wikipedia is increasingly used as a computational resource for tasks of relation extraction (Wang et al. 2007), text categorisation (Gabrilovich and Markovich 2006), and named entity disambiguation (Bunescu and Pașca 2006). The Wikipedia corpus has also been used to construct a hierarchical semantic network in multiple studies, e.g., Ponzetto and Strübe (2007b), Zesch et al. (2007a). The use of Wikipedia in computational tasks has been facilitated through the use of resources such as the web-based Wikipedia API\(^2\) for accessing the live version of Wikipedia, and regular XML

dumps of the entire content of Wikipedia.\footnote{Wikipedia dumps: \url{http://dumps.wikimedia.org}} This section describes how Wikipedia was used as a static resource, the specific version of Wikipedia used throughout this thesis, and how it was adapted to hierarchical semantic network.

### 3.2.1 Wikipedia XML Dumps

Wikipedia provides a PHP based API used for returning live queries to Wikipedia. Using this API, information regarding Wikipedia User pages, Talk pages and Articles can be retrieved from the current online version of Wikipedia. This API is a powerful tool, but is unsuitable for experimentation for two reasons: (1) results based on these queries will not be repeatable as they will be based on the structure of Wikipedia at a specific (unrecorded) moment in time, and (2) continual querying of the PHP API will result in excessive bandwidth usage, particularly during the construction of a hierarchical semantic network comprising all categories in Wikipedia. For these two reasons, the preferred alternative is to use a local, static version of Wikipedia.

Wikimedia (the organisation that hosts Wikipedia, Wiktionary, Wikiquotes, etc.) routinely backs up all content for all language versions of Wikipedia and provides it for download at \url{http://dumps.wikimedia.org}. A number of files are generated for each dump, and these files can range in size from a few kilobytes (the list of Wikipedia articles which did not dump correctly), to over a terabyte (the complete dump of the entire history of every English Wikipedia article).

For the experiments conducted in this thesis, only the \texttt{pages-articles} dump file is used. This file contains an xml representation of the full text and meta-data of all current version articles (containing no revision history, only the Mediawiki marked up raw text of the current version of each article). The full article text includes elements
Porcupines are [[rodents]] with a coat of sharp [[Spine (zoology)|spines]], or quills, that defend or camouflages them from predators. They are indigenous to the Americas, southern Asia, and Africa. Porcupines are the third largest of the rodents, behind the [[capybara]] and the [[beaver]]. Most porcupines are about {{convert|25|-|36|in|cm|abbr=on}} long, with an {{convert|8|-|10|in|cm|abbr=on}} long tail. Weighing between {{convert|12|-|35|lb|abbr=on}}, they are rounded, large and slow. Porcupines come in various shades of brown, grey, and the unusual white. Porcupines' spiny protection resembles that of the unrelated [[Erinaceomorpha|erinaceomorph]] [[hedgehog]]s and [[monotreme]] [[echidna]]s.

Figure 3.3: A piece of raw text with markup from the Wikipedia Porcupine article. Note the markup indicating templates, unit conversion and article links.

such as inter-article links, templates, category membership, and links to equivalent versions of the article in other languages (language links). An example of a portion of this markup can be seen in Figure 3.3.

The version used at all stages of this thesis (unless otherwise noted) is the September 9th, 2009 dump of Wikipedia.4

Using the category membership information of these articles, a hierarchical seman-

---

4The name of the specific dump file used is enwiki-pages-articles-20090909.xml. This file is available at the Wikimedia dump page (http://dumps.wikimedia.org).
tic network was built. The construction of this network is described in the following section.

3.2.2 Wikipedia Category Hierarchy

Wikipedia classifies articles using categories. Every article in Wikipedia must be the member of at least one category (Wikipedia 2011b). Categories are organised in a hierarchy with broader categories subsuming more specific categories. At each level, categories can contain both child articles and child categories. For example, Jupiter and Saturn are both classified as members of the Gas giant planets category, which also contains subcategories of Hot Jupiters and Hot Neptunes. This membership also demonstrates another aspect of Wikipedia categories: that articles are classified by categories with as great a specificity as possible. Jupiter and Saturn are not explicitly classified as members of the Planets, Bodies of the Solar System or even Planets of the Solar System. All planets of the Solar System are members of their own individual categories (e.g., the Earth will be a member of the Earth category and Mars will be a member of the Mars category), which are in turn members of the Planets of the Solar System category.

Using the article dump described in the previous section, a computational representation of the category hierarchy was constructed. I used a bottom up approach to construct this hierarchy: using the category membership of a set of seed articles as starting points and tracing a path to the root via category subsumption. In the Mediawiki markup, category links are a sub-type of link. Links between articles in the same wiki are denoted in the raw text of an article by double square brackets. For example, when “[[Fish]]” is written when editing an article, it will appear as a hyperlink to the article on Fish. The category membership of an article is included at the bottom
of articles as a list of links of the form \[[Category:Category name]\]; each of these links is on a separate line. These links were identified as the parent categories of a given article, and were expanded until a unique root node was reached. Due to the multiple inheritance property of category membership (some articles were members of over 10 categories), this hierarchy became very large. All categories are subsumed by two root nodes: FUNDAMENTAL and MAIN TOPIC CLASSIFICATION. These two categories are further subsumed by an administrative category: ARTICLES, which, in addition to containing the categories FUNDAMENTAL and MAIN TOPIC CLASSIFICATION, contains a number of other administrative categories. I utilise the ARTICLES category as the root node of a constructed category hierarchy, with the restriction
that only subcategories including encyclopedic content are included as children (i.e., the removal of administrative categories such as CATEGORIES FOR DELETION).

Many studies utilise this hierarchy for measuring the semantic relatedness of article pairs (e.g., Ponzetto and Strübe (2007b), Zesch et al. (2007a)), however none describe the explicit process used to construct this hierarchy, nor which version of Wikipedia is used in the construction. Importantly, the discussions of the resolution of cycles in graph structure and how a unique root node is adapted from the category hierarchy are omitted. While Wikipedia editing guidelines discourage the inclusion of cycles in category inheritance, they also acknowledge that in some cases cycles are acceptable (Wikipedia 2011b). Logical cycles, such as EDUCATION $\rightarrow$ ACADEMIA $\rightarrow$ ACADEMIC DISCIPLINES $\rightarrow$ EDUCATION ($\rightarrow$ denotes a subcategory-of relationship) have been the subject of much discussion and have been consciously removed from Wikipedia (Wikipedia 2011a). However, a casual exploration of the category network presents additional, naturally occurring, cycles. For example, a cycle involving the EDUCATION is present in: EDUCATION $\rightarrow$ KNOWLEDGE $\rightarrow$ LEARNING $\rightarrow$ EDUCATION.\(^5\)

I resolve cycles in Wikipedia by putting in place the restriction that a parent category must be closer to the root node (using the shortest path to root) than its child. For example, EDUCATION is a valid child of KNOWLEDGE as KNOWLEDGE is grandchild of the root node, ARTICLES, (via the path KNOWLEDGE $\rightarrow$ FUNDAMENTAL $\rightarrow$ ARTICLES), and EDUCATION a valid child of KNOWLEDGE. However, LEARNING is not a valid parent of KNOWLEDGE as its shortest path to the root node is via EDUCATION (i.e., the shortest path from LEARNING to the root is longer than EDUCATION to the root).

\(^5\)As at 1.41pm 9/2/2011 GMT
In addition to the restriction that a child be further from the root than its parent, I remove any administrative category memberships (e.g., CATEGORIES FOR DELETION ARTICLES NEEDING CITATIONS, LISTS, etc.) as these are not part of the conceptual inheritance structure, and create an artificial shortening of the distance between articles and the root node. This is due to their being direct children of the root ARTICLES category.

3.3 Evaluation

This section describes the computational resources required for evaluation of the experiments performed in Chapter 4 and Chapter 5. The evaluation methodologies themselves and their respective application are described in Section 2.7. This section describes the considerations taken in annotator selection, methods of data collection, and the frameworks used for the collection of evaluation data.

3.3.1 Annotator Selection

For the purposes of the construction of the gold standard data set and the manual evaluation of extracted relationships, the selection of the annotators must be taken into consideration. These considerations include the familiarity of the annotators with the data being evaluated, the annotators’ understanding the evaluation procedure itself and the reliability of the annotators. The datasets used in this thesis are taken from the collection of exhibits at Melbourne Museum, and familiarity with this collection is a desirable property of annotators, particularly in constructing a collection of judgements of exhibit relatedness.

In Chapter 4, the construction of the gold-standard data set is based on the pro-
cess described by Rubenstein and Goodenough (1965) in which the annotators were native English speakers, and thus highly familiar with the domain they were annotating (English words). The construction of a set of exhibit relatedness judgements necessitates that the annotators have a high familiarity with the Melbourne Museum collection. For this purpose, the use of Melbourne Museum members is ideal: they are highly familiar with the collection.

The evaluation performed in Chapter 5 is not as strongly tied to Melbourne Museum itself (but still uses the Wikipedia articles associated with the museum exhibits identified in Chapter 4), and as such does not require as great a familiarity with Melbourne museum. For this reason, alternate groups of annotators were used in addition to those familiar with the collection of Melbourne Museum. These alternate sources were comprised of Computer Science postgraduate students and Mechanical Turkers. Melbourne Museum staff were also used as a group of annotators due to their capacity as domain experts familiar with the content represented by the Wikipedia articles used (as their selection was based on their alignment to museum exhibits).

### 3.3.2 Data Collection and Online Surveys

The data collection performed in Section 4.2, and parts of the evaluation performed in Section 5.3, were done using an online survey. The specifics of each survey is described in their respective sections, however, they were all constructed using the same method. Each survey is a cgi-script that produces a series of web pages containing a randomised allocation of data for each participant. These web pages were hosted on a University of Melbourne server. The annotated data produced in these surveys was recorded into an SQL database, again hosted on a University of Melbourne server. In each of these surveys, annotators were identified by an md5 hash of
the combination of their IP address and the timestamp of when they commenced the survey. This allows data annotated by a single user to be identified, but also preserves the annotator’s privacy by anonymising their details. The layout and instructions for each iteration of the evaluation process is provided in Appendix B.

3.3.3 Mechanical Turk

The purpose of Amazon’s Mechanical Turk and its reliability are described in Section 2.7.3. This section describes the specifics of constructing a HIT, and how the test is disseminated to Turkers.

Mechanical Turk has multiple interfaces for creating and managing Human Intelligence Tasks (HITs). The simplest of these is a wizard-like web interface that allows a requester to construct the layout of the HIT using a html editor or a visual editor, upload data for annotation, set annotator requirements, and set the number of times each data point is to be annotated (to measure inter-annotator agreement, or obtain a majority class judgement).

This interface is much more limited in scope than the command line interface and the Mechanical Turk API, as it does not allow the automated scheduling of events (e.g., automatic approval of a Turker’s HIT) and does not allow for the creation of customised qualifications. However, for the approach taken in Section 5.3 these additional features are not needed and the web interface is sufficient for the creation of the HIT.

HITs are composed of a template that is common to all HITs in a series, and a set of data that fits this template. For example, in testing geographical knowledge, a HIT may ask *In what country is the city* [city name] *located?* and replace [city name] with a city name taken randomly from a list of city names in the associated data file.
Furthermore, each HIT in a set must contain a uniform number of data points to be annotated, this is due to each HIT needing a single frame into which data variables are placed, thus making a single template into which all data points are able to be placed. This allows the same unit of work to be performed by each Turker that performs the task.

The specific design of the Mechanical Turk HITs used in the evaluation of the relationship extraction methods is presented in detail in Section 5.3.

3.4 Summary

In this thesis, conceptual combination of diversely structured entities is approached on two fronts: those of semantic relatedness and relationship extraction. The choice of domain presented in this chapter was that of a museum collection with diverse content, specifically the collection present at Melbourne Museum. Melbourne Museum affords the benefits of a diverse collection of physical entities with which to experiment, and a collection of domain experts (repeat visitors and museum staff) for use in evaluation and construction of data sources.

The catalogue of museum exhibits at Melbourne Museum uses an attributive database (a purpose constructed catalogue using the KE-EMu catalogue architecture), describing each exhibit in terms of a collection of attributive values and a defined collection of possible relations between exhibits. This rigidity does not allow for complete representation of exhibit qualities (due to the diverse nature of the exhibits no subscribing to a common representation), nor a complete quantification of the relations that exist between them.

To represent the contextual diversity between entities used in the experiments in
this thesis, I use Wikipedia to represent the content of museum exhibits. Wikipedia provides an encyclopedic basis for the representation of complex physical entities, and the relations that exist between them. I have utilised a static XML dump of Wikipedia to provide the basis for a semantic network of Wikipedia articles, organised using Wikipedia’s category hierarchy and the links between articles. This semantic network is as the basis for experiments in Chapter 4 and Chapter 5.

To determine the effectiveness of a computational measure to calculate the relatedness of museum exhibits, I present a newly constructed dataset of museum relatedness judgements. The evaluation methodology used in studies relating to the calculation of semantic relatedness utilises gold-standard human judgements of entity relatedness (Rubenstein and Goodenough 1965; Miller and Charles 1991; Finkelstein et al. 2002). This methodology is the foundation of the creation of a gold-standard data set of museum exhibit relatedness, constructed with the co-operation of Melbourne Museum visitors. The full detail of this construction is described in Section 4.2. The evaluation of semantic relationships between Wikipedia articles requires a post evaluation framework. Furthermore, I have described multiple approaches to the evaluation of extracted semantic relationships, in particular the establishing of framework for the evaluation of extracted relationships. This is first done via a purpose constructed series of web pages, and later by using the Amazon Mechanical Turk annotation service.

The resources presented in this chapter describe the computational foundation for the representation of data, the adaptation of Wikipedia into a computational semantic network, and the construction of evaluation data sets and evaluation interfaces.
Chapter 4

Semantic Relatedness

The research presented in this thesis provides a computational methodology for identifying the conceptual connections humans make between entities. These conceptual interactions have the ability to be used in applications such as recommender systems (e.g., constructing a personalised museum tour, or recommending an appropriate film) or narrative construction (e.g., tying together a group of museum exhibits that possess a common narrative thread). In terms of the domain discussed in Chapter 3, the semantic relatedness of museum exhibits can be used to identify groups of exhibits that revolve around a common theme, both for tour construction and for curating purposes (e.g., constructing an exhibition or gallery around a specific theme). Narrative construction and recommendation explanation make use of specific relationships between entities. Relationship extraction is discussed in Chapter 5, and this chapter will focus on the measurement of semantic relatedness between contextually diverse entities.

In this chapter, multiple measures of semantic relatedness and similarity are used to identify the relatedness of museum exhibits (using the exhibits described in Sec-
tion 3.1.1), and compared to museum visitor relatedness scores collected in Section 4.2.

Established methods of document and ontological similarity (created for WordNet and Wikipedia) are compared with a method created specifically for this task. These measures are compared to two physical distance baselines that based on the physical layout of exhibits at Melbourne Museum. I also describe a purpose created measure of semantic relatedness that takes into account Wikipedia article links and category membership to calculate the semantic relatedness of two articles, and demonstrate the performance of this measure in comparison to established measures of semantic relatedness over the museum exhibit evaluation set.

The remainder of this chapter is structured as follows: the primary considerations used in calculating semantic relatedness between museum exhibits and their adaptation to a computational measure of semantic relatedness, presented as the Related Article Category Overlap (RACO), are described in Section 4.1. Section 4.1 also includes a description of how established measures of semantic relatedness described in Section 2.4.2 can be adapted to the Wikipedia category hierarchy. The construction of a gold standard set of exhibit pair judgements from annotators familiar with the domain is described in Section 4.2; this dataset is used as the evaluation dataset in later sections of this chapter. The comparative performance of the established measures of semantic relatedness and RACO, using the constructed exhibit pair dataset, is presented in Section 4.3. RACO is then compared to another measure of semantic relatedness created specifically for Wikipedia, this time using two established datasets for the task of Word Sense Disambiguation, in Section 4.4. As a final demonstration of the versatility of RACO, I present an analysis of how it may be used in the task of generating candidate labels for use with the task of topic modelling in Section 4.5. Finally, the chapter is summarised in Section 4.6.
4.1 Semantic Relatedness of Museum Exhibits

The semantic relatedness of pairs of word senses, as described in Section 2.3.1, can be used as the basis for disambiguating word senses. The semantic relatedness of entity pairs is also used as the basis of many content-based recommender systems (see Section 2.6 for a discussion). The principle of identifying the conceptual relatedness as performed in tasks of measurement of lexical relatedness and used in content-based recommender systems forms the basis for the experimentation in this chapter: the measurement of relatedness of contextually diverse entities.

The concepts in this research are based on specific instances of physical entities (specific museum exhibits) rather than abstract concepts (such as words), and as such, there are additional factors to consider in the construction of methods to calculate the semantic relatedness of museum exhibits. Qualities such as physical layout of the museum, groupings of exhibits, exhibit popularity, exhibit content, and the semantic organisation of these exhibits are aspects to be taken into account when considering exhibit relatedness (Boisvert and Slez 1995; Grieser et al. 2007; Zukerman and Albrecht 2001).

Computational measures of semantic relatedness are able to use the structure of existing knowledge bases (e.g., WordNet (Fellbaum 1998), Cyc (Lenat 1995) or Wikipedia) to identify connections between concepts. For example, synonymy in WordNet can be used to identify semantically similar word senses. These relationships are used for the purpose of emulating the cognitive organisation of human knowledge (Miller et al. 1990). For example, the use of synset glosses in WordNet can be used to identify overlapping content, and hence relatedness of meaning (Banerjee and Pedersen 2003). I extend the use of term glosses by utilising the descriptions of entities present in Wikipedia articles. These descriptions of entity content provide an
advantage over glosses used in WordNet or Cyc to describe entities: in addition to being organised in a taxonomic structure, they possess explicit reference to related articles in the form of article links, rather than relying on textual overlap to determine relatedness.

Wikipedia uses categories to group articles, thus defining a quality that is common to all articles it subsumes (Zesch et al. 2007a). Links present in the body text of an article provide an explicit statement of a relation between two articles. Also contained in Wikipedia articles are collections of attributive links that appear in the form of lists of linked information, or article infoboxes (Wu and Weld 2008).

Through the combination of inter article links and articles’ category membership, I present a purpose created measure of semantic relatedness for use with contextually diverse entities. Specifically I examine the number of categories that linked articles have in common, identifying the overlap in related categories between two articles. In this case, linked articles is the set of articles to which a source article links, i.e., its out-links. This limitation is made, partly, for reasons of efficiency: the number of in-links to an article greatly exceeds the number of out-links from that same article (Milne and Witten 2008a), thus greatly increasing the computational resources required to identify all related categories. Using solely article out-links to identify relevant categories maintains a single source of identifying links, thus maintaining a consistent level of quality; using article in-links draws from the entirety of Wikipedia, and thus from articles of varying quality. In order to maintain a consistent level of linking quality for an article’s relevant links I only use out-links in the formulation of a measure of semantic relatedness. The use of article in-links could be beneficial, but they are not explored in this thesis for reasons of computational feasibility and the desire to maintain a consistent level of quality for article links.
The number of categories that are common to the sets of categories of out-linked articles of two articles, \( a \) and \( b \), can be defined as follows:

\[
\text{Category-Overlap}(a, b) = |(\bigcup_{p \in O(a)} C(p)) \cap (\bigcup_{p \in O(b)} C(p))| \tag{4.1}
\]

where \( O(a) \) is the set of out-links from article \( a \), and \( C(p) \) is the set of categories of which article \( p \) is a member.

In cases where an article contains a large number of links, the above function (Equation 4.1) will compute higher relatedness scores due to the larger number of resulting member categories than from articles with fewer links. For example, a long and wide ranging article such as Art, contains links to a diverse range of articles (including Video Games, the Olmec civilisation, and Death Metal); without weighting, all links (no matter how briefly mentioned or insignificant) will be given equal weight. As the number of links per article is not uniform across Wikipedia, the number of links an article possesses is a factor in computing article relatedness (as there is a greater chance of obtaining category overlap). Normalisation is a common practice in statistical analysis, intended to remove unwanted influencing factors (in this case, the variability of article length). As Equation 4.1 is based on set intersection, I normalise the equation by Dice’s coefficient (Dice 1945):

\[
D(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{4.2}
\]

The final form of the Related Article Conceptual Overlap (RACO) method is given in Equation 4.3.

\[
sim_{\text{RACO}}(a, b) = \frac{2 \times |(\bigcup_{p \in O(a)} C(p)) \cap (\bigcup_{p \in O(b)} C(p))|}{|\bigcup_{p \in O(a)} C(p)| + |\bigcup_{p \in O(b)} C(p)|} \tag{4.3}
\]
resulting in a relatedness score between 0 and 1. This form of the equation is the one used for the calculations presented in Table 4.2 and Table 4.3.

The remainder of this section will cover additional measures of semantic relatedness used to relate the Wikipedia article-aligned museum exhibits. Specifically, this section demonstrates the adaptation of established measures of semantic relatedness to the Wikipedia taxonomy, and how the physical environment of the museum can be exploited to identify exhibit relatedness.

4.1.1 Physical Distance

The museum space at Melbourne Museum is grouped into themed collections by curatorial staff, placing semantically related exhibits in spatially cohesive groups, e.g., in galleries and exhibitions. Based on this themed distribution, physical distance between exhibits and galleries can be used to identify groups of exhibits with common themes. To this end, physical distance is an indication of the relatedness of exhibits as specified by the curatorial layout of exhibits. As a set of relatedness scores that is produced by domain experts (museum curators), I define the physical distance between museum exhibits as my benchmark for the task of identifying museum exhibit relatedness. In machine learning tasks, the benchmark is the method or score that any developed systems or methods aim to achieve or exceed.

Two measures of distance derived from the physical layout are used: one derived from the actual locations of the exhibits (Exhibit distance), and one derived from the arrangement of galleries (Gallery distance). Gallery distance is based on the assumption that exhibits appearing in the same gallery are considered to have a level of common content, and that the distance between galleries increases based on the separation of content (as placed by curators). This is opposed to the use
of individual inter-exhibit distance where each exhibit is considered an independent entity. To calculate the distances, a digital representation of the museum layout was placed into an SVG file and mapped onto a graph structure which preserved the physical layout of the museum (preventing paths from passing through walls or ceilings). Exhibit distances are calculated on this graph using the physical location of the exhibits in the museum space. To obtain gallery distances, the centroid of each gallery was calculated based on the physical location of all exhibits in that gallery, and then determined the gallery-to-gallery distance for each exhibit pair as the distance between the centroids of the galleries the exhibits are placed in (in the case of two exhibits being in the same gallery, the distance between them was equal to zero).\footnote{Special thanks go to Dr. Fabian Bohnert of Monash University for the creation of the museum plan SVG files, and the calculation of inter-exhibit distances. Dr. Bohnert was a collaborator on the Kubadji project, and co-author on published articles resulting from the work presented in this thesis.}

### 4.1.2 Document Similarity

As the museum exhibits used for this experiment are aligned to bodies of text (contained in the respective Wikipedia articles), another method useful for determining exhibit relatedness is to identify the similarity of the text of both documents. Document similarity is a method of identifying the level of similarity between two documents by identifying the amount of textual content they have in common. If two documents contain large amounts of identical words the bodies of text are likely to refer to similar content. To calculate the document similarity of two documents, the documents are decomposed into term vectors, and weighted by the overall occurrence of terms across an entire document set. Each article is represented as a vector of terms contained in the document, and the document similarity calculated using a\( tf \cdot idf \) weighting scheme (Salton and McGill 1983).
The term ($t_i$) frequency for a document ($d_j$) is defined as the number of times the term occurs in the document, divided by the sum of number of occurrences of all terms $t_k$ in the document $d_j$:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4.4)$$

and the inverse document frequency is defined to be the logarithm of the total number of documents ($|D|$) divided by the number of documents containing the term:

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (4.5)$$

The $tf\cdotidf$ term weight is the product of these two values, i.e., $tf_{i,j} \cdot idf_i$.

The simple cosine similarity of the document vectors is insufficient as it does not take into account the frequency of terms across a large document set. This means that documents that contain very different content could be scored as highly similar because they have many words in common that are not representative of their content (e.g., words such as the, and, or, etc.). By normalising the frequency of a term within a document by its frequency across the entire document set (using $tf\cdotidf$ weighting), the impact of very common terms is reduced.

Here, two document sources are used to describe the exhibits: (1) the Wikipedia article text to which each exhibit is aligned, and (2) the exhibit descriptions present on the Melbourne Museum website (described in Section 3.1.2). Recall that not all exhibits used in these experiments had a corresponding document on the Melbourne Museum website (see Section 3.1.2). Of the 40 exhibits used in this test set, only 26 had an equivalent document on the Melbourne Museum Website.
4.1.3 Hierarchical Relatedness

As described in Section 3.2 the article categories in Wikipedia are used to generate a hierarchy of Wikipedia categories and member articles, and a representative set of lexical similarity measures, described below, is applied to the derived graph. The majority of these measures were designed to work in regards to lexical similarity and lexical relatedness (and hence used over data structures such as WordNet), however they have been adapted to Wikipedia in previous literature (namely Ponzetto and Strübe (2007a)) and are presented here as a bank of established measures of semantic relatedness against which RACO can be compared.

In keeping with standard practice in the lexical relatedness literature (e.g., Budanitsky and Hirst (2005), Ponzetto and Strübe (2007c)), two basic approaches to the calculation of hierarchical relatedness are used: path-based and information-based measures of relatedness in addition to the RACO measure. The formulations of these measures are described in Section 2.4.2, and are revisited here to describe how they work with the Wikipedia category hierarchy.

Path-Based Relatedness

These path-based measures of relatedness consider all edges in the Wikipedia category hierarchy to have unit weight, and measure the distance between nodes to be the count of the category nodes in the path between a given pairing of article nodes in Wikipedia. This count is used as the basis for measures of path based similarity.

The simplest approach to path-based relatedness is to take the simple shortest path between two nodes (Shortest-Path), i.e., the count of the number of nodes in between two given nodes via the category hierarchy. As this is a distance measure
rather than a relatedness measure, the $\rho$ value of any correlation listed in Section 4.3.1 has its sign reversed to account for distance vs. relatedness comparisons.

This path between the two nodes will travel via a common subsumer category of the two nodes, however this subsumer may not be the deepest category which appears in both nodes’ ancestry. To ensure that the shortest path between two nodes traverses the most specific (deepest) subsuming category of both nodes, the Shortest-LCS-Path method is used: first identify the LCS (Lowest Common Subsumer) for a given pair of articles, then return the distance between the articles via that LCS. Note that the shortest path through the LCS can be different to the overall shortest path between the nodes due to the inheritance structure of Wikipedia categories (see Section 3.2.2).

The Leacock-Chodorow (Leacock et al. 1998) and Wu-Palmer (Palmer and Wu 1995) measures of semantic similarity also utilise the LCS. These two measures (described in Section 2.4.2) are also used to estimate exhibit similarity (and are readily adapted to Wikipedia, cf. Strübe and Ponzetto (2006)). Recall that Leacock-Chodorow determines similarity by scaling the shortest path between two nodes and scaling this length by the depth of the hierarchy, whereas Wu-Palmer also takes the depth of the LCS of the two nodes into consideration.

**Information Content-based Ontological Similarity**

Information Content (IC) based ontological similarity measures weight edges in the hierarchy by estimating the relative semantic difference between the concepts they represent. This is conventionally interpreted by the synset priors (for lexical ontologies), based on analysis of the token frequency of senses (e.g., relative to a pre-existing corpus of term usage such as SemCor). In this case, a separate data collection
exercise was performed (described in detail in Bohnert et al. (2009)) of which part was dedicated to tracking museum visitors as they toured the museum. This tracking exercise consisted of following a first time, single, adult visitor around Melbourne Museum as they toured the collection. Data collected included the number of exhibits used, time spent at each exhibit and path taken while touring the museum. This was followed up with a questionnaire designed to determine the visitor’s interests as well as information on their level of education and background. From these visitor tours, the prior probability of a visitor visiting an exhibit was calculated. The assumption is that exhibits appearing in the same visitor tour are alike as the visitor is interested in them both in some way, similar to the assumption used in word sense disambiguation that words that appear in the same sentence possess some semantic similarity.

Using an information theoretic interpretation of the exhibit-visit priors, we estimate exhibit similarity based on the Lin (1998) and Jiang and Conrath (1997) methods (Section 2.4.2). Recall that JIANG-CONRATH took the semantic similarity of two concepts to be the Information Content of the subsumer of both nodes with the greatest IC, and that LIN scaled the individual IC of both nodes by the IC of their Lowest Common Subsumer to obtain semantic similarity.

4.2 Evaluation Data collection

The evaluation of methods for Lexical Semantic Relatedness most often utilises standardised sets of word pairings, such as those defined by Rubenstein and Goodenough (1965) and Finkelstein et al. (2002), by comparing scores generated by these methods to gold standard scores created by domain experts. As there exists no such evaluation set for evaluating the semantic relatedness of Cultural Heritage items, one
must be constructed. This section describes the creation of such a set, and the process used to gather the expert annotations required.

4.2.1 Experiment Design

The construction of an evaluation dataset requires a set of annotators familiar with the content and domain of the experiment. The data in this set consist of pairings of museum exhibits, and those most familiar with the collection are museum curators and museum visitors. For the purpose of this data collection exercise museum visitors were chosen to evaluate the relatedness of museum exhibits, as museum visitors are the users that any personalisation would be intended to benefit (e.g., when used as a component of a tour guide system). Melbourne Museum has a membership scheme whereby visitors sign up to become Melbourne Museum members, hereafter referred to as museum members or members. Melbourne Museum allowed me to use the museum members for the task of evaluating exhibit relatedness. This was done by sending a annotation request, in the form of an online survey, to the Museum Member email list.

The survey required the members to grade the relatedness of randomised pairings of museum exhibits, from Melbourne Museum, on a scale of 0 to 4 (similar to the methodology used by Rubenstein and Goodenough (1965) in their human judgements of semantic similarity of words). The visitor was also asked to explain the reason for their relatedness judgement (what criteria they used to compare the two exhibits).

Prior to beginning the Survey, each member was asked to answer a number of questions relating to their visiting habits (e.g., how often they visit the museum, how they browse the collection, and their main reasons for visiting the museum). These questions were asked in order to identify members that were highly familiar with the
Figure 4.1: An example of an exhibit pair used in the data collection exercise. In this case, the two exhibits being compared are An Amethyst Geode and Hadrosaur Fossil. The progress bar at the bottom of the image indicates the Member’s progress through the Survey.

collection as well as to more deeply analyse results in the case that the ratings were highly varied.

After each member completed this brief questionnaire, they were presented with instructions as how the remainder of the Survey would proceed. This was followed be a series of fifteen randomised exhibit pairings of which the member was asked to gauge the relatedness. In this set of 15 pairs, 3 predetermined pairs were included to provide an indication of the agreement between annotators. The three pairs given to every annotator taking part in the survey were **Gorilla Diorama & Trilobite Fossil**, **Ant Colony & Gold Mine Model**, and **Rifles and Pistols & Sauropod**
Bone. An example of an exhibit pair as presented in the online survey can be seen in Figure 4.1. As each pair was rated, the results were entered in an SQL backend which was checked to ensure that the member was not given repeat pairings. The placing of the three predetermined pairs was also randomised in order to account for people showing a high interest in the process early on, and possibly losing interest towards the end of the Survey. At the end of the Survey, the member was able to leave any feedback, following which they were able to continue rating pairings if they wished. The survey was anonymised, but session details were kept track of in order to match starting questionnaire responses with rated exhibits.

4.2.2 Results

The survey received a response from 500 unique IP + timestamp combinations (used by the system to identify a single annotation session) over a period of 3 weeks. Each participant annotated an average of 14 pairs. This was due to some annotators choosing to end their evaluation early but partially balanced by others annotating in excess of 15 pairs. No single annotator annotated all 820 pairs. The most pairs a single member annotated was 75, however this annotator did not vary in their gradings, grading all 75 pairs as having 0 relatedness. As a result of observing the gradings of this annotator, a decision was made to remove all annotations of a member who made no variation in their gradings (i.e., graded every pair they encountered with exactly the same score). A member making uniform gradings in this manner was deemed either to have misunderstood the task, or to have deliberately attempted to skew the results. 198 annotated pairs were removed using this criteria, leaving exactly 7000 annotated pairs.

The average relatedness scores of the three exhibits pairings given to all partici-
Chapter 4: Semantic Relatedness

<table>
<thead>
<tr>
<th>Exhibit Pair</th>
<th>Relatedness</th>
<th>Mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla Diorama – Trilobite Fossil</td>
<td></td>
<td>1.53</td>
<td>1.18</td>
</tr>
<tr>
<td>Rifles &amp; Pistols – Sauropod Bone</td>
<td></td>
<td>0.50</td>
<td>0.88</td>
</tr>
<tr>
<td>Ant Colony – Gold Mine Model</td>
<td></td>
<td>1.89</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Table 4.1: Average relatedness score of exhibit pairs given to all survey participants.

pants is presented in Table 4.1 along with the standard deviation of each pair. Note that the standard deviation for each score is high relative to the respective relatedness scores, suggesting low inter annotator agreement. However, when all 820 exhibit pairings are taken into account, the agreement between annotators is much greater.

Here, annotator agreement is calculated by the Spearman rank-order correlation (see Section 2.7.1) of the scores of each annotator compared with the average score of each pair that annotator graded (with their own scores removed). The final correlation is the averaged Spearman correlation of the n-fold cross validation of all sets of annotator gradings.

The overall inter annotator agreement of exhibit pair relatedness scores gathered in this exercise was $\rho = +0.507$. Interestingly, there was variation in the inter annotator agreement when gradings were partitioned by the familiarity of the Member with the exhibits considered: Members that could recall seeing both exhibits at Melbourne Museum had a higher agreement (with other visitors that recalled both exhibits), $+0.532$, than visitors that recalled neither exhibit, $+0.393$. One potential explanation for this agreement is that visitors that are familiar with the content of the exhibits are able to make a more informed or confident decision about the specific level of relatedness between given exhibits. This contrasts with visitors that didn’t recall seeing either exhibit at Melbourne Museum that were more likely to be unfamiliar with the content of the exhibit, and hence weren’t aware of any particular relationships
that may exist between the content of the exhibits, and more likely to pick a random or lower value for exhibit relatedness. Additionally, the inter annotator agreement gives an idea of the upper-bound on achievable correlations with this data set, i.e., the scores produced by a measure can only agree with the gold standard as much as the gold standard agrees with itself.

Clustering was performed over the 820 exhibit pairs, using basic point-wise distance between exhibits. The results of this clustering using the average relatedness between exhibits is shown in Figure 4.2 and the clustering using the minimum relatedness score (selected from the set of scores provided by all annotators of an exhibit pair) is shown in Figure 4.3.\(^2\)

From these images it is evident that the annotators have given exhibit pairings relatedness scores that place the exhibits into thematic groups. Furthermore, these groups do not reflect the physical structure of the museum, indicating a conscious examination by the members of the relatedness of the exhibit content irrespective of where the exhibit is placed in the museum. For example, The Giant Squid (exgisq) and the Lake Tanganyika Aquarium (extang) received an average relatedness score of 2.75 and are clustered together as part of a larger “Living World” cluster in Figure 4.2. The full list of exhibit codes are presented in Appendix A. These relatedness scores are used as an evaluation dataset for the experiments presented in the following section.

With regard to the benchmark of physical distance between exhibits, the clustering demonstrated in Figure 4.2 and Figure 4.3 provides a demonstration of the differing perspective on the relatedness of exhibits: physical layout is used as an inference of the curators’ groupings of related exhibits, while members demonstrate the ability to provide an alternate grouping. While only an implicit statement of the relatedness of

\(^2\)In Machine Learning, the process of selecting the \textit{most} distant score between two nodes as the representative distance to be used in clustering data points is referred to as the \textit{complete} method.
Figure 4.2: Clustering of exhibits based on the average pairwise relatedness ratings. The Y axis is the dissimilarity of exhibits, and the points on the X axis are the individual exhibit codes of the exhibits in the test set of 40 exhibits (see Appendix A for definitions of each exhibit code). Coloured branches represent a closely clustered group of exhibits. For example, the Yellow branches represent fossils and the prehistoric world, and the Black branches are exhibits that revolve around forests and beaches.

exhibits, physical distance provides a set of relatedness scores based on the decisions of domain experts (museum curators).
Chapter 4: Semantic Relatedness

4.3 Experiments

In order to evaluate the different methods’ ability to score the semantic relatedness of museum exhibits, the Pearson correlation (as used in similar studies, e.g., Ponzetto and Strübe (2007c)) between each of the aforementioned measures of semantic relat-
edness (Section 4.1) and the set of gold-standard relatedness scores (Section 4.2) is used.

In addition to evaluating the overall performance of a given method, I calculate its performance over three equal-frequency bands of physical exhibit-to-exhibit distance. This was in order to investigate the hypothesis that there is a direct inverse relationship between physical distance and relatedness (i.e., the further apart two exhibits are, the less related they become), as predicted by the curator’s view of the museum. The 820 exhibit pairs were partitioned equally into the three bands: 274 pairs in the near band, and 273 pairs in each of the mid-range and far bands.

When considering the document similarity of exhibit pairs, this partitioning was altered to account for the reduced number of exhibits with corresponding Melbourne Museum web pages. Of the 40 exhibits, 26 had an equivalent web page at the Melbourne Museum website, and the results presented for document similarity over the Museum documents are based on this subset of the exhibits. Due to the smaller set of exhibits used in the Museum document similarity measure, there were only 325 overall exhibit pairs for which both exhibits possessed a corresponding Museum web page, resulting in an alternate distance-based partitioning of 109 pairs for the near distance band, 108 pairs for the mid-range band, and 108 pairs for the far band.

### 4.3.1 Results

Presented in Table 4.2 are the Pearson correlation (and p-value) between the gold-standard dataset and the relatedness estimates from each of the proposed methods. The physical distance between exhibits is used as the baseline against which the other measures of semantic relatedness are judged as it is an established set of relationships (exhibits placed in the same collection by curators are share a common
Chapter 4: Semantic Relatedness

Table 4.2: Overall Pearson correlation ($\rho$) and statistical significance (p-value) between the gold-standard relatedness scores and the various exhibit relatedness estimation methods (the highest correlation is highlighted)

<table>
<thead>
<tr>
<th>Overall</th>
<th>Overall</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ontological Similarity</strong></td>
<td><strong>Overall</strong></td>
<td><strong>Overall</strong></td>
</tr>
<tr>
<td>RACO</td>
<td>+0.404 ($1.6\times10^{-33}$)</td>
<td>+0.404 ($1.6\times10^{-33}$)</td>
</tr>
<tr>
<td>Shortest-Path</td>
<td>+0.212 ($7.5\times10^{-10}$)</td>
<td>+0.212 ($7.5\times10^{-10}$)</td>
</tr>
<tr>
<td>Shortest-LCS-Path</td>
<td>+0.133 ($1.3\times10^{-4}$)</td>
<td>+0.133 ($1.3\times10^{-4}$)</td>
</tr>
<tr>
<td>Leacock-Chodorow</td>
<td>+0.263 ($1.9\times10^{-14}$)</td>
<td>+0.263 ($1.9\times10^{-14}$)</td>
</tr>
<tr>
<td>Wu-Palmer</td>
<td>+0.009 ($7.9\times10^{-1}$)</td>
<td>+0.009 ($7.9\times10^{-1}$)</td>
</tr>
<tr>
<td>Lin</td>
<td>+0.007 ($8.4\times10^{-1}$)</td>
<td>+0.007 ($8.4\times10^{-1}$)</td>
</tr>
<tr>
<td>Jiang-Conrath</td>
<td>−0.022 ($5.3\times10^{-1}$)</td>
<td>−0.022 ($5.3\times10^{-1}$)</td>
</tr>
</tbody>
</table>

| **Document Similarity** | **Overall** | **Overall** |
| tf·idf (Wikipedia) | +0.209 ($1.6\times10^{-9}$) | +0.209 ($1.6\times10^{-9}$) |
| tf·idf (Museum) | +0.294 ($6.6\times10^{-8}$) | +0.294 ($6.6\times10^{-8}$) |

| **Physical Distance** | **Overall** | **Overall** |
| Exhibit distance | +0.196 ($1.5\times10^{-8}$) | +0.196 ($1.5\times10^{-8}$) |
| Gallery distance | +0.144 ($3.2\times10^{-5}$) | +0.144 ($3.2\times10^{-5}$) |
| Upper Bound | +0.507 | +0.507 |

To test the influence that the layout of Melbourne Museum has on visitor judgements of relatedness, the distance partitions described at the beginning of this section are used to segment this relatedness scores into three bands of equal size. These split scores are then compared to the gold-standard visitor scores which are split in the same fashion (by physical distance between exhibits). The similarity measures may be able to identify when exhibits have a lot in common (i.e., in the same gallery), but for exhibits that are separated by large distances they may be unable to identify

theme). Five measures outperform this baseline (RACO, Shortest-Path, Leacock-Chodorow, and Document similarity using both Melbourne Museum website pages, and Wikipedia articles), however the RACO measure is nearest to the Inter-Visitor Correlation upper bound (see Section 4.2). The effectiveness of these methods is further examined below.


Table 4.3: Pearson correlation ($\rho$) and statistical significance (p-value) between the gold-standard relatedness scores and the various exhibit relatedness estimation methods (for exhibit pairings of differing physical distance, based on 3-class equal-frequency discretisation; the highest correlation is highlighted in each column)


any possible relationships. The band-wise correlations (Table 4.3) identify for which measures this is the case.

The best correlation achieved with the band-wise comparison is again that of the RACO measure. RACO outperforms all other measures at all three bands, and comes closest to the inter-visitor correlation upper bound in all bands.

4.3.2 Analysis

Unsurprisingly, the physical distance based methods performed better for Near exhibit pairings (where the exhibits are generally in the same gallery) than for Mid-distance and Far pairings. This supports the hypothesis that physical distance is an effective representation of exhibit relatedness only within galleries. What was
surprising was that the best performing of the ontological similarity and document similarity methods at each band of physical distance outperformed the best of the physical distance-based methods. This is not an indictment of the curators’ placement of exhibits, but a demonstration of the differing interpretations that a visitor may take of the collection. Exhibits regarded in isolation (as they were presented in Section 4.2) take on a different meaning than when presented as part of a collection. Curators place exhibits together to highlight potentially unseen relationships exhibits which are not evident to non-experts. Falk and Boyd (1992) describe the dichotomy between curators’ views of exhibits and their relationships and visitors’ views: specifically mentioned is the difficulty in placing exhibits so that visitors, of differing levels of experience, are able to understand why a collection of exhibits is placed together. The placement of exhibits in a collection will be interpreted differently by visitors with different levels of experience (Falk and Boyd 1992), but when regarded in isolation, visitors of all experience levels are forced to regard them without any reference to the context they appear in in the museum. Being made to regard an exhibit in isolation, and hence based entirely on its own content, is a potential reason for the difference between curator placement of exhibits as part of a collection, and the visitors’ views of the relatedness of singular exhibits.

The measures utilising the simple path distance between nodes (Leacock-Chodorow and Shortest-Path) outperform methods taking into account the deepest common ancestry of the two nodes (Wu-Palmer and Shortest-LCS-Path). This is true for the overall correlation as well as at each level of the band-wise correlations.

Detailed analysis of each article’s set of ancestor categories (the set of all subsuming categories) indicates that even for relatively shallow articles the ancestor set can be as large as 1500 Wikipedia categories. When using the shortest path to the root
node to define a category’s depth, there are many instances where a category is sub-
sumed by a deeper category (the multiple inheritance used in Wikipedia’s category
can cause parent categories to have a longer path to the root node than its child;
see section Section 3.2.2). Additionally, the branching structure of the category hi-
erarchy means that there is a large subset of nodes that are common to all articles’
ancestor sets. The combination of these two factors (categories that are deeper than
their children, and a large set of categories common to all articles’ ancestor sets)
creates a situation where the LCSs arrived at for all exhibit article pairs is a set
of approximately 5 categories. In the majority of cases, this node is the same: the
category Synapsids. The next two most common occurring LCSs are Carl Jung
and Territories by Language, with only 2 pairs reaching LCSs of the desired
specificity and relatedness to the original articles (these categories were Passeri and
Coelurosaurs). The removal of these hub nodes produces an identically aligned
new set of hub nodes (meaning that there exists a large set of categories at that
depth), and the removal of a hub node just means selecting the next in the list of
candidate nodes at the desired level. Without proper analysis of how the subsuming
category relates to the two articles on an informational level, rather than solely on-
tological, the use of the LCS to identify relatedness across the Wikipedia Category
hierarchy is inappropriate.

By placing the restriction that the LCS of two articles can only be shallower than
the shallower of the two articles, a new set of LCSs arises. This set of LCSs produces
a $\rho = +0.119$ correlation for the basic LCS measure, and a $\rho = +0.004$ for Wu-
Palmer. Analysis of the LCSs reached shows that there is a single LCS for all pairs
that meet at a level. The impact of defining the maximum depth for the LCS of two
articles simply produces a hub category for each level. For example, the LCS of a
pair of articles of depths 8 and 5 will always be at depth 4 (one node shallower than the shallowest of the pair) no matter what the pair; furthermore it will always be the same category no matter what the starting articles (at depth 4 the LCS was always the Category CHECK IT, simply because it was the first in a list of previously traversed subsumers). Without specifically examining the information that is associated with the category and the articles, using the LCS to compute article association will be unsuccessful (as it will always be the same node for a given depth).

Comparing the best of the established state-of-the-art ontological similarity methods (LEACOCK-CHODOROW) with the best of the document similarity methods (tf·idf (Museum)), the difference in overall correlation is small despite the two methods using completely different elements of Wikipedia to score similarity (textual content vs. the category hierarchy). Additionally, LEACOCK-CHODOROW is much more consistent across the different distance bands, and is applicable to all exhibit pairings (recall that not all exhibits had a web page on the Melbourne Museum website). However, RACO outperforms both of these state-of-the-art measures. Analysis of the band wise separation of scores (Table 4.3) demonstrates that the RACO measure is superior to all other tested measures across respective bands. The band with the smallest correlation using the RACO measure (the mid-range band at $\rho = 0.304$) achieves comparable performance to the next best correlation across all bands (the near band of tf·idf (Museum) at $\rho = 0.318$).

One can notice a pattern in the band-wise correlation obtained for the path-based measures: measures using the shortest path distance (SHORTEST-PATH, LEACOCK-CHODOROW) do not have as high a correlation with visitor judgements in the mid-range band as at the Near and Far distance bands. One potential explanation is that the measures are able to identify when a pair is highly similar or dissimilar (very near
or far), but mid-distance exhibits being borderline cases, because the galleries containing the two exhibits are both members of a larger exhibition (e.g., Dinosaurs and Human Anatomy galleries are both members of the Natural Science exhibition).

The performance of the Information Content based measures have the potential to exceed the performance of the path based and document similarity measures, as they directly take into account the amount of information shared between two concepts. This insight is only gained with the use of an appropriate corpus. It is evident that the corpus of exhibit visit frequencies did not provide the correct insight into common groupings of exhibits. The criteria of first time visitors to the museum meant that the visitors were unfamiliar with the exhibit space and considered all locations equal, resulting in a distribution that gave a higher Information Content to more visible exhibits and a lower Information Content to more obscure exhibits.

The category subsumption problem present with the LCS-based measures is again the primary cause of the failure of these methods. With the majority of categories being present in all articles’ ancestor nodes, the IC of many categories approaches 1. Thus the IC of many LCSs provides no additional information in calculating either the Jiang-Conrath or Lin scores.

In contrast to the Information-based measures, the measures of document similarity demonstrate that a correlation exists between article content and opinions of exhibit relatedness. Both document similarity measures outperform the physical similarity baseline over all pairs, and the Museum document similarity measure shows the second highest correlation for the near band. This is an unsurprising result when one considers that the documents have been tailored to the specific context of the museum (and hence contains some information that the visitor may have already encountered in the museum space). This performance is not present in either the mid-distance or
far bands, where the $tf \cdot idf$ score using Wikipedia documents shows a higher correlation with the survey pairs. The cause of this anomaly is that documents for exhibits in a given gallery (which inevitably end up in the near band of exhibit pairings) are generally authored by the same gallery curators, and hence a better reflection of the relative exhibit relatedness. As we cross gallery boundaries, however, this effect disappears, and the general-purpose Wikipedia documents are a superior representation of exhibit content. Additionally, the relative sparsity of documents present on the Melbourne Museum website makes it less preferable to Wikipedia when being used for further exhibit alignment.

These two measures exclusively examine the content of the documents, not taking into account the organisation of these documents. Even with this restriction, the $tf \cdot idf$ measures outperform the majority of ontological measures. However, by regarding direct links between articles and their ontological grouping, RACO offers a significantly higher correlation with the exhibit relatedness scores.

### 4.4 Evaluating RACO Over Existing Data

In this section I present the use of RACO over more widely used datasets, to which the measures described in Section 4.1 are better suited. The word pair sets of Rubenstein and Goodenough (1965) (R&G) and the WordSim353 set (Finkelstein et al. 2002) are used for this experiment. Agirre et al. (2009) split the WordSim353 into separate Similarity and Relatedness datasets, consisting of 203 pairs and 252 pairs respectively. Using these sets, I demonstrate the performance of RACO in comparison with the WLVM measure of semantic relatedness (Equation 2.7).

As with the museum exhibit dataset, the words used in the R&G and WordSim353
must be aligned to Wikipedia articles for comparison to take place. I also present the
evaluation of RACO and WLVM using two independently manually aligned versions of
the WordSim353 dataset in order to provide contrast with the automated alignment
of entities using a diverse knowledge base (Wikipedia).

As reported in Milne and Witten (2008a) manual alignment of words to spe-
cific Wikipedia articles improves the performance of semantic relatedness measures.
However this experiment demonstrates the vulnerability of Wikipedia due to the large
number of Named Entities present in disambiguation articles in that they may be able
to draw spurious, but strong, connections between articles present in disambiguation
articles.

The experiments presented in this section correlate scores generated by the RACO
and WLVM methods with human judgements in line with the methodology described
by Budanitsky and Hirst (2005). In this case the correlation used is the Spearman
rank-order correlation coefficient. The word pairs in each dataset are aligned auto-
matically by the following method:

- To identify the article name, the word is normalised by Wikipedia convention
  (capitalising the first letter). If no article exists by that name, it is unalignable
  and all pairs that contain this word are assigned a score of zero.

- If the resulting article is a redirect, then the redirect link is followed until an
  article or a disambiguation article is reached.

- In the case that a word aligns to a disambiguation article, all articles that the
  word disambiguates to are used to compute the score. The disambiguation that
  yields the largest score is used as the representative score for that pair.

In addition to this automated alignment, two sets of manual alignments for the
Table 4.4: The Spearman correlation between computational measures of similarity (raco and wlvm) and human gold standard judgements over two datasets. For the WS353 set, multiple annotation methods were used: automatic computational alignment (auto), and manual alignments performed by two annotators (ann-1 and ann-2).

WordSim353 set are also used (as well as for the Relatedness and Similarity partitions of that dataset). Both annotators identified the single Wikipedia article that they believed best represented the sense expressed by the word in a word pair. In 86% of cases, both annotators aligned a word to the same Wikipedia article.

The results of the raco and wlvm methods using these three alignments are presented in Table 4.4. The best performing score for each dataset (and partition) is shown in bold font. The scores were calculated using the September 9th, 2009 Wikipedia dump.

As expected, the overwhelming trend is that manual alignment improves the correlation of both measures, however the degree to which the measures are affected differs greatly depending on the dataset and the annotator. Most interesting is that both raco and wlvm perform better at the task of semantic similarity than semantic relatedness. When the WordSim353 set is partitioned, both measures show a higher
correlation with the Similarity partition. This is further backed up by both measures’
correlation with the R&G word pair scores, which is comprised of semantically similar
word pairs. Wikipedia provides many links to related articles and these articles are
often not semantically similar, merely just related. The facility for determining the
relatedness of two articles exists within Wikipedia’s many link types, however in this
case the measures used perform poorly at the task.

In all cases WLVM outperforms RACO, but overall, both measures show a greatly
improved correlation when provided with manual alignments. The subset of 252
relatedness word pairs shows a low correlation with human judgements when using
automatically aligned articles, but shows an improvement when using human aligned
scores. The improvement in correlation when using the Similarity partition of the
dataset also demonstrates improvement. One possible interpretation is that selecting
the highest scoring sense pair for two words is a valid methodology for computing
semantic similarity as a similar article will contain many similar links to the same
articles (and most likely each other), but that there exists a degree of relatedness
between all articles within Wikipedia, and that arbitrary relationships can cause
higher relatedness scores than human judgements.

The effect of manual alignment on word pairs prior to their relatedness score
being computed will invariably improve their score as the annotator is aware of which
particular sense is being referred to by the word. It remains to be seen if this trend
exists in relation to other manually annotated datasets.

Despite the encyclopedic quality of Wikipedia article content (which extends be-
yond the simple description of entity meaning) and the numerous links present be-
tween articles, both measures of relatedness performed better at the task of semantic
similarity than at the task of semantic relatedness. This is demonstrated by the cor-
relations over the Rubenstein & Goodenough dataset and the similarity portion of the WordSim353 dataset in comparison to the relatedness partition of the WordSim353 dataset.

This experiment is only provided as a brief aside to demonstrate the effectiveness of Wikipedia based measures of semantic relatedness over established datasets and the effect that manual alignment has on measure performance. This experiment could be expanded to examine the effect of removing Named Entities from the Wikipedia disambiguation pages, and further manually aligned datasets are needed to identify if there is any significance in the manual versus automatic alignment improvement. The contrast between manual and automated disambiguation of word senses over Wikipedia has the potential to be expanded, and while this is not further explored in this thesis, provides a valid area for future research.

4.5 Application of RACO to Topic Labelling

Recently, the field of topic modelling has emerged as a popular research area in Natural Language Processing. Topic modelling is a form of probabilistic classification of documents: for a given document collection, the multinomial distribution of terms in that collection is used to calculate the probability of a document belonging to a group of terms, or topic. Each topic is in essence described by a probabilistic distribution of terms, these terms are clustered using the process of Latent Dirichlet Allocation or LDA (Blei et al. 2003). For example, the co-occurrence of the terms pitcher, batter and base in multiple documents may indicate that these terms are representative of a single topic (e.g., a topic based on BASEBALL). Thus, any documents that contain a large number of these terms are likely to be members of the
same topic. Typically, a topic is described in terms of an ordered list of the top $N$ ranked terms (ranked by probability of being relevant to the topic: Blei et al. (2003), Griffiths and Steyvers (2004)).

When presented as a list of terms, a burden of interpreting the subject of this topic is placed on the end-user encountering the topic. Given a sufficiently diverse set of terms that represent a topic, this interpretation can be a complex task, subject to the understanding of the end-user. For this reason, it is beneficial to explicitly label the topic with a title that best describes its content. For example, the set of terms stock market investor fund trading investment firm exchange companies share, would be the top 10 terms from a topic relating to Stock Market Trading. However, the labels Stock Market, Stock Investment or Stock Trading may also be equally valid, but may be too general, too specific or not entirely capture the nuanced subject of the topic. Determining a label for these topics is a subjective task that makes human annotation difficult. To eliminate this subjectivity, there have been recent studies aiming to automatically supply appropriate labels for topics (Mei et al. 2007; Magatti et al. 2009; Lau et al. 2011).

Existing methods of applying labels to topics have used methods such as the selection of the first ranked term in the topic (e.g., stock would be the topic label in the above instance, as it is the first term in the ranked list of topic relevant terms) (Hoffman 1999; Lau et al. 2010), or simple post-hoc manual annotation of topics (e.g., as used by Mei et al. (2006) as an aid to describe the outcome of topic modelling experiments). However, in the former case, these labels may not capture the entire nuanced meaning of a group of terms, and in the latter case are subjective and are not reproducible.

During the course of this thesis, I collaborated on work targeted at the task of
automated topic labelling. I present here a brief overview of this work, and present a further analysis of the contribution of RACO to this work. The principle contribution of the work presented by Lau et al. (2011) was the use of the set of Wikipedia article titles as a set of candidate labels that could be used to describe the content of topics. The 350,000+ article titles provided a large set of candidates, covering a wide variety of content from which to select labels for topics.

The approach taken by Lau et al. was to identify a set of primary candidates for the label of a topic by searching for articles within Wikipedia that contained terms from the topic being labelled. This was done by performing a basic document search using (1) a Google query of the top 10 ranked terms of the topic, over the Wikipedia article collection, and (2) the same 10 terms used as a query in the native Wikipedia search API. For each of these queries, the top 8 ranked article titles were used as the primary candidates for that topic, producing between 8 and 16 primary label candidates (article titles were only included once if they were discovered by both the Google query and the Wikipedia query).

In order to increase the number of candidates labels for each topic and include more generalised and specific versions of article titles, the set of primary candidates was further used to generate a set of secondary candidates: each primary candidate was broken down into individual noun chunks that occurred as article titles. These chunks were then recombined to produce further, multi-chunk secondary candidates, these combinations were subject to the condition that the chunks appear in the same order that they appeared in the primary candidate. For example, given the primary candidate STOCK MARKET TRADING, generated secondary candidates would be STOCK, MARKET, TRADING, STOCK MARKET, MARKET TRADING and STOCK TRADING, and the combination TRADING STOCK would be invalid. Using this method, an
additional 30-40 secondary candidates were generated per topic. The set of the top
5 words for each topic was also included as candidate labels for each topic, provided
that they existed as article titles in Wikipedia.

To identify a set of topics for which labels were selected, 4 existing topic modelling
datasets were used to identify a set of topics.

- A collection of blog articles from August to October 2008, taken from the
  Spinn3r blog dataset.\(^3\)

- A collection of English language books from the Internet Archive’s American
  Libraries selection.\(^4\)

- A collection of New York Times news articles from July to September 1999,
  from the English Gigaword corpus, a corpus of 1.2 billion words of English
  taken from news articles.


These are referred to, respectively, as BLOGS, BOOKS, NEWS, and PUBMED. The
specifics of the topic identification process is described in detail in Lau \textit{et al.} (2011); this section describes the use of RACO in the task of identifying topic label candidates,
not the topic modelling process itself.

\subsection{RACO Thresholding}

Due to there being a large number of primary and secondary candidates extracted
for each topic, it was prohibitively expensive to manually evaluate all candidate labels
with respect to the topic they described. To reduce this set, the set of secondary

\footnote{\url{http://www.icwsm.org/data}}

\footnote{\url{http://www.archive.org/americana}}
candidates was pruned using RACO to calculate their semantic relatedness to the primary candidates for that topic, and discarding all secondary candidates that fell below a certain threshold.

The RACO score for each secondary candidate was calculated by determining the average RACO score between the secondary candidate (automatically aligned to a Wikipedia article) and all primary candidates for that topic. In the case that a secondary candidate was aligned to a disambiguation article, the RACO score between the primary candidate and all disambiguation targets was calculated, and the highest selected.\footnote{Due to an oversight, the formulation of RACO used in this study neglected to include the $2 \times$ multiplier used in Dice’s coefficient, thus causing each calculated RACO score to be between 0 and 0.5, rather than the intended 0 and 1 as intended. This is a cosmetic difference that does not impact on the deeper conceptual overlap formulation of RACO, nor the basic normalisation premise of Dice’s coefficient. Hereafter, the results are reported as what they would have been using the correct formulation of RACO.}

After a manual analysis of the RACO score calculated for each secondary candidate, a threshold of 0.2 was selected as the point below which candidates would be discarded. In an effort to determine if this threshold was a well placed one, I present a manual analysis of the appropriateness of this threshold based on a sample of secondary candidates taken from just above this threshold, and below it. This experiment is also presented as a more general test of RACO’s performance in evaluating the quality of topic labels compared to equivalent human judgements: does there exist a positive correlation between RACO scores of candidate labels and human judgements of the quality of these same labels.

As the threshold was chosen after manual analysis of the RACO scores between primary and secondary candidates, there exists the possibility that the threshold was not an ideal one, thus eliminating desirable candidate labels from the pool. To determine the validity, I present a statistical analysis of the candidate label scores.
computed by RACO and by manual evaluation of two samples of candidates: one sample from below the threshold, and one sample from above the threshold.

To evaluate the candidate labels above and below this threshold, I use the same 4 point evaluation scale (0-3) used by Lau et al. (2011):

0. Label is completely inappropriate, and unrelated to the topic.

1. Label is semantically related to the topic, but would not make a good topic label.

2. Reasonable label, but does not completely capture the topic.

3. Very good label; a perfect description of the topic.

I select 13 secondary candidates from each dataset (BLOG, BOOKS, NEWS and PUBMED) that obtained a RACO score of just above 0.2, and 13 that obtained a score around 0.1, producing a total of 52 candidate topic label samples at 0.2 and 52 at 0.1. The order of this selection of 104 candidates was randomised and distributed to 8 annotators, each annotating half of the data, thus producing 4 complete evaluations of the dataset. Each annotator evaluated the appropriateness of each candidate against its respective topic using the above 0 to 3 scale. The annotators used for this evaluation task were Computer Science postgraduate students, hereafter referred to as the “In-House” annotators.

4.5.2 Results

The overall inter-annotator agreement between the four sets of annotations, calculated using Fleiss’ Kappa (see Section 2.7.2), is 0.34. A possible factor in this low agreement is the use of a numerical scale to evaluate label applicability: despite a very
Table 4.5: Inter-annotator agreement of evaluated topic labels, above and below the 0.1 threshold used by Lau et al. (2011), calculated using Fleiss’ Kappa.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Above</th>
<th>Below</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOGS</td>
<td>0.53</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>NEWS</td>
<td>−0.06</td>
<td>0.76</td>
<td>0.45</td>
</tr>
<tr>
<td>BOOKS</td>
<td>0.01</td>
<td>0.41</td>
<td>0.26</td>
</tr>
<tr>
<td>PUBMED</td>
<td>0.01</td>
<td>0.23</td>
<td>0.13</td>
</tr>
<tr>
<td>OVERALL</td>
<td>0.19</td>
<td>0.47</td>
<td>0.34</td>
</tr>
</tbody>
</table>

small level level of difference in an annotator’s evaluation (e.g., one annotator selected a 0, while another annotator selected a 1 for the same label), using Fleiss’ Kappa, they are treated as independent classes. Upon reviewing individual annotations, there existed 10 cases where two annotators’ evaluations of a label’s applicability differed by 2, and one case in which the evaluations differed by 3 (i.e., one annotator evaluating a label’s applicability to a topic as 0 and another evaluating that same label as a 3). Despite there existing very few cases where annotator judgements differed by a large amount, agreement is low due fine-grained distinctions in annotator judgement as each score is treated as a separate class rather than a point on a scale. As an additional method of evaluating the agreement and consistency of annotator evaluations I use another set of evaluations over the same data for comparison. A dataset containing Mechanical Turk evaluations of all generated candidate labels was published as part of the work by Lau et al. (2011); this dataset only included the evaluation of secondary candidates above the 0.2 threshold. In this section, an analysis of these evaluations is included as a comparison to the In-House annotation.

The individual segmentations of inter-annotator agreement, calculated using Fleiss’ Kappa (see Section 2.7.2), separated by dataset and score sample (at 0.2 and 0.1) are given in Table 4.5.
When the candidate labels are separated into their scoring range, a pattern presents itself: the overall agreement for the topic candidate evaluations at the 0.1 level is greater than that at the 0.2 level. When examining the agreement of annotators over individual datasets, this case remains true (excepting the BLOGS dataset), with three out of the four datasets showing an increased agreement for the topic labels with lower RACO scores. The BLOGS dataset is the only dataset for which the agreement remains respectable, and roughly similar, over both score bands. When examined in combination with the average evaluation scores extracted for each partition (seen in Table 4.6), a potential explanation for this disparity in agreement arises.

Over all partitions of the candidate labels, the average topic label score was between 0.85 and 1.42, the lower end of the evaluation scale. When combined with the variability in annotator agreement, the indication given by the results is that while all labels at these ranges were poor labels on average, annotators were able to agree that labels taken from the 0.1 RACO score sample were at the bad label end of the rating range, while there was disagreement on whether labels just above the RACO threshold used could be used as appropriate topic labels. For example, the topic described by the terms art, artist, picture, painter, painting, paint, museum, drawing, study, and colour from the iABooks dataset identified MARY CASSAT as a candidate label.

Table 4.6: Average evaluation score for label candidates, separated by data set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Above</th>
<th>Below</th>
<th>Turk (Above)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOGS</td>
<td>0.85</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>NEWS</td>
<td>1.31</td>
<td>1.08</td>
<td>1.21</td>
</tr>
<tr>
<td>BOOKS</td>
<td>1.33</td>
<td>1.40</td>
<td>1.23</td>
</tr>
<tr>
<td>PUBMED</td>
<td>1.42</td>
<td>1.29</td>
<td>1.64</td>
</tr>
<tr>
<td>OVERALL</td>
<td>1.22</td>
<td>1.19</td>
<td>1.22</td>
</tr>
</tbody>
</table>
Chapter 4: Semantic Relatedness

Table 4.7: Correlation between annotator evaluations and RACO scores, calculated using Spearman correlation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>In-House–Turk</th>
<th>In-House–RACO</th>
<th>Turk–RACO</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOGS</td>
<td>+0.27</td>
<td>+0.31</td>
<td>+0.58</td>
</tr>
<tr>
<td>NEWS</td>
<td>+0.45</td>
<td>0.00</td>
<td>+0.27</td>
</tr>
<tr>
<td>BOOKS</td>
<td>+0.46</td>
<td>+0.59</td>
<td>+0.09</td>
</tr>
<tr>
<td>PUBMED</td>
<td>+0.52</td>
<td>−0.04</td>
<td>+0.02</td>
</tr>
<tr>
<td>OVERALL</td>
<td>+0.06</td>
<td>−0.05</td>
<td>+0.26</td>
</tr>
</tbody>
</table>

All four annotators gave this label a different rating (0, 1, 2 and 3). Mary Cassatt was an American Impressionist painter, and based on personal knowledge individual annotators may have been able to identify this connection to the painting profession. This trend also existed in the evaluations gathered through Mechanical Turk: the Mary Cassatt label was evaluated once as 0, twice as 1, twice as 2, and once as 3. This topic label was obtained a RACO score just above the 0.2 threshold. Incidences of this occurring in the 0.1 range label candidates did not occur in the sample chosen for this experiment. That is not to say that all labels at the 0.1 range are bad, merely that, based on this sampling, ambiguous labels are statistically less likely to occur.

The correlation between manual evaluation and RACO scores of candidate labels varies between dataset, with both the data annotated by Turkers and the In-House annotators demonstrating a positive Spearman correlation with RACO scores for some datasets and zero correlation for others. The full breakdown for correlations between RACO and each evaluation source, by dataset, can be seen in Table 4.7. Despite achieving a level of agreement with RACO over some datasets, both sets of evaluation achieve a low overall correlation with the RACO scores and disagree with one another. This is evidenced by the $\rho = +0.06$ correlation between Turk annotations and In-House annotations. This result is not erroneous: the correlation between annotations
over individual datasets are positive, but scattered. The $r$ values (the likelihood of another distribution achieving the same correlation) for each correlation are relatively high, ranging between 0.07 for the PUBMED dataset and 0.37 for the BLOGS dataset. When all datasets are combined this scattering is compounded, resulting in a dataset with such a great variance that no correlation is present. This is backed up by regression analysis of the two sets of annotations: when these evaluations are plotted, the line of best fit has a gradient of 0.005, again indicating that there exists no relationship between these two sets of annotation. Despite this disagreement, each of the annotation sets achieve a positive correlation with RACO scores over different partitions of the candidate labels.

To determine the overall trend in the correlation between each set of annotation scores and RACO, I perform regression analysis over the individual RACO–Turk and RACO–In-House score pairs. Figure 4.4 plots these points, and identifies the lines of best fit for both sets of annotation. Note that the range of plotted values over the x-axis of Figure 4.4(a) (RACO score) is limited, with the only one data point achieving a RACO score above 0.5. This is largely due to the composition of RACO: unless both articles contain almost entirely identical sets of links, the set of related categories for each article will differ greatly due to the large number of potential categories of which articles can be members, resulting in a small amount of category overlap. To clarify the overall relationship between these two sets of scores, regression analysis is useful.

In both cases, it can be seen that there exists a positive relationship between the evaluations and RACO scores. This increase is slight in both cases: the evaluations provided through the In-House annotation were over a narrow range of RACO scores, and the differences in average scores for each sample over all datasets (shown in Table 4.6) were minor. However, when individually compared with the individual
Figure 4.4: Linear regression plots of candidate label scores.

RACO scores for each label, the slight positive relationship between the In-House annotation and the RACO scores becomes apparent. The gradient of the line of best fit to the Mechanical Turk evaluations is 1.22, greater than that of the In-House
annotation (0.53) indicating a stronger positive trend between the two score sets. Additionally, the broad scattering of data-points (shown in Figure 4.4(a)) indicates that there still exists some discrepancy between manual evaluation of the points, and the RACO scores. The $R^2$ values for each of these lines of best fit are 0.01 for the Mechanical Turk scores and $3.30 \times 10^{-4}$ for the In-House scores, where $R^2 = 1$ indicates a perfect fit with the data, and $R^2 = 0$ indicates no linear relationship between the two score sets. Ideally, the gradient for each of these lines of regression should be closer to 3 (i.e., a RACO score of 1 equalling an annotator score of 3), but their shallow gradient indicates (combined with low $R^2$ values) that there exists variability between what annotators consider valid, and equivalent RACO scores.

A potential explanation for this variability is, again, the ambiguity present in some label candidates. Recall that the label of MARY CASSATT possessed this ambiguity over both evaluations. The Mechanical Turk evaluations were not performed on label candidates below the 0.2 threshold, so it is unknown if this ambiguity would persist below the threshold, or whether annotator agreement would increase at lower scores as the In-House annotations did.

This section has examined the use of RACO in generating label candidates for topic models. RACO was used to calculate the semantic relatedness of candidate labels generated from Wikipedia articles. In this case, Wikipedia’s diversity makes it desirable as a source of topic labels and RACO is able to provide a reliable method of pruning unrelated label candidates. When manual evaluations are performed at different samples of candidate labels, it can be seen that at lower RACO scores labels are more assuredly judged as inappropriate, but at RACO scores closer to the threshold used there exists ambiguity. Furthermore, there exists a positive relationship between the RACO score of a candidate label and the manual evaluation score of the same label,
indicating that RACO provides a reasonable approximation for identifying determining
the validity of topic label candidates.

4.6 Summary

Using a comprehensive ontology of information-rich documents allows one to iden-
tify associations between concepts that are not present in simple document collections.
Link structure and category membership of Wikipedia articles are used to measure
the strength of relationships between articles. The evaluation data took the form
of relatedness scores between pairs of museum exhibits, in an experiment similar to
Rubenstein and Goodenough’s word pair experiment. By aligning a collection of
museum exhibits to corresponding Wikipedia articles, the correlation between the
survey respondent scores and the relatedness scores produced by a set of state-of-
the-art similarity measures was measured. The methods of SHORTEST-PATH and
LEACOCK-CHODOROW (both using Wikipedia’s category hierarchy) and the \( tf\cdot idf\)-
weighted cosine similarity of the Wikipedia articles, exceeded the performance of
the baseline metric of physical distance between the exhibits within Melbourne Mu-
seum. Additionally, the proposed RACO measure of article similarity over Wikipedia
that utilises the category overlap of article out-links significantly outperforms these
state-of-the-art similarity measures, and approaches the upper-bound correlation de-
termined by the inter visitor agreement.

The arrangement of the exhibits at Melbourne Museum is based on thematic
areas, placing exhibits with common content together. This placement represents
the classification of exhibits from a curatorial viewpoint, and in many cases these
exhibits are constructed specifically to fit within a gallery theme. Even when created
specifically for a gallery or exhibition, these exhibits still maintain connections with exhibits in other collections within the museum, or even in other Cultural Heritage sites. By using measures such as RACO to identify exhibits that are closely related, tour planning systems, such as those described in Chapter 6, can be used to create tours which centre around visitors’ interests.

When this methodology is extended to the existing task of word sense disambiguation, the diversity of content present in Wikipedia, which makes it desirable for measurement between explicit articles, makes the task of disambiguation difficult. The large variety of term senses present in Wikipedia disambiguation articles, combined with the diversity of inter-article links, causes a situation in which a pair of disambiguation articles will be able to identify a pair of senses that is highly related or similar (even in the case where human annotators would judge the same pair of terms as unrelated). Further experimentation is required to determine if this over-diversity makes Wikipedia unsuitable as a disambiguation resource.

RACO has been successfully used as part of a larger system to identify topic labels, and in this RACO was used to prune topic labels that fell below a certain threshold. In an analysis of this threshold, I have demonstrated that RACO provides a valid indication of the applicability of topic labels. This is a further demonstration of the advantage of using RACO as its use of Wikipedia as a basis for calculations allows it to be used on tasks where diversity and size of content is an important factor.

In extending the ability to identify how much two exhibits have in common, it is possible to identify the nature of the relationships that exist between exhibits. Computational measures of relatedness such as RACO or WLVM contain explicit reference to elements of Wikipedia that identify relatedness (e.g., Links and Categories). Existing studies (e.g., Nastase and Strübe (2008)) have used language analysis to
identify subsumption relationships within Wikipedia, but do not extend to the other semantic content provided by Wikipedia. In Chapter 5, I extend the practice of using Wikipedia documents to identify specific inter-article relationships by co-opting qualities that comprise measures of semantic relatedness as described in this chapter.
Chapter 5

Semantic Relationship Extraction

The identification of specific reasons for relatedness can aid in demonstrating the links that exist between conceptually diverse entities. Inter-entity relationships can be identified by examining attributive overlap by comparing the attribute values of entity descriptions as performed in studies of exhibit comparison such as Wang et al. (2008), Stock et al. (2005) and O’Donnell et al. (2001). These projects aim to utilise the overlap in attribute value to identify museum exhibits with a common conceptual structure (e.g., paintings or sculptures) to identify groups of museum exhibits that pertain to a common theme. These relationships can further be used to describe how exhibits relate to a visitor’s interests when planning a museum tour (Aroyo et al. 2007).

Entities with diverse conceptual structure possess multiple relationships with other diverse entities, but identifying these relationships presents an additional challenge over that of the identification of relationships between entities with a common conceptual structure. Diverse collections of entities, such as the multiple exhibit types that exist in a natural history or science and technology museum, require digital rep-
resentations that are able to encapsulate the diverse nature of their exhibits at an appropriate level of detail (Keene 1998), however, there are no universally accepted standards of exhibit representation (Tedd and Large 2005). In this chapter, I present a solution to the problem of identifying relationships between diverse, information-rich entities by using an existing encyclopedic ontology as a representation of the content of these diverse entities and their taxonomic organisation.

The exhibits at Melbourne Museum represented by the Wikipedia articles Giant Squid and Dendroclimatology are conceptually diverse exhibits that, at first glance, do not have anything in common. However, from the methods used in this chapter, The age of a giant squid can be determined by “growth rings” in the statocyst’s “statolith”, similar to determining the age of a tree by counting its rings, can be identified as a valid relationship between these two entities. Through the use of an appropriate knowledge source and exploitation of the organisation of this knowledge, relationships that describe complex interactions between conceptually diverse entities can be identified and extracted.

Chapter 4 utilised Wikipedia articles to evaluate the strength of relatedness of contextually diverse entities. The measures presented in that chapter were able to assess semantic relatedness by identifying the categories with which each article was associated. The attributes and compositional elements of each article used to calculate this relatedness also can be used to provide knowledge of the nature of the semantic relationship that exists between two objects. Wikipedia possesses multiple connections between articles in the form of article lists, sub-articles, article links, article categories, and inherited category ancestry. Some of these features describe relationships by themselves (e.g., using immediate category membership, one can describe Saturn and Jupiter as both being Gas Giant Planets), identifying simple is-a
Chapter 5: Semantic Relationship Extraction

relations. The example presented at the beginning of this chapter uses a combination of inter-article links, and category ancestry to identify the relationship.

In this chapter, I describe how the information structure of Wikipedia articles and the Wikipedia Category hierarchy can be used to extract semantic relationships between articles and how the taxonomic links used as the basis of the RACO measure of semantic relatedness (presented in Chapter 4) can be further exploited to identify these relationships. Section 5.1 describes the process of adapting features of Wikipedia for use in identifying the semantic relationships between articles. These features are then adapted into methods for the extraction of semantic relationships between entities, and the resulting relationships extraction task is described in detail in Section 5.2. The evaluation of the extracted relationships is a problem that required multiple iterations to consolidate; the development of a process for evaluating extracted relationships is described in Section 5.3. The outcome of the relationship annotation process is described in Section 5.4. Finally the content presented in this chapter is summarised in Section 5.5.

5.1 Relationship Identification

The measures of semantic relatedness presented in Chapter 4 utilise specific elements of Wikipedia articles and the articles’ organisation to determine the degree of relatedness between articles. The links between articles in Wikipedia are not featureless, unlabelled links: they contain valuable information about the nature of the relationships between articles (Kamps and Koolen 2008). For example, articles that are both members of the same category can be classified as both being a type of that category, just as the articles Jupiter and Saturn are both members of the
Gas giant planets category. Hence the relationship that exists between Jupiter and Saturn is that they’re both Gas Giant Planets. Using the measures presented in Chapter 4 this section describes the adaptation of the article components used to score article relatedness to the identification of specific relationships that exist between articles.

5.1.1 Category Membership

Wikipedia’s hierarchical category membership is a basic system of inheritance and, while not adhering to a strict system of is-a relations, still preserves the property of category generalisation: specific categories are subsumed by categories that possess a broader classification. For example, the category Crops originating from China is subsumed by the category Crops by country; note that this subsumption is not an instance of an is-a relation, but does provide a more generalised classification of the original category. In this respect the identification of common category membership between a pair of articles presents the basis for a relation between them. This relation is untyped due to the non-uniformity of inheritance type present in the Wikipedia category hierarchy.

Due to the high degree of separation between semantically similar articles at deeper levels of the Wikipedia hierarchy (see Section 3.2.2 for a discussion of Wikipedia category membership policies), immediate category overlap is insufficient to identify common ancestor categories at an appropriate level of specificity. To extend the search for common ancestor categories beyond the immediate parent categories of a node, I introduce a rudimentary scaling factor to determine how many levels of ancestor categories are included when determining the semantic similarity of two articles. The additional levels of category ancestry considered are expressed by the relative depth of
two articles \( A \) and \( B \) compared to the overall depth of the taxonomy (Wikipedia, \( W \)). The number of levels of category ancestry used when determining category overlap of two articles is given by the square of the sum of the depths of the two articles \( (d(A) \) and \( d(B) \) respectively) scaled by the square of the total depth of Wikipedia \( (d(W)) \):

\[
\text{Ancestors}(A, B) = \text{int} \left( \frac{(d(A) + d(B))^2}{d(W)^2} \right)
\]

(5.1)

The final result of this equation is rounded to the nearest integer in order to discretise the number of ancestor nodes considered (i.e., fractions of ancestors are nonsensical).

This scaling factor is intended to account for the increased branching factor present at deeper levels of the Wikipedia hierarchy, and thus in identifying a reasonable search distance from deep nodes to find common ancestor categories. From examination of the formula, when both nodes are close to the maximum depth of the hierarchy, the potential search space for ancestor nodes can extend beyond the root node (the \textsc{Fundamental} category). However, in practice, this is rare and for the purposes of this experiment does not exert undue influence on the results.

Using this scaling metric ensures that deeply buried nodes are able to generalise a sufficient amount to find relationships, and very general articles (close to the root category node) only consider their immediate parent categories when computing category overlap. This scaled ancestry use limits the effect of the inheritance cycling discussed in Section 2.4.1: at shallower depths, only immediate parent categories are considered, and the category network is much less densely populated the deeper the article is (Yu \textit{et al.} 2007).

Another potential approach to determining the distance from each article to search for relevant ancestry is to utilise the information content of articles and their common ancestors (see Section 2.4.2 for a discussion of Information Content with relation to
Figure 5.1: Common ancestry membership. In this case, only Category 2 is a common ancestor of Article 1 and Article 2 as it is the only category that subsumes both articles. Thus Article 1 and Article 2 are deemed to be related as they are both classified as types of Category 2.

hierarchical semantic networks). However, due to the lack of a corpus with which to populate the information content of each article and category in the hierarchy, this is infeasible. For this reason, Equation 5.1 is used to determine the relevant search depth for article pairs.

I utilise the property of common ancestry to identify the relationships between article pairs.

5.1.2 Article Text

As Wikipedia is an encyclopedic resource, the majority of the content present on Wikipedia is text that comprises the body content of the article. In the case of some articles, this text may only be a stub containing a sentence or two, requiring further expansion (Wikipedia 2011h). In other cases the article can extend to many thousands
words of text, requiring sections of an article to be placed into their own separate articles. For example, the article Ecology has a subsection titled Ecological niche and the habitat which contains a several paragraphs on this topic, and links to the two separate main articles Ecological niche and Habitat for further explanation.\footnote{Wikipedia has a template for instances where sections of articles contain enough material to support their own article. The guideline states that the sub-article should only be described briefly in context of the article in which it appears and the main article on this topic is referenced using the {{Main}} template.}

Article text describes the relationships that exist between the current article and other articles or categories (Milne and Witten 2008a). These relationships may be directly specified by containing a link to another Wikipedia article with the nature of the relationship described by the surrounding text, or they made be referred to without the explicit reference to an article. The Wikipedia Linking guidelines (Wikipedia 2011d) state that an editor should In general, only link to the first occurrence of an item, leading to a highly relevant article only being explicitly linked to once in an article. Only utilising explicitly linked articles means that repeat instances of relevant content are ignored. In order to expand the search space beyond only explicitly linked phrases in the content of an article, I utilise the entire textual content of an article for mention of relevant articles.

To identify relative sentences, I first parse the textual content of all Wikipedia articles by decomposing them into their constituent sentences with markup removed. Links appearing in the sentence are expanded by following the link target to its final destination (resolving redirects), and the full article title appended to the end of the sentence. This is due to some links in Wikipedia text being presented as abbreviated or alternate versions of their full article titles. This is essentially the same as hyperlinking a word on a webpage to a link target rather than presenting the full path of the link itself. For example, in the marked-up sentence The [[bird]] [[Family...
Petroicidae includes roughly 45 species in about 15 genera, the link [[bird]] will be presented in text as “bird” and the link [[Family (biology)| family]] will be presented as “family”. Sentences with fewer than 4 words are removed from the set, as well as sentences that occur as part of lists (as these are usually lists of links).

Each of these sentences is placed into a single file using TREC notation: an SGML markup standard that contains the text of the document within <DOC> tags, and identified by a unique <DOCNO> tag. The unique document id for each document consisted of the title of the article from which the sentence was taken, and the sentence number within that article. For example, the 4th sentence of the Tarantula article was placed into the document file as:

```xml
<DOC>
  <DOCNO>Tarantula, Sentence 4</DOCNO>
  <#SENTOL:Genus#> Some genera of tarantulas hunt prey primarily in trees; others hunt on or near the ground.
</DOC>
```

Note that the outlink to the article Genus is represented using a SENTOL tag appended to the beginning of the sentence, and the markup denoting its original placement (originally linked-to by the word genera) has been removed from the sentence.

The September 9th, 2009 Wikipedia dump (described in Section 3.2) was parsed into over 150 million individual sentences.

The Information Retrieval system Zettair (Billerbeck et al. 2004) was used to index the sentences and search through the sentences using the desired relationship targets (the two articles that are to be related) as the search terms. The basic out-of-the box configuration of Zettair was used for this process: stop words are removed for indexing, and document similarity is computed by using Okapi term weighting (using the formulation provided by Billerbeck et al. (2004)).
5.1.3 Article Links

Links in the body text of Wikipedia articles provide direct reference to related articles. The inclusion of a link in an article provides an explicit statement of how the current article relates to the article at the target of the link. This link can be unidirectional, or the articles may share a strong enough connection that they both link directly to one another (Milne et al. 2006; Milne and Witten 2008a).

The method of textual search described in the previous section identifies direct article links between pairs of articles (as article link targets are included in the sentence representation used), however, the taxonomic structure created by these inter-article links is able to be further exploited to identify relationships between articles. This section describes two methods of utilising inter-article links to identify relationships between Wikipedia articles.

Common Third Article

The Wikipedia Link Vector Model (WLVM) described by Milne and Witten (2008a) determines article relatedness based on how often two articles link to common articles, and how often those articles are linked to by all other articles in Wikipedia. Using this principle of identifying the common article, rather than an explicit link from one article to another, this measure can be adapted to isolate shared qualities of the original two articles in terms of how they relate to a third article.

This relationship is expressed by explaining how the two articles related to the third article individually, and then defining these two relationships as a shared quality. The original two articles share a uni-directional relationship to this common third article.

For example, the articles Beetle and Mummy both contain links to The British
Figure 5.2: Relationship identification using a reference to a common third article.

Museum as both of these articles contain instances of their topics located at The British Museum (a beetle collection, and as part of the Ancient Egypt exhibit, respectively). In this case, the relationship that exists between Beetle and Mummy is one of location: both of these articles have instances in the same physical location.

Sibling Reference

As described in Section 2.3.3, the classification of articles into related categories is intended to describe an article using the most specific categories available. Whilst this makes articles that share a small element of relatedness very unlikely to have categories of the same specificity or place them in a similar location in the category tree, articles that are members of the same category possess a very high degree of relatedness. This highly specific categorisation of articles can be exploited to identify relationships to closely related sibling articles.

Identifying relationships from articles to another article’s sibling or parent is done by determining if any links in an article share a common category with the target article.
The relationship between Dendroclimatology and Giant Squid presented at the beginning of this chapter is an instance of this relationship type.

Only a single level of category ancestry was used to extract relationships using this method instead of the depth scaled method presented in Section 5.1.1, thus only identifying relationships with the closely related immediate sibling articles or articles of a target’s parent categories. This enabled the identification of relationships where the relationship was to the hypernym of an article. Relationships from an article to an article that was a member of a target’s parent category were less fine grained or intricate, and able to provide a clearer definition of how the initial article related to a general concept surrounding the second article, rather than just the second article in isolation.

This link is essentially a reference to a sibling article of the target article: an article that shares the same parent category as the target article. The relationship described at the introduction of this chapter is an instance of this kind of relationship. The Giant Squid doesn’t relate directly to the Dendroclimatology or Dendrochronology articles, but does relate to the Tree article. Dendroclimatology
is a member of the **DENDROLOGY** category which is a member of the **TREES** category (of which **Tree** is a member).

## 5.2 Relationship Extraction Task

A subset of 20 article pairs out of the 820 exhibit pairs described in Chapter 4 was chosen as the set of pairs for which relationships would be extracted. 16 of these pairs were chosen at random from the 820 pairs, and an additional 4 pairs were selected manually. These manually selected pairs were deemed to have a high likelihood of identifiable and extractable relationships, and that a person with casual knowledge of both of the articles would be able to easily relationship a relationship between them. The four manually selected article pairs were: **Sauropoda & Hadrosaurid**, **Human Skeleton & Human Anatomy**, **Gold Nugget & Medal**, and **Frieze & Urn**. Hadrosaurs and Sauropods are both dinosaur families, the human skeleton is part of the human anatomy, medals can be made from gold, and friezes and urns were both prominent in classical art and sculpture.

Each of the extraction methods was applied to this set of 20 pairs. Only a maximum of 10 reasons were accepted for each extraction method, and any which were shorter than 4 words or were part of a list were excluded. The method of Text Retrieval was the most prolific measure, and in only 2 cases did it fail to produce any reasons at all. Closer examination of some of these results indicated that it identified many articles that were simply large lists of articles or topics. These results were pruned from the final results as they did not describe relationships.

The final number of relationships extracted for each method were: 94 relationships identified by basic text search; 29 relationships identified from links in the source
article that linked to a sibling of the target article; 40 relationships that related both articles to a third article; and 5 relationships where the articles had common ancestor categories. The number of extracted relationships for individual article pairs is given in Table 5.1. The following sections describe the relationships extracted by each method.

### 5.2.1 Extraction Discussion

Relationships between articles were identified by performing a plain text search using the two article titles as the search terms. The top 10 ranked documents (sentences) returned that were greater than 4 words long, or part of a list of links, were selected as being the resulting relationships between the two articles. The restriction on short sentences and list elements was to eliminate the identification of relationships that were not fully explained, and sentence fragments that occur as part of an itemised list of links. An examination of the reasons extracted using this method, reveals that many of the reasons extracted simply repeat a single one of the search terms multiple times, resulting in a high document similarity score. The majority of the time, these extracted relationships simply refer to one of the pairs without mentioning the other at all. For example, for the article pair **Glyptodon** and **Phar Lap** the sentence *Phar Lap is now part of IntervalZero, formerly Ardence, which produces, among other products, the Phar Lap ETS real-time operating system, used for instance on LabVIEW real-time targets* does not relate to **Glyptodon** at all, and to **Phar Lap** by name only. The identification of textual terms to identify relationships between entities fails to take into account the ambiguity of terms involved (in this case, a reference to the Phar Lap software package, rather than the racehorse). This
<table>
<thead>
<tr>
<th>Article Pair</th>
<th>Extraction Method</th>
<th>RACO score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouthparts &amp; Sauropoda</td>
<td>Common Ancestor</td>
<td>0.02</td>
</tr>
<tr>
<td>Dinosaur &amp; Mummy</td>
<td>Common Ancestor</td>
<td>0.05</td>
</tr>
<tr>
<td>Ant colony &amp; Gondwana</td>
<td>Common Ancestor</td>
<td>0.04</td>
</tr>
<tr>
<td>Ant colony &amp; Human genome</td>
<td>Common Ancestor</td>
<td>0.02</td>
</tr>
<tr>
<td>CSIRAC &amp; Hummingbird</td>
<td>Common Ancestor</td>
<td>0.03</td>
</tr>
<tr>
<td>Diprotodon &amp; Mummy</td>
<td>Common Ancestor</td>
<td>0.02</td>
</tr>
<tr>
<td>Crystal &amp; Mouthparts</td>
<td>Common Ancestor</td>
<td>0.01</td>
</tr>
<tr>
<td>Beetle &amp; Mummy</td>
<td>Common Ancestor</td>
<td>0.03</td>
</tr>
<tr>
<td>Giant squid &amp; Sauropoda</td>
<td>Common Ancestor</td>
<td>0.06</td>
</tr>
<tr>
<td>Dendroclimatology &amp; Giant squid</td>
<td>Common Ancestor</td>
<td>0.01</td>
</tr>
<tr>
<td>Blunderbuss &amp; Sail</td>
<td>Common Ancestor</td>
<td>0.02</td>
</tr>
<tr>
<td>Bushfire &amp; Petroicidae</td>
<td>Common Ancestor</td>
<td>0.02</td>
</tr>
<tr>
<td>Petroicidae &amp; Sauropoda</td>
<td>Common Ancestor</td>
<td>0.04</td>
</tr>
<tr>
<td>Cyatheales &amp; Gorilla</td>
<td>Common Ancestor</td>
<td>0.06</td>
</tr>
<tr>
<td>Glyptodon &amp; Phar Lap</td>
<td>Common Ancestor</td>
<td>0.03</td>
</tr>
<tr>
<td>Coral &amp; Diprotodon</td>
<td>Common Ancestor</td>
<td>0.07</td>
</tr>
<tr>
<td>Hadrosaurid &amp; Sauropoda</td>
<td>Common Ancestor</td>
<td>0.15</td>
</tr>
<tr>
<td>Gold nugget &amp; Medal</td>
<td>Common Ancestor</td>
<td>0.03</td>
</tr>
<tr>
<td>Human anatomy &amp; Human skeleton</td>
<td>Common Ancestor</td>
<td>0.16</td>
</tr>
<tr>
<td>Frieze &amp; Urn</td>
<td>Common Ancestor</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>-</strong></td>
</tr>
</tbody>
</table>

Table 5.1: Article pairs used in testing, number of relationships extracted by respective methods and corresponding RACO scores. The exhibit titles correspond to the exhibit definitions in Appendix A.

ambiguity can be resolved by explicit reference to the desired article. The use of links between articles resolves this ambiguity.

The use of a depth scaled ancestor category overlap has multiple advantages over
a uniform use of category membership independent of article depth. For example, when considering only a single level of ancestry above the immediate parent categories of an article (i.e., using only parent and grandparent categories) only 5 Common Ancestry relationships were identified across the 20 article pairs. Four out of five of these extracted reasons were between the article pair Human Anatomy and Human Skeleton. Human Anatomy and Human Skeleton can be classified as members of the same entity (the human body), with many similar attributes. The Human Skeleton is a component of the Human Anatomy, thus making them very closely related entities and increasing their categorical overlap. The use of solely grandparent categories produces relationships for articles with a high semantic relatedness, but a depth scaled approach is able to identify relationships between a greater number of article pairs. For example, using grandparent categories there is no Common Ancestry relationship identified between the articles Hadrosaurid and Sauropoda, but using the depth scaled approach, the articles are identified as both being Dinosaurs, Mesozoic Animals and Prehistoric Reptiles. The shift to the use the scaled ancestry extraction was made after observing the poor levels of extraction for this set of exhibits. Using the scaled ancestry extraction method, 23 relationships were identified in the above set of article pairs. Unfortunately, this shift was made midway through the evaluation process, and the first 3 iterations of the evaluation process (described in Section 5.3) utilise datasets with Common Ancestry relationships determined by grandparent-level only category overlap.
5.3 Evaluation

The evaluation of the extraction methods presented in Section 5.1 requires an approach that examines the qualitative nature of the extracted relationships. Chapter 4 described the use of gold standard semantic relatedness scores to evaluate the performance of computational measures of semantic relatedness and similarity. That evaluation methodology is a straightforward one: compare a set of generated numbers with gold standard numbers and determine if there exists a correlation between the two sets. Examining the specific semantic relationships that have been identified between concepts requires a qualitative analysis of each annotation. When analysing the extracted relationships, the appropriateness of the relationship (whether or not it is a valid relationship between the two entities), and the quality of the relationship need to be taken into account. In a best case evaluation scenario, the evaluation of these extracted reasons requires a judgement of whether these reasons are the best possible reasons out of all potential reasons that could have been extracted. Knowledge of the quantity and types of relationships that exist for all article pairs would allow for the explicit evaluation of the extraction methods based on the relationships retrieved. Given this gold standard set of relationships, extraction measures would be able to be evaluated in terms of the reasons extracted as a proportion of the set of valid reasons for this pair over all of Wikipedia, and consequently, whether the methods were extracting invalid reasons, or the best possible reasons.

In these respects the evaluation of these methods parallel the evaluation of Information Retrieval systems: retrieving a set of documents (relationships) for a given query (article pair) over a large document collection (all relationships in Wikipedia).

Information Retrieval systems have a well defined methodology for testing a system’s performance: given a large collection of documents and a set of queries, there
exist a finite number of gold standard documents within the collection relevant to each query (van Rijsbergen 1979). Testing the performance of a new IR system is done by comparing the set of documents returned by the system against the set of gold standard documents for the query.

This approach is the standard one for evaluating IR systems, and it relies on heavily curated sets of documents. The exact numbers of relevant documents for each query, and in some cases the ordering of the documents’ relevance to the queries, must be known in advance (Baeza-Yates and Ribeiro-Neto 1999). The production of new evaluation sets is an expensive task.

To utilise an IR evaluation methodology, we require a document set for which each document has been evaluated in relation to each query, and the documents which are relevant to the queries identified as the desired set. In essence it is required that every sentence of Wikipedia be examined to determine if it posses a valid relationship between a given two articles.

For this relationship extraction task, the document collection consists of every sentence in Wikipedia. Once blocks consisting of solely markup are stripped, this collection still consists of over a hundred and fifty million separate documents (each sentence in Wikipedia is a document). Performing an IR style evaluation for this set would involve manually examining all 150-million+ sentences to determine the exact sentences that fulfil the queries.

This presents a situation in which extraction measures are able to identify potentially relevant relationships, but there exists no gold standard with which to compare their candidates (as the cost of generating a complete set of gold standard documents is prohibitive). Manual evaluation of the extracted reasons is a more feasible task:
human annotators are able to decide whether a relationship is valid, and the validity of each relationship is given by the agreement of the annotators.

I approach the task of manual evaluation in stages, using multiple annotation techniques and annotator groups. Primarily, this section presents a discussion of the necessary considerations taken into account in the construction of a methodology for the evaluation of relationships between diverse conceptual entities. The annotation format was refined over several iterations to focus on the property of relationship validity: the property that an extracted relationship possessed a relation that linked the two conceptual entities aligned to respective Wikipedia articles. The results of this annotation process and the analysis of the performance of the relationship extraction methods formulated in this chapter are described in Section 5.4.

5.3.1 Evaluation Design

The primary goal of this evaluation is to determine the validity of an extracted relationship. The relationship may possess other qualities that distinguish it from other extracted reasons (e.g., a very broad relationship, or a slightly convoluted relationship), but these are secondary to the validity of the relationship. I identify the proportion of valid relationships that an extraction method extracts versus the total number of reasons the method has extracted as the primary indicator of an extraction method’s performance (i.e., the ratio of valid relationships to invalid relationships).

In addition to the binary property of relationship validity (i.e., valid or invalid), the specificity and clarity of the relationship were also used as evaluation criteria. The measurement of clarity was presented as a 4-point scale, ranging from 0 (unintelligible) to 3 (complete clarity), while the classification of specificity was presented as a binary choice (broad or specific)
Figure 5.4: The second iteration of the annotation interface. It presents two article blurbs (comprised of the first two sentences of the Wikipedia article), and a series of potential relationships between the two articles.

The process of annotation was performed using a standard template for each article pair. Annotators were presented with a two sentence blurb of each article (comprised of the first two sentences of the respective Wikipedia articles), and the set of relationships extracted for that article pair. The annotator was asked to grade the relationship on the qualities of *validity*, *clarity* and *specificity*.

To test the efficacy of this annotation interface, a trial annotation task was performed using a subset of 6 article pairs from the 20 used in the relationship extraction experiment. The article pairs used in this testing were *Mouthparts & Sauropoda*, *Ant Colony & Gondwana*, *CSIRAC & Hummingbird*, *Beetle & Mummy*, *Dendroclimatology & Giant Squid* and *Glyptodon & Phar Lap*. Six annotators (computer science postgraduate students) took part in this annotation task, and
Table 5.2: Fleiss’ Kappa statistic for different partitions of the first round of annotation. There was only one reason extracted using the Common Ancestor method in the sample set of six pairs used in the pilot annotation.

<table>
<thead>
<tr>
<th>Method</th>
<th>validity</th>
<th>specificity</th>
<th>clarity</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Ancestor</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Article Text</td>
<td>0.37</td>
<td>0.32</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>Common Third Article</td>
<td>0.27</td>
<td>0.28</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>Sibling Reference</td>
<td>0.58</td>
<td>0.50</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>Combined</td>
<td>0.45</td>
<td>0.39</td>
<td>0.21</td>
<td>0.23</td>
</tr>
</tbody>
</table>

In order to determine the overlap in annotator evaluations of relationship validity, it is necessary to calculate the annotator agreement. To measure the annotator agreement, I used Fleiss’ Kappa statistic (see Section 2.7.2). Fleiss’ Kappa was used to analyse the results of the annotation task to determine the agreement of the annotators over multiple partitions of the data. The results of this analysis are presented in Table 5.2.

The inter-annotator agreement was low when using this set of annotation guidelines, this is largely due to the large number of classes which reduces the likelihood of agreement (Sim and Wright 2005). When using an ordinal scale to evaluate data, slight differences of opinion between annotators (e.g., one annotator rates a relationship as a 3 on the clarity scale, while another annotator rates the same relationship as
a 4 while agreeing with every other aspect of the first annotators judgement) results in complete disagreement as determined by Fleiss’ Kappa. When there are a large number of finely separated classes, the disagreement will be higher (Sim and Wright 2005).

However, on the key evaluation criteria of relationship validity, annotators demonstrated a level of agreement that indicated a Moderate level of agreement as described by Landis and Koch (1977). This annotator agreement is consistent with manual annotation tasks utilising a similar manual evaluation methodology (e.g., the evaluation of truthfulness of statements as described by Rubin (2007)). This agreement is encouraging in that it demonstrates the agreement of annotators on the primary target criteria of relationship validity. The overall validity of the relationship extraction methods is presented in Section 5.4.

Refinement of the Annotation Interface

Based on the feedback from the annotators and the kappa statistic, a new interface was created to reduce the large number of classes. The evaluation of relationship clarity was eliminated from the annotation interface and a set of labels, independent from the validity and specificity measures, was introduced as a secondary classification scheme. The labelling structure, nominated by the annotators, was included as a method of separating the relationships. These labels were Broad, Specific, Obvious, Obscure, Tenuous and Interesting. The layout of this annotation interface is presented in Figure 5.4.

In addition to this restructuring, a set of deliberately invalid relationships was included to ensure the reliability of annotation. Distractor relationships for an article were constructed by selecting first sentence from the first article linked to by the can-
didate article. For example, The first link that appears in the Mummy article is to the Vatican Museums article, and thus the distractor relationship for Mummy is the first sentence from the Vatican Museums article (specifically, the sentence, The Vatican Museums, in Viale Vaticano in Rome, inside the Vatican City, are among the greatest museums in the world, since they display works from the immense collection built up by the Roman Catholic Church throughout the centuries). Using this approach, a given article pair had two distractor relationships included in the list of extracted relationships (one constructed from each of the articles in the pair).

The revised annotation interface was used to perform the relationship annotation task using the same group of annotators and the same data sub-set. The inter-annotator agreement on the question of relationship validity increased to $\kappa = 0.56$, with 100% agreement over the distractor relationships. However, the agreement of annotators over the labels was very small, ranging between $\kappa = -0.03$ and $\kappa = 0.23$ depending on the label. This increase in validity was not solely due to the increased agreement over distractor reasons: with distractor relationships removed from the dataset, the annotators maintained an agreement of $\kappa = 0.53$.

Further feedback provided by the annotators identified that the labelling structure implemented in the updated interface did not suit the content of relationships, and while labels applied in some circumstances, they were not consistently applicable to provide valid separation between relationships. Due to the separation provided by variations in the proportion of validity of the methods of relationship extraction and the feedback provided by annotators, secondary classification criteria were removed from the annotation interface.
5.3.2 Mechanical Turk

In order to evaluate the complete set of relationships extracted for the 20 article pairs used in the extraction experiment, the annotation criteria decided upon based on the previous two iterations was disseminated to a larger group of annotators. For this task, the Amazon Mechanical Turk (see Section 2.7.3, Section 3.3.3) was used to source annotators and provide an infrastructure for recording relationship evaluations.

The feedback given by annotators based on their experience with the pilot annotation tasks described above was used to construct a Human Intelligence Task (HIT) for use over the Amazon Mechanical Turk. Specifically, feedback on question phrasing and criteria used to evaluate the data-points was used to clarify the instructions given to potential annotators. Relationship validity was the only criteria used to evaluate relationship quality in this interface. Mechanical Turk requires that all HITs consist of the same number of annotation points. Each HIT comprised two article blurbs (as described above) and three relationships for evaluation. The relationship partitioning was conducted in such a way that if a Turker chose to annotate every set of three relationships included in the data set, the Turker would only encounter each extracted relationship once, however distractor relationships were included multiple times in order to meet the requirement that there be three relationships per HIT. A distractor relationship was included in every HIT to ensure that Turkers weren’t answering ‘yes’ to the validity of every relationship (and thus introducing erroneous evaluations). The final design of this HIT can be seen in Figure 5.5.

The time taken to complete the evaluation of the relationship data set was a little over 6 hours. This Task consisted of 92 individual HITs containing three relationships each, with the requirement that each HIT be completed 5 times, resulting in 1380 individual relationship judgements.
One Turker failed to perform the annotation correctly and annotated every relationship in the data set, including distractors, as valid. This user was also the user with the most responses, evaluating every relationship in the experiment. As this user did not properly complete the task as it was designed, their HITs had to be removed from the data. This removed a fifth of the annotated data, and hence eliminating any deciding vote in cases where the other four annotators could not reach consensus (two annotators evaluating a relationship as valid, and two annotators evaluating the relationship as invalid).

Fleiss’ Kappa was calculated for the results of the Mechanical Turk evaluation after the removal of the erroneous annotations. The resulting annotator agreement was lower than that of the annotators in the pilot annotation. The overall Kappa score for the Mechanical Turk evaluation data set was $\kappa = 0.08$. The reason for
this inconsistency was unclear as there was no feedback left by annotators, however, analysis of the analysis of the validity judgements indicated that in a majority of cases, the majority class for a relationship’s validity (based on the 4 annotation points remaining) was undecided, receiving evaluations of 2 valid and 2 of invalid. The complete results of the validity evaluation for all annotation tasks can be seen in Table 5.6, presented in Section 5.4. To reduce the likelihood of undecided data points, a second iteration of Mechanical Turk annotation was undertaken with a smaller set of relationships, but with the proviso that each relationship be annotated by twenty separate annotators.

Separately, there existed the possibility that the relationships in the 20 article pairs were ambiguous or unclear. To remove this possibility, a new set of article pairs was selected at random and a new set of relationships extracted. This annotation task utilised a newly generated set of relationships, comprised of 5 randomly chosen relationships from 5 newly chosen article pairs. Each set of 5 relationships was placed into a separate HIT.

In contrast to the amount of time taken to complete the previous Mechanical Turk task, this task took a week to be fully annotated. This was potentially due to the smaller amount of money offered per HIT: $0.05 per 5 relationships annotated as opposed to $0.05 per 3 relationships annotated as with the previous task. This hypothesis is consistent with the ratio of work to payoff described by Snow et al. (2008). In this evaluation task, 10 HITs were rejected due to annotators failing to specify the validity of a relationship, or for consistently annotating all relationships as valid.

The resulting annotator agreement was little improved from the previous Mechanical Turk task, achieving a Fleiss’ Kappa score of $\kappa = 0.18$. The occurrence of
relationships with undecided validity was greatly reduced, but many were borderline cases, with 2 of the 25 relationships still being undecided (including a distractor relationship). These borderline cases (e.g., 8 annotators declaring a relationship as valid, and 12 as invalid) caused a situation in which the validity of a relationship could be decided (due to the unlikeliness of a 10-10 split of annotators), but this judgement was not reliably accurate.

### 5.3.3 Museum Staff Annotation

There are multiple potential reasons for the disagreement of the Turkers: the annotators were not familiar with the content of the articles as the blurbs were not enough to fully contextualise the conceptual content of both articles; individual annotators may have interpreted the relationships in different ways based on life experience with the concepts involved; or the relationships were not clear enough. To remove (or at the very least, limit) the variable factor of life experience, the decision was made to identify a group of annotators with a similar level of familiarity with the concepts on which the article pairs were based, and knowledge of the potential interaction between them. For these reasons, the staff at Melbourne Museum were selected as group of

---

**Table 5.3: Fleiss’ Kappa statistic for each extraction method, separated by annotation source.**

<table>
<thead>
<tr>
<th>Extraction Method</th>
<th>Annotation Source</th>
<th>Pilot 1</th>
<th>Pilot 2</th>
<th>M-Turk 1</th>
<th>M-Turk 2</th>
<th>Museum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td></td>
<td>0.27</td>
<td>0.34</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.47</td>
</tr>
<tr>
<td>Third Article</td>
<td></td>
<td>0.27</td>
<td>0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0.25</td>
</tr>
<tr>
<td>Ancestor</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Sibling</td>
<td></td>
<td>0.58</td>
<td>0.56</td>
<td>0.03</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td>Distractor</td>
<td></td>
<td>–</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.34</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>0.45</td>
<td>0.56</td>
<td>0.08</td>
<td>0.18</td>
<td>0.33</td>
</tr>
</tbody>
</table>
annotators familiar with the collection of concepts considered (as they were based
on museum exhibits from the collection at Melbourne Museum) and the connections
between them.

The set of article pairs used in the second Mechanical Turk task were used for
this evaluation task, but instead of restricting the number of relationships for each
article pair, the full list of extracted relationships was used for each pair. To remove
the factor of evaluation fatigue, a subset of 5 relationships for each article pair was
presented to the annotator. The selection of these relationships was randomised
to ensure equal distribution of annotation across all relationships. The response to
this annotation task was smaller than the previous iterations of the task, and each
relationship was only annotated twice.

The museum annotators achieved a Kappa score of $\kappa = 0.33$, greater agreement
than that of the Turkers but lower than that of the pilot annotations. Table 5.3
shows the Kappa statistic for each iteration of the annotation task, partitioned into
the Kappa for individual methods. Note that due to the randomised selection, re-
lationships extracted by the common ancestry method were rarely selected. The
sparsity of immediate parent category overlap prompted the inclusion of the scaling
method described in Equation 5.1, however this alteration was not performed in the
first three iterations of the annotation interface: the two pilot interfaces, and the first
Mechanical Turk interface. In all cases where the common ancestry method was eval-
uated, the agreement was unanimous, resulting in Kappa statistics of $\kappa = 1.0$. This
includes the Mechanical Turk evaluation over the complete data set of relationships
which contained 4 common ancestry relationships. Further evaluation of the common
ancestry extraction method for larger collections of relationships must be performed
before any definitive statements can be made about its performance, however it was
the extraction method that achieved the highest degree of agreement between annotators across all iterations of the evaluation task, achieving unanimity of evaluation in all cases.

5.4 Comparison of Method Performance

Ultimately the annotation tasks described in the previous section were unable to reliably evaluate the extracted relationships. This was largely due to the lack of annotator agreement for larger collections of relationships. However, the annotation process was not entirely without merit. The five separate annotation tasks all utilised the criteria of relationship validity when evaluating the performance of extraction methods. A cross task analysis provides useful insight into the consistency of relationship extraction methods when evaluated by differing groups of annotators, and over different collections of relationships. There exists a great disparity between the quantity of relationships extracted by the more successful methods (e.g., Common Third Article) and the more prolific methods (Article Text). This can be seen in Table 5.1: Article Text is able to identify a large number of relationships, but the majority of these relationships were not deemed valid by annotators. The total relationships extracted by each method and the respective number of relationships termed valid, invalid and undecided (as determined by majority class evaluation) is shown in Table 5.4. Figure 5.6 contains a graphical representation of the breakdown of the percentage of relationships classed as valid, invalid and undecided based on the majority class of annotated relationships extracted by the respective methods.

Although there exists variability for the classification of relationships for a given method depending on the annotation source, the performance of multiple methods
Table 5.4: Evaluation counts for relations extracted by each extraction method, and as a total, separated by annotation group. Each entry is a comma separated set of validity, invalidity, undecided. In the percentage column and row, these three values are presented as an average percentage of the extracted relationships’ validity.

demonstrates patterns of validity. Most noticeable is the complete validity displayed by the common ancestor method (shown in Figure 5.6(d)): relationships extracted by this method achieved 100% validity in all annotation tasks. As described in the previous section this is largely due to the sparsity of relationships extracted by this measure selected for evaluation, however, the validity of these relationships was determined unanimously in all cases for multiple annotators, and multiple relationships. Ultimately this cannot be taken as a definitive statement of the effectiveness of this method in identifying valid relationships between articles due to the small size of the data set used.

Distractor relationships were intended as a normalising tool to identify which annotators were paying close enough attention to identify relationships that were designed to be invalid. When distractor relationships were included (shown in Figure 5.6(e)), annotators were able to identify this invalidity a majority of the time (50%-100%) with the remaining classifications skewing towards undecided. This is a
positive indication of the engagement of annotators in the task of evaluation. In only a few cases (11% in the first Mechanical Turk annotation and 20% in the Museum Staff annotation) were distractor reasons determined to be valid. I acknowledge that the method used to select distractor relationships could potentially identify a valid reason between an article pair, particularly if the articles are closely related to begin with, and this is a potential explanation for the identification of valid distractor reasons. For example, the distractor relationship, *Insects (Class Insecta) are arthropods, having a hard exoskeleton, a three-part body (head, thorax, and abdomen), three pairs of jointed legs, compound eyes, and two antennae*, between *Beetle* and *Trilobite* was determined to be a valid relationship by museum staff.

The remaining methods of relationship extraction exhibit varying degrees of validity. The method of extracting relationships based on textual content (shown in Figure 5.6(c)) obtained the lowest average validity percentage of the three, achieving less than 35% validity for any given annotation group. The relationship between *Glyptodon* and *Phar Lap* described in Section 5.2.1 is an example of an invalid relationship extracted using this method. The reasons for the invalidity in this case is the confusion of the software package *Phar Lap* with the racehorse by the same name. The inability of this method to disambiguate the entities in the relationship and thus identifying them as relevant relationships is a weak point in this relationship, demonstrating the advantage in utilising the existing structure of Wikipedia to disambiguate entities.

The sibling reference method of relationship extraction also suffers from many extracted relationships being evaluated as invalid. A potential explanation for this is the separation present in some Wikipedia categories. The *Giant Squid* and *Dendroclimatology* is an example of when this succeeds: *Tree* is closely related enough to
Dendroclimatology that the relationship is relevant and valid, but among broader categories this connection may not be as evident. For example, the category Living People subsumes in excess of half a million articles, causing the identification of a relationship between Architecture and Michael Bublé using the British architect couple Brenda and Robert Vale as the sibling article.

In contrast, relationships extracted using the common third article extraction method achieve a high validity percentage across multiple groups of annotators. A potential explanation for this improvement over the sibling reference method of relationship extraction is that relationships extracted using this method demonstrate both articles’ connection to the third article, ensuring both articles possess a relationship in their own right before joining them together to demonstrate the common link. The use of explicit links between articles (rather than relying on category membership to identify links) reduces the burden of interpretation on the reader as it doesn’t require them to make the mental connection between a link and a sibling article.

A qualitative analysis of the extracted relationships indicates that the extraction methods identify different aspects of relatedness when describing the connection between articles. Common Ancestry is the most predictable of these, as it most often describes an organisational similarity between the two articles (e.g., Beetle and Trilobite are both Arthropods, or Crystal and Geode are both Crystalline solids), but the remaining methods also exhibit this behaviour. Over all pairs tested and relationships extracted, there existed only one relationship that was identified by more than one extraction method: the relationship Some scholars suggest that the Egyptians’ practice of making mummies was inspired by the brooding process of the beetle, between Beetle and Mummy. The Article Text method also identifies this relationship (using a different wording) in another Wikipedia article. Unfortunately, as this was the only
such instance of extraction overlap, no conclusions can be made about the correlation between extraction overlap and relationship validity.

The above **Beetle–Mummy** relationship appears in the **Mummy** article, making it identifiable by the Sibling extraction method, but also contains explicit reference to mummies and beetles, making it appear in a text search. In addition to succinct statements referencing both article titles, the method of identifying relationships by article text introduced many invalid and noisy candidate relationships. In the majority of cases this was due to there being no sentence in Wikipedia that contained both article titles, meaning the method only identified sentences as candidates based on a large number of occurrences of a single article title in a sentence (e.g., the Phar Lap software package sentence described in Section 5.2.1).

Using text of the aligned articles as the source of any extracted relations, in the Sibling and Common Third Article methods, removed the sense ambiguity that contributed to the high proportion of invalid relationships extracted by the Article Text method. Despite the use of one of the original articles for which relationships are being extracted as the source of text for extraction, the Sibling method only succeeds in extracting valid relationships 28% of the time. Analysis of the relationships evaluated as invalid reveals that sentences selected as relationships often do not refer to the context in which they appear combined with the lack of a direct reference to the target article. For example, in the extracted relationship **In March 2006, the body of the Greek Orthodox Monk Vissarion Korkoliacos was found intact in his tomb, after fifteen years in grave**, between **Mummy** and **Dinosaur**, the link between **Dinosaur** and its sibling, **Monk**, is unclear. In this case, the link between these two articles is that they were both members of the **Greek Loanwords** category. Unless the
reader were attempting to understand the concepts in the sentence in terms of their linguistic roots, this relationship would not be apparent.

The Common Third Article method avoids this problem by only identifying sentences that link explicitly to a target, rather than relying on an inferred connection. These links are not between the desired targets, but their individual link to the common third article is explicitly given. For example, in the relationship between Dinosaur and Mummy:

Dinosaur and Mummy are both related to Mongolia:
In Dinosaur: The Mongolian oviraptorid “Citipati” was discovered in a chicken-like brooding position in 1993, which may mean it was covered with an insulating layer of feathers that kept the eggs warm.
And in Mummy: Natalia Polosmak discovered the Siberian Ice Maiden in a sacred area known as the Pastures of Heaven, on the Pontic-Caspian steppe in the Altay Mountains near the Mongolian border,

the relationship of each article to Mongolia is stated, reducing the ambiguity present in relationships extracted based on article siblings.

Aside from Common Ancestry, which extracted too few relationships to be evaluated reliably, relationships extracted through the identification of a Common Third Article achieve the highest average validity when evaluated. However, the multiple extraction methods highlight different aspects of article content in their relationships (as is evidenced by the small amount of overlap in relationships extracted by different methods), and each contribute to the overall understanding of how entities represented by the articles interact with one another.

Ultimately, the basis for the relationships between articles identified by the methods in this chapter is myriad. Examples provided in this chapter have included relations such as age measurement (Dendroclimatology & Giant Squid), cultural practice based on natural phenomena (Beetle & Mummy), same biological
classification (Beetle & Trilobite) and similar digestive method (Hadrosaurid & Sauropoda). These relationships are difficult to classify in terms of existing rule sets, and exemplify the intricate relation types that exist between contextually diverse entities. While additional processing is required to extract the precise relation in each of these relationships, these methods provide a base for the identification of semantic relations between contextually diverse entities without relying on pre-defined relation patterns.

5.5 Summary

The identification of relationships between entities can aid in educational tools and recommender systems in the Cultural Heritage domain and elsewhere. However the diverse nature of entities in the Cultural Heritage domain makes the use of common templates to represent the contextually diversity inadequate. Existing studies utilise standardised conceptual frames to identify exhibit relations using overlap in attribute values (e.g., O’Donnell et al. (2001), Aroyo et al. (2007)). In an effort to identify a broader range of inter-entity relationships I utilise a more contextually diverse representation of entities: Wikipedia articles. I extend the basic principle of semantic relatedness between Wikipedia articles, as described done by RACO, to take advantage of the taxonomic structure of Wikipedia for the purpose of identifying the specific reasons that relate articles.

Wikipedia articles utilise a semi-structured representation of content, and do not utilise a standard structure for representing relationships between entities. Relationships to other concepts are described in the body text of the article but do not utilise a common representation. In order to overcome this limitation, I approach the prob-
lem of relationship identification between entities by utilising the existing taxonomic structure of Wikipedia to identify inter-article links. I demonstrate four methods of relationship extraction utilising existing Wikipedia components: mutual term occurrence in text (text search), the subsumption of article by ancestor categories (ancestor overlap), a link between one article and a sibling of the second article (sibling reference), and the mutual reference of both articles to a common third article (common third article).

The extraction methods were tested over a subset of the exhibit pair dataset described in Chapter 4. The evaluation of the performance of the individual extraction methods took the form of a manual annotation exercise. After two iterations of a pilot annotation interface the choice was made to use relationship validity as the primary criteria for evaluation. The evaluation exercise was performed 5 times with differing groups of annotators and over different sets of extracted relationships. The level of agreement between annotators varied greatly depending on the group of annotators used, and the relationships used in the evaluation process. Despite the uncertainty of annotation agreement among later iterations of annotation, patterns regarding the validity of relationships extracted by different methods are evident. Relationships extracted by common ancestor and third article reference methods have a higher likelihood of validity than those extracted using sibling references or article text search. A definitive statement about the ability of these methods to identify valid relationships between entities with diversely structured content is difficult without further, more rigorous, annotation of extracted relationships. However, the methods presented in this chapter demonstrate that the structure of the semantic network containing entities with diverse conceptual frames is able to be used as the foundation for contextually diverse relationships between these entities.
The calculation of the Semantic Relatedness between contextually diverse entities, and the identification of relationships between them has been demonstrated over the past two chapters in the context of the Cultural Heritage domain. However, the methods presented in this thesis are intended to be applicable to alternate domains. In the following chapter, I present a discussion of how the methods described in this chapter are directly applicable to the domains of Film and Tourism, and how alternate semantic networks to Wikipedia are able to be utilised to calculate the RACO score of entity pairs, and to identify relationships between these entities.
(a) Percentage of Common Third Article relationships judged as valid.

(b) Percentage of Sibling Reference relationships judged as valid.

(c) Percentage of Article Text relationships judged as valid.

(d) Percentage of Common Ancestor relationships judged as valid.

(e) Percentage of Distractor relationships judged as valid.

Figure 5.6: Validity percentage of relationship extraction methods determined by majority class of manually annotation of relationships, separated by annotation source. Graphs with missing columns ((d) and (e)) did not include relationships of that type during that round of annotation.
Chapter 6

Domain Diversity and Resource Portability

This thesis has described a measure for computing the semantic relatedness of conceptually diverse entities and a set of methods utilising taxonomic links between entities for the purpose of identifying semantic relations between entities. These methods have been constructed using Wikipedia as a digital representation of conceptual entities and as a representation of their conceptual organisation. Chapter 4 and Chapter 5 demonstrate the usefulness of this broad scale, collaboratively-edited knowledge base and its use as a basis for identifying semantic connections between entities.

In addition to Wikipedia there exist alternate semantic networks, both cross-domain and single domain, that describe large collections of entities and their conceptual organisation. These alternate semantic taxonomies possess differing organisational structures to Wikipedia and different entity representations. The difference in representation and structure of these semantic networks to that of Wikipedia is
examined in Section 6.1. This discussion includes an examination of the viability in using these semantic networks as a replacement for Wikipedia in identifying semantic relations and semantic relatedness between large collections of contextually diverse entities.

The Cultural Heritage domain has been used throughout this thesis for developing these methods, however the Cultural Heritage domain is not the only collection of contextually diverse entities that would benefit from the methods proposed in this thesis. Section 2.6 described the use of relatedness scores and semantic relations to User Modelling tasks, and briefly examined the benefit of their application to music and online shopping recommender systems. In this chapter I present an analysis of two alternate domains of information rich entities that can benefit from the methods proposed in this thesis: Film and Tourism. I present a discussion of the attributes that these domains possess to make them desirable candidates for semantic relatedness calculations and semantic relation identification. These qualities include the benefit of applying semantic relatedness and extracted relationships to the domain, resources for representing the content of entities, and semantic networks describing the organisation of these entities. Section 6.2 describes these two domains in detail, available domain specific resources, and the applications of semantic relatedness calculation and relation extraction to these domains.

### 6.1 Alternate Semantic Resources

The semantic resources used as the basis of the experimentation in this thesis are comprised of collaborative encyclopedias (Wikipedia), document collections (e.g., the Melbourne Museum website documents) and manually annotated data sets. In
this section I present a discussion of the alternate semantic resources that can be used to represent large collections of contextually diverse entities. This includes domain-specific wikis, digital encyclopedias, the Cyc knowledge base, and document augmentation through wikification.

6.1.1 Domain-Specific Wikis

There are few domains that Wikipedia does not contain (e.g., Milne et al. (2006) demonstrates a high degree of consistency between Wikipedia and a specialised agriculture semantic network, Agrovoc), but as an encyclopedic semantic network its stated goal is to provide encyclopedic definitions of entities. Section 6.2.2 describes Wikitravel, a travel based wiki that describes entities in terms of sights to see, accommodation and transport. The use of a domain specific wiki reduces the amount of contextual diversity for a given entity; restricting it to contextual diversity with regard to other domain entities.

Domain-specific wikis create a communal space for domain enthusiasts to create and edit content. This approach has advantages and drawbacks: ultimately, there is a smaller group of editors familiar with the content being documented, however, these editors are more familiar with the domain and are able to describe entities in terms of other domain content rather than generalised content and descriptions (e.g., cross-domain encyclopedic descriptions as in Wikipedia) (White et al. 2008).

Depending on the description style used to construct the semantic network, the calculation of semantic relatedness and the extraction of semantic relations will be performed with regard to the domain. This property will have a greater effect on the calculation of semantic relatedness as the separation of two entities within a domain is greater than when the same two entities are compared using a cross-domain semantic
network. For example, the fictional world of Tamriel used in the series of *Elder Scrolls* computer games contains a dedicated domain wiki located at http://www.uesp.net. In contrast, Wikipedia contains only a few articles on The Elder Scrolls series. When representing a location in the fictional world, Wikipedia (depending on the version of Wikipedia used, it may even be a single article) describes each game location using reference to the games in the series in which it appears. When comparing two locations using Wikipedia, RACO scores may be high due to large numbers of links to the same articles. This same comparison may differ greatly when using the articles from the Elder Scrolls wiki representing the same locations, due to a greater amount of content associated with each article. This is an issue of granularity of description: domain specific wikis place a different level of importance on more fine-grained entities relevant to the domain. An entity that may not satisfy Wikipedia’s notability criteria (or only warrant a short mention in an associated article) may have a large article devoted to it on a domain-specific wiki. Domain-specific wikis contain fewer articles due to the reduced number of entities as compared to a broad coverage wiki (e.g., Wikipedia), but as each entity is regarded at a closer level of granularity it will contain more associated information and hence more links to in-domain entities. This increased linking, greater description length, and reduced number of entities are all factors that have potential to effect the performance of RACO when used over these alternate resources, but experimentation is required to determine the precise nature of this effect.

Domain-specific wikis provide the advantage of more comprehensive description of domain entities compared to the representation used in cross-domain wikis, however, they lack desirable elements of Wikipedia such as a deep and well populated category hierarchy, instead being restricted by the sparsity of content and choosing
to forgo a hierarchical system (e.g., the flat classification scheme used in Wikitravel) or implementing their own minimalist categorisation scheme (e.g., based on the editing progress of articles). There do exist some domain-specific wikis that utilise a hierarchical organisation, similar to Wikipedia, eg the *Elder Scrolls* wiki described above. Whether this restricted size of category hierarchy impacts on the effectiveness of RACO or relationship extraction methods requires additional experimentation.

The contextual diversity of domain-specific wikis is less than that of Wikipedia as they lack reference to cross-domain entities, but this may be advantageous as relations to out-of-domain entities can result in irrelevant relations or when broad-scale relations are not desired. Domain-specific wikis are able to use the methods constructed in this thesis due to the same manner of construction as Wikipedia, however, calculations of semantic relatedness and relations extracted will be relative to the domain of representation. The domain-specific wiki, Wikitravel, is further analysed with regard to the applicability of methods developed in this thesis, in Section 6.2.2. Ultimately, further experimentation is required to determine the exact effect of methods in this thesis on domain-specific wikis, however this is outside the scope of this thesis.

### 6.1.2 Digital Encyclopedias

Wikipedia is an instance of a digital encyclopedia that utilises the wiki paradigm in its construction: collaborative editorship and peer review. Other digital encyclopedias are constructed in a more traditional fashion: professional editors create articles and the links between them. This approach yields smaller amounts of content in articles, but also yields article of a more consistent quality and structure (Giles 2005). Online digital encyclopedias such as Encarta and Encyclopedia Britannica consist of a similar
structure to that of Wikipedia as described in Section 2.3.3: entities have contextually diverse content, and are interlinked in the body of the article’s content.

There have been multiple analyses of the reliability of Wikipedia versus that of professionally constructed encyclopedias, e.g., Giles (2005), Rector (2007), Rosenzweig (2006). These investigations found that articles within professionally constructed encyclopedias are more reliable than their corresponding Wikipedia articles. Inconsistencies such as unsourced statements and minor factual errors were prevalent in Wikipedia articles when compared to their counterparts in other digital encyclopedias (Giles 2005; Rector 2007), but the amount of information present in Wikipedia articles was found to be greater than that of professionally constructed digital encyclopedias (Rosenzweig 2006). Increased reliability comes at the trade off of coverage: Wikipedia is the largest online encyclopedia and is able to provide a diversity of content many times greater than professionally constructed encyclopedias.

In addition to the factual accuracy of professionally constructed digital encyclopedias, the organisational structure of the encyclopedias also differ from that of Wikipedia. Digital encyclopedias contain inter-article links in the body of the text, pointing to related articles within the same encyclopedia, just as Wikipedia does, however, Wikipedia has the added benefit of the category hierarchy that has evolved from collaborative editorship, and is required for the use of the methods created in this thesis. Without this hierarchical network, methods of relation extraction that rely on category membership are no longer functional. This is not to say that there do not exist digital encyclopedias that do not use a hierarchical taxonomy similar to Wikipedia, merely that the larger, well researched digital encyclopedias of Encarta, Britannica Online do not utilise a hierarchical system of categorisation to organise entities. Some smaller, more rigorously reviewed encyclopedias incorporate a hierar-
chical classification scheme, however this comes at the cost of increased data sparsity compared to resources with open editorships. For example, the professionally edited, encyclopedic wiki, Citizendium,\(^1\) utilises the same hierarchical classification scheme to organise peer reviewed encyclopedic articles, but contains only 156 expert reviewed articles, with an additional 15,000 in various stages of review, far below the millions of articles offered in the English version of Wikipedia.

In using the larger digital encyclopedias so as to maintain a maximised level of coverage, the identification of a relation between entities by identifying a reference to a common third article and basic text search are still valid, but identifying relations to sibling articles or common ancestor categories are no longer possible due to the lack of categorisation in these encyclopedias. Smaller encyclopedias that utilise a hierarchical categorisation scheme are able to extract relationships using all of the methods formulated in Chapter 5. The Common Third Article method identified valid relationships with greatest consistency (aside from the Common Ancestry method for which there was not enough data to draw reliable conclusions), and the use of professionally inserted article links is likely to maintain this quality. Conversely, the Article Text method was the least reliable in identifying valid relationships over Wikipedia. The combined effect of the professionally edited document set and the reduced size of the content available in digital encyclopedias will impact on the proportion of valid relationships extracted by these methods, but further experimentation is required to determine the precise effect on the overall performance of these extraction methods.

---

\(^1\)English version of Citizendium, a free, peer reviewed, wiki-based encyclopedia: [http://en.citizendium.org](http://en.citizendium.org)
6.1.3 Cyc

The methods in this thesis have been developed to utilise semi-structured representations of entities, relying on large quantities of text and in-text links in addition to a taxonomic organisation of the entities. However, highly structured semantic networks still have the facility to represent the content of contextually diverse collections of entities (recall the Cultural Heritage semantic networks described in Section 2.3.4). These networks require alternate approaches to identifying unforeseen semantic relations between entities, and calculating the semantic relatedness of these entities. In this section I present a highly structured semantic network that represents a cross-domain collection of entities: Cyc.

Cyc is a computational knowledge base, the purpose of which is to become a machine readable representation of all human knowledge (Lenat et al. 1990). The structure of Cyc is hierarchical, and new entities are defined entirely by logical combinations of existing, more general entities (Lenat et al. 1985). Entities in Cyc are called individuals, and are defined in terms of pre-existing, more general individuals using a system of logical predicates. For example, to insert the logical notion of a brother into Cyc, a new individual is created that is a specialisation of the sibling individual, and possesses an argument frame that takes an individual that is defined as being a specialisation of a male, that has the same parent individual as the brother individual. The use of logical predicates to specify individual relationships and attributes allows Cyc to be queried by searching for individuals that satisfy certain truth conditions, e.g., searching for all individuals within Cyc that have brothers.

This logical structure has been used to identify word sense, and sentence structure: by testing the validity of applying a given property to an individual, the validity or correct interpretation of a larger statement can be identified. An example given
by Lenat et al. (1985) is the following two sentences: *Fred saw the plane flying over Zurich*, and *Fred saw the mountains flying over Zurich*. The individuals in Cyc can be used to identify the specific properties of the objects in these sentences (specifically that planes have a flying property and mountains do not), based on which of the interpretations is logically possible it is possible to eliminate interpretations of these sentences that violate these properties (e.g., flying mountains).

The *individual* entity structure of Cyc allows an author to potentially create any relationship between other individuals. Sarjant et al. (2009) used this property to augment Cyc using relationships extracted from Wikipedia. Using an alignment process described by Medelyan and Legg (2008), Wikipedia articles were aligned to Cyc individuals and features such as the article infoboxes, category membership and syntactic patterns in the articles’ text were used to identify relationships between the aligned individuals. This approach allowed individuals to be automatically added to Cyc, rather than undergoing the time consuming task of manually adding a large collection of new individuals.

The comprehensive nature of Cyc allows it to be used for many tasks of computational intelligence such as question answering (Chu-Carol et al. 2003; Curtis et al. 2005) and word sense disambiguation (Curtis et al. 2006). However, the information within Cyc is rigidly defined and while containing many specifically identified Named Entities (such as political leaders, organisations and historical events), lacks the descriptive aspect that desirable for identifying contextual links between entities used by digital encyclopedias (Weyer and Borning 1985). This is exemplified by several studies using encyclopedic resources to augment individuals in Cyc (e.g., Medelyan and Legg (2008), Sarjant et al. (2009), Zirn et al. (2008)). While Cyc does contain the facility for textual descriptions of individuals, these descriptions rarely exceed a
sentence in length, and are intended to be a brief statement about describing the entity (Lenat 1995). Large amounts of unstructured textual content are not in keeping with Cyc’s purpose as a logical, formal representation of knowledge.

Cyc has the ability to describe an individual’s contextual diversity using semantic relations rather than textual descriptions. The limitation of this approach is that the complexity of representation needed to describe a brief discussion of how, for example, a giant squid’s age can be measured similarly to that of a tree, would require many individuals fitting into multiple argument frames rather than a single sentence textual explanation. Cyc does have the potential to utilise the methods proposed in this thesis, however, issues arise from the coverage of Cyc: using the alignment process described by Medelyan and Legg (2008), Cyc contains the equivalent of 250,000 Wikipedia articles (based on the 2009 release of Cyc 2.0 (Matuszek et al. 2006)). Furthermore, the complexity of description limits its ability to represent large amounts of contextual diversity without extensive adaptation to take into account the logical representation used by Cyc.

Cyc’s strict use of the is-a relation to define each entity results in a viable taxonomic structure for use with RACO. Entities are defined using multiple inheritance, analogous to category membership present in Wikipedia. Beyond this inheritance structure, Cyc possesses little facility for identifying related entities, thus complicating the adaptation of RACO. As a substitute to the use of entity-entity links, the property of subsumption may be used. Recall that article subsumption was not present in Wikipedia as only categories possessed children, but as entities in Cyc also constitute the hierarchy of the network, an additional set of relations is now available. Theoretically, child articles could be identified in Wikipedia through use of the main article link in the corresponding category (e.g., the Earth is the main article
for the EARTH category), but main article links are not universal across Wikipedia categories, meaning that child articles cannot be identified in all cases.

The calculation of a RACO-like measure of semantic relatedness for use over Cyc could be formulated by first identifying all child entities of the two entities being compared, and then computing the overlap of the parent entities of the set of child entities. This calculation would need to be normalised to account for entities with differing numbers of child nodes, as RACO does. The effectiveness of this measure of semantic relatedness over such a collection of rigidly defined entities (representing contextually diverse entities) may differ greatly from its equivalent calculation over Wikipedia, but experimentation is required to determine the viability of this shift in representational format. The extension of contextual overlap to Cyc using this adapted methodology is a desirable direction of future research, but is not explored in this thesis.

The ability of Cyc to identify diverse relations between entities is hindered by its rigidity and absoluteness of relation definition: relations between entities are clearly stated using the is-argument property of entities. The example of the Giant Squid and Dendroclimatology relation provided above requires that the dendroclimatology entity refer to tree rings as a measurement of age, that the statolith be defined as part of a giant squid, and that the rings in a statolith are a way of measuring age. Given the situation where all of these entities are completely defined in Cyc, identifying how these two methods of age measurement are similar would require identifying where the ancestry of the statolith rings and tree rings overlapped (potentially an entity such as age measurement). As the statolith and tree rings would be child entities of the giant squid and dendroclimatology entities, respectively, the path to identifying age measurement could be potentially highly tortuous. Given the
highly specific nature of this relationship, it may be unreasonable to expect that it be represented in Cyc, and as such the discovery of such a relationship would still be necessary. However, the extraction methods presented in Chapter 5 require plain text describing each entity, and the context in which they can occur, and do not directly translate to the Cyc hierarchy.

Cyc provides a valuable resource in identifying logical definitions of entities and their logical inheritance, however due to the relative sparsity of coverage, the level of conceptual overlap is required for relationship identification and relatedness calculation is less than that available when using Wikipedia. Cyc does have the potential to be used for identifying semantic relatedness using conceptual overlap, similar to RACO over Wikipedia, but additional experimentation is required determine the effectiveness of such an adaptation given the relative sparsity of Cyc and the complexity in determining an equivalent form of categorical overlap.

6.1.4 Resource Augmentation and Wikification

So far, this chapter has examined the representation and utility of alternate domains of application, and the semantic resources that are able to provide a similar representative framework to the semantic network used throughout this thesis (Wikipedia). However, there exist domains of entities that may not have a representative semantic network that describes the content of entities, an organisational structure, and semantic relations between entities. There exist two approaches in the creation of semantic network for such a domain: the manual creation of a new semantic wiki from scratch, and the adaptation of existing resources into a semantic network. The former approach is a time intensive task that requires the creation of new content describing each entity in the domain, the representative framework
of the semantic network, and the placement of semantic relations between entities. The latter of these two utilises an existing form of representation (e.g., a collection of textual documents, each describing an entity) and places these entities into an existing semantic network (Wikipedia) by applying semantic relations to the existing representation (e.g., linking words in the document to relevant Wikipedia articles). This process is known as wikification (Mihalcea and Csomai 2007). This is desirable as it allows the use of collections of plain text documents, specific to a reduced collection of entities, to be used in conjunction with Wikipedia’s inter-article links and those articles’ category membership, thus allowing methods such as RACO and the relationship extraction methods presented in Chapter 5 to be used over external document collections.

Wikification is performed by identifying terms in the body of a document of an article that are relevant to the document’s overall content. This is done by first identifying named entities in the document that correspond to the title of a Wikipedia article, and then testing that the named entity is relevant to the central subject of the document. If the named entity is considered relevant, then that term is linked to the corresponding Wikipedia article. For example, when Wikifying a document about cheesemaking, named entities such as cheese, curds and butter may be identified as being related and linked to the corresponding Wikipedia article, but stopwords (words such as is, the and and) and irrelevant named entities (e.g., water or wood) would not be linked. It is possible that some of these perceived irrelevant named entities may be validly linked to in certain situations, and the thresholding for determining named entity relevance for the purpose of wikification has been examined by Milne and Witten (2008b). The level of wikification is ultimately up to that of the annotator.
constructing the set, however Milne and Witten (2008b) were able to achieve levels of link insertion equivalent to that existing in Wikipedia articles.

Document collections are authored without regard to the editing guidelines stipulated by Wikipedia’s manual of style, and as such are unlikely to utilise Wikipedia style document structure or adhere to content guidelines. Thus there may exist some discrepancy in wikifying document collections that do not utilise an encyclopedic description style (Csomai and Mihalcea 2008). Furthermore, it is unlikely that each factual statement in these document collections will contain a reliable source of reference to support. A domain-specific document collection need not adhere to these guidelines provided each document describing an entity possesses enough contextual diversity to describe its relationship with cross-domain entities. The primary requirement of the document collection is that each document describing an entity contains enough contextual diversity to describe the relations to cross-domain entities.

The wikification of a document allows it to be described in the context of existing cross-domain articles in Wikipedia. This inserts the article into Wikipedia’s semantic network, but only using outgoing links from the document as the document does not exist as a Wikipedia article and thus cannot be pointed to by other articles. These inserted semantic relations allow the document collection to be used in conjunction with the methods described in this thesis: RACO utilises the category membership of linked-to articles to determine the semantic relatedness of the represented entities, but does not require that these representations be part of the category hierarchy. Using this property, any two documents describing entities may be compared using RACO. Firstly the documents must be wikified using the process described by Mihalcea and Csomai or Milne and Witten, after which the two articles may be compared directly using RACO to determine the category membership of the wikified named
entities within each article. Additionally, the basic text search and common third article methods of relation extraction may be used to determine the relevant semantic relationships between these two documents.

When using a document collection that has been wikified as the basis for the calculation of semantic relatedness via RACO or to identify the semantic relationships between documents, the results of these two processes will be dependent on the quality of the wikification itself: if named entities are erroneously linked to incorrect documents, calculations based on these links will also be erroneous. Using the process of wikification, existing document collections are potentially able to use the methods constructed in this thesis to calculate the semantic relatedness of contextually diverse entities and identify semantic relations that exist between them.

6.2 Alternate Domains of Application

In using the Cultural Heritage domain as a basis for the experiments in this thesis, certain domain properties were required: it was comprised of a set of discrete entities that represented distinct concepts (museum exhibits); and the entities in this set were multi-faced and could be related to other entities for a variety of reasons. For example, the Quartz reef mining model exhibit used in experiments in this thesis was a self-contained object represented in the museum space, appearing as part of a larger exhibit collection on gold mining. This exhibit contained relations to other exhibits in the collection by virtue of a common theme for the exhibit collection (that of gold mining), but in addition to these central themes the model contained references to geology, geography, and immigration.

This contextual diversity of a collection of entities is the chief factor in considering
alternate domains of application. Recall the wooden canoe example from Chapter 1: depending on the entity to which the canoe is compared, different elements of its content are identified as being relevant to the context in which it is placed. The contextual diversity of a collection of entities can be restricted by the format of description used to represent a domain (e.g., the WordNet hierarchical inheritance of synset meaning), thus restricting the diversity of relations available. For this reason, it is necessary that a domain appropriate semantic network exist for a domain to benefit from the methods described in this thesis.

It is desirable that any alternate domains are comprised of contextually diverse entities, and that there exists a need for identifying the semantic relatedness of domain entities and the semantic relations between them. In order to demonstrate the adaptability of the methods constructed in this thesis, I profile two alternate domains of contextually diverse content. These alternate domains of application were chosen based on the requirements that they contained entities with contextual diversity, and demonstrated a need for the semantic comparison of entities and discovery of relations. Based on a review of application areas of entity comparison, the domains of Film and Tourism were chosen as alternate domains of application. Both of these domains are described in detail in the following sections, containing a description of the entities within the domain, a discussion of the relational diversity of the domain, an examination of the domain specific representational resources available, and a discussion of how RACO and methods of relation extraction are applicable to these domains.
6.2.1 Film

The domain of Film and film related entities pertains to individual films and their content, actors, directors, producers and production crew. Associated resources include film review articles and film ratings as provided by viewers. There is a large level of interaction between person entities in the Film domain and individual films (as directors, actors and production crew work on multiple films), but very little interaction between the films themselves: beyond sequels (e.g., the Star Wars films) or shared universe films (e.g., Marvel superhero films), there exist few direct film-film interactions. This differs from the Cultural Heritage in which all entities have the potential to interact. Resources such as film reviews go some way to identifying these film-film relations in terms of genre and similar themes, but are often presented in unstructured form and require additional processing to identify these relationships, making it a desirable target for the application of RACO and the methods of semantic relationship extraction presented in this thesis. The remainder of this section describes how these Film specific resources can be adapted for use with the methods in this thesis, and the applications of the identified film-film relations to tasks of film relatedness, film recommendation and explanation of recommendation.

Benefit to Domain

Films are conceptually rich entities: viewed solely from a finished product perspective, they contain characters that tell a story from start to finish. Using this simple characterisation, relations based on the story, genre and characters may be drawn between individual film entities. Given the diversity of content associated with each film entity, there exists scope for identifying qualities of relatedness between films
(e.g., using aspects such as same genre or geographic locations occurring in multiple films).

Of the two domains presented in this chapter, Film contains the most parallels with the Cultural Heritage domain: the identification of relations between Film entities may be used for educational purposes (e.g., in a high school English class where a film or film style is being analysed, knowledge of related themes is beneficial), or entertainment purposes (e.g., identifying the exhibit or film most relevant to a visitor/film goer’s interests). In this respect, the identification of relationships between film entities is beneficial for identifying relations between extra-film entities (directors, actors, etc.) and films as well as film-film relations (e.g., identifying a common narrative device or film style). For example, identifying directors that utilise similar filmographic style or similar genres such as identifying that Roland Emmerich and Michael Bay film disaster movies with many explosions.

In addition to identifying the semantic relations between Film entities, content-based film recommender systems are able to make use of RACO to calculate the semantic relatedness of Film entities. This task requires the identification of semantic relations between films and film-related entities, and as such requires the use of a contextually diverse semantic network to represent the domain. The following section describes an approach to identifying such a semantic network.

**Domain Specific Representation and Resources**

Qualities outside the limited narrative aspects of a film can include content related to the production of the film: actors, directors, producers, film crew as well as award nominations (e.g., multiple films nominated for a best film award in the same category may be considered related). The quantity of information used to represent a film
entity will depend on the application. For example, when trying to solve the six
degrees of Kevin Bacon problem, wherein the task is to link a given actor to Kevin
Bacon using co-starring links, only a subset of relations are required as a basis to solve
this task. Specifically, only the \textit{costar-of} relationship would be required to construct
a semantic network to use as a basis for identifying the separation between actors.
For other tasks the content of each entity utilised and the relations identified between
them would be different.

Films contain a great amount of contextual diversity, and their use in specific
applications utilise these facets for different purposes. The scope for entity diversity
can be extended beyond solely films: entities such as actors, film studios and fran-
chises are all relevant to the Film domain and can be validly included in a digital
representation of the Film domain.

The most comprehensive digital representation of the Film domain is the Internet
Movie Database (IMDb), which is comprised of film and person entities. Person
entities contain lists of films and television episodes in which the person has acted,
written for, directed or played some part in. Film entities contain descriptions of the
film’s plot, cast and crew of the film, images of the film’s production and promotional
images and user reviews of the film. IMDb has been constructed using a folksonomic
approach, similar to that of Wikipedia, and while there exist fewer checks and balances
to ensure entity quality and validity, multiple studies use it as a de facto standard
of film representation (e.g., Cantador \textit{et al.} (2008), Malin \textit{et al.} (2005), Szomszor
\textit{et al.} (2007)). IMDb contains only domain related entities, films and people working
in the film industry, and forms a semantic network of film-person (e.g., \textit{acted-in},
directed, wrote). These film and person definitions contain explicit links between
one another, however the content of films and biographical information about the
person entities contains only textual data, requiring additional analysis to identify any relevant relations contained within.

In addition to the film-person relations, films are classified using a tagging system to label films by their themes, genres and central plot points. This classification system differs from a hierarchical classification system, such as used in Wikipedia, in that there exists no inheritance between individual tags and forms a flat classification scheme.

In an effort to refine the content and relation definitions in IMDb, Hassanzadeh and Consens (2009) adapted inter-entity relations in IMDb to a traditional ontological
definition of a semantic network, with all relations defined as RDF Triples. This added structure allowed the network structure of IMDb to be easily queryable but ignored any contextual relations that may be present in plot descriptions and user reviews.

While existing work utilising the Film domain touches on aspects of film relatedness (e.g., movie recommender systems), the use of relationship extraction to identify links between films has thus far been unexplored. The following section goes on to describe how RACO and methods of relation extraction constructed in this thesis can be used in conjunction with the Film domain.

Applications and Computational Use

As with Cultural Heritage, measures of semantic relatedness and semantic relations can be used as the basis for user modelling applications and recommender systems. Where the Film domain differs from Cultural Heritage is in domain resources: there exists a large amount of expert annotation of film content in the form of film reviews. This annotated content exists in the form of online film websites (e.g., Rotten Tomatoes\(^2\) and IMDb) and from professional film reviewers. Another difference is the physical distribution of entities: a Quartz reef mine model may be present at multiple museums in different forms, and thus will be regarded as separate cognitive entities. This differs from the Film domain as a film entity is regarded as the same entity no matter where it is encountered (i.e., Jurassic Park is always Jurassic Park). This conceptual definitiveness allows for a great deal of overlap in annotation. This abundance of user ratings allows film domain recommender systems to utilise collaborative-filtering user modelling techniques.

Content-based recommender systems are able to directly make use of the con-

\(^{2}\)http://www.rottentomatoes.com
tent of film entities to determine relatedness, and have been used in tasks of film recommendation (Cantador et al. 2008; Salter and Antonopoulos 2006). Hybridised recommender systems that utilise both content-based and collaborative-filtering in their construction have also been demonstrated to perform well over these same film recommendation tasks (Cantador et al. 2008; Salter and Antonopoulos 2006; Szomszor et al. 2007). Content-based recommender systems are able to utilise the calculation of semantic relatedness between film entities to calculate film relatedness.

As with relations between Cultural Heritage entities, recommender systems that utilise explanation of recommendation are able to utilise extracted semantic relations between film entities. Herlocker et al. (2000) uses the example of a supporting evidence of a film critic to increase user trust in a film recommendation. The construction of explanations of how a user’s interest model overlaps with a film recommendation is able to utilise relationship extraction methods, such as those described in this thesis, to automatically inform a user of a system’s reasoning, thus increasing a user’s trust in the reasoning of the system. The following section describes how these methods of relationship extraction and RACO can be adapted to utilise film specific resources and semantic networks.

Adaptation of Methods

The description format used to represent film entities in Wikipedia differs from that of IMDb, and thus the level of adaptation required to use RACO and the methods of relation extraction used in this thesis will differ dependent on the organisational structure of each of these semantic networks. The use of Wikipedia as the semantic network over which the methods are used requires no adaptation of methods as the principles of conceptual overlap (the basis of RACO), and the taxonomic links required
for the methods of relation extraction presented in Chapter 5 are unchanged. The domain of Film is directly able to use Wikipedia as a basis for representation in calculating RACO scores and identifying relationships between film entities, as there exist film, actor, director, producer, etc. entities in Wikipedia. However, the sole use of Wikipedia to represent these entities disregards the additional valuable information possessed by IMDb (e.g., reviews, tags, plot keywords, etc.). To take advantage of this additional in-domain content it is desirable to consider IMDb as an alternate semantic resource over which to calculate the semantic relatedness of entities and identify the relationships that exist between them.

An approach to augmenting the content of IMDb to function with the RACO and the relation extraction methods, is to only examine the textual content associated with each film, actor, director, etc., and use a process of Wikification (see Section 6.1.4) to place it in context of Wikipedia content. This approach eschews all existing links and categorisation system present in IMDb in favour of Wikipedia’s semantic network, and simply uses IMDb as a collection of simple text documents. This approach is undesirable as it does not take into account the taxonomic structure of IMDb.

IMDb possesses a broad classification system, classifying films in terms of genre (e.g., Romance, Adventure, etc.), and people in terms of film role (e.g., producer, actor, etc.). This flat classification system does not possess enough diversity to act as a simple substitute for Wikipedia categories when determining the conceptual overlap between film entities. Qualities that films have in common are people, genres and plot keywords. As described, the property of genre is a restrictive classification system (there are a total of 26 genres that subsume all films), and while there exists overlap in genre classification (e.g., a film can be classed as ADVENTURE and COMEDY) the diversity is not great enough to provide enough variation in relatedness measurement.
People such as actors and producers rarely restrict themselves to a single genre, and thus identifying the interaction of two groups of actors (each group associated with an individual film) would yield films they worked on together, rather than a group of common classifications (as is done by RACO).

The plot keyword system of characterising the content of a film provides the best parallel to the use of categories to provide conceptual overlap. Plot keywords are manually provided for each film in the form of simple textual descriptors. There exist over 10,000 plot keywords such as dinosaur, experiment-gone-wrong and hit-in-crotch. By themselves, examining the keyword overlap is similar to computing the cosine similarity for two term vectors, and while studies have demonstrated this to be effective (e.g., Grieser et al. (2007)), it does not utilise the contextual diversity of a films content by identifying the categories (plot keywords) of related entities (in this case, related films).

In order to overcome this lack of an organisational hierarchy, other aspects of IMDb must be used to determine an equivalent form of category membership. In addition to an equivalent form of category membership, a method of identifying links to related films is required. Fortunately, there exist elements of film entities within IMDb that may be used as a substitute for these Wikipedia qualities: each film entity in IMDb contains a list of film recommendations as part of its definition in the form of a list of 5 films deemed to be similar to the current film. This list of related films may be utilised in the same manner as inter-article links are used in the formulation of RACO. Using film recommendations, and their respective plot keywords, a measurement of Related Film Plot Overlap (RFPO) may be formed to identify the relatedness of film plots. This creates an equivalent method of identifying the conceptual overlap between film entities, over a film specific semantic network.
Experimentation is required to determine if this substitution an adequate parallel to RACO, and is a valuable area of future research.

The additional content provided by users on IMDb (e.g., film reviews) provides valuable insight into the subtext of films and their overall perception by film viewers. These additional bodies of text provide contextual diversity in describing elements of production not captured in plot summaries or lists of actors. Identifying relations to plot themes or details of the film production process (e.g., using specialised filming techniques) are the diversely structured relationships sought after by the methods constructed in Chapter 5. Beyond the identification of relationships using simple textual search, the methods described in Chapter 5 require a taxonomic structure and marked up text to be able to identify relations. All bodies of text in IMDb are plain text (e.g., plot summaries, reviews) and contain no hyperlink markup, making them unsuitable for taxonomic analysis. In order to approach the identification of relations in bodies of text by utilising the taxonomic structure of an existing semantic network, the text must first be augmented. This can be done by identifying existing entity names in IMDb (e.g., film or actor names used in a review) and linking these to the appropriate entity. Using this method, identifying relationships based on reference to a common third entity is possible, but any methods that utilise the category hierarchy are still unable to be used over this limited taxonomy. Another approach is to utilise the Wikipedia semantic network by wikifying plot descriptions and film reviews. This approach allows all methods described in Chapter 5 to be used, but relies on a semantic network external to IMDb. Ultimately the requirement that an extensive hierarchical taxonomy exists causes a subset of the methods of semantic relation extraction described in this thesis to be restricted to semantic networks with an inherited hierarchy. Specifically, the Common Ancestor and Sibling Reference
methods of relationship identification require the existence of a system of hierarchical categorisation to identify common membership of categories.

The domain of Film is a valid area of application for the methods proposed in this thesis as it is comprised of contextually diverse entities and, depending on the source of description used, is interwoven with multiple semantic relations. Methods of relationship identification and semantic relatedness calculation developed in this thesis are able to utilise Wikipedia or alternate semantic networks, but require adaptation when used outside semantic networks with a hierarchical inheritance structure. Ultimately, experimentation over this domain is required to determine the effectiveness of these methods, but in this section I have performed an analysis of how RACO may be adapted to utilise IMDb-specific properties to calculate the semantic relatedness of films, and how the existing content of IMDb may be augmented using wikification so the methods of relation extraction constructed in this thesis may be used in conjunction with IMDb content.

6.2.2 Tourism

As used in this thesis, the domain of Tourism refers to entities that relate to travel destination, methods of travel, sightseeing locations, accommodation and leisure activities one partakes of while on holiday. I refer to this collection of entities as Tourism rather than Travel due to the application of RACO and relationship discovery being useful for planning holidays, and to extend beyond the simple identification of relationships between geographical locations. This distinction is important in that the analysis of the relatedness of geographical locations is potentially reducible to the simple discovery of relationships between encyclopedic descriptions, making it too similar to the Cultural Heritage domain, and thus requiring little or no adaptation
to RACO or the relationship extraction methods produced in this thesis. In this section, I present an analysis of the usefulness of identifying semantic relations between Tourism entities and their semantic relatedness, and the process of applying RACO and the relation extraction methods constructed in this thesis to Tourism-specific resources.

**Benefit to Domain**

The Tourism domain is comprised of entities that represent physical locations, related activities and modes of transport. Users in this domain plan tours, shop for accommodation and book travel tickets. Often referred to as *e-tourism*, the use of tourism websites in planning these trips and related activities has become more prevalent in the past decade (Prantner *et al.* 2007). These resources are used by individuals wishing to create their own travel plans or travel agents planning a tour for customers (Jakkilinki *et al.* 2007). There are, however, alternate resources that a user planning their own holiday may use to elucidate or better inform their choices. These alternate resources (described further in the following section) describe connections between tourism domain entities that are not present in simple booking systems.

Identification of how tourism entities such as destinations, accommodation and modes of travel can be used by users to better plan a trip, and to demonstrate an interaction between a user’s interests and activities at a location or properties of a tourism destination. Tourism-specific semantic networks allow relations specific to the domain to be identified, reducing the likelihood of irrelevant relations being identified (recall the discussion of domains in Section 2.2). For example, Wikipedia contains definitions for geographical locations, methods of transport and other entities relevant to the Tourism domain, but the cross-domain definition of these entities (as opposed
to a tourism centric representation) can result in the extraction of relations irrelevant to the tourism domain such as comparing destinations on the basis of Gross Domestic Product. While valid, this relation is of little meaning to the tourism domain.

The use of domain specific resources (e.g., Wikitravel) in identifying domain-specific relation reduces the likelihood of off-topic relations, i.e., leads to contextual restriction. For the purposes of identifying relations intended for tourism purposes (e.g., activities at a location, places to stay, cultural events, etc.) the use of a Tourism domain specific semantic network is desired over an encyclopedic network such as Wikipedia. For this purpose, Tourism specific semantic networks are better suited to providing a knowledge base from which to draw semantic relations and to identify the (tourism specific) semantic relatedness of entities. The organisational structures of various Tourism domain semantic networks differ from that of Wikipedia, and RACO and the methods of relationship extraction described in Chapter 5 require alteration to utilise this structure. The following section details the alternate semantic resources available for this task, and the application of the measures developed in this thesis to these semantic networks.

**Domain Specific Representation and Resources**

When one considers Tourism as a domain of information rich entities, the most common class of entities is that of geographical locations (i.e., travel destinations). Unlike film entities, geographical locations possess less facility for attributive representation. Qualities such as latitude, longitude and country of location are possessed by all geographical entities, but description of their content will vary considerably from location to location. For example, a **city** and a **museum** are both geographic locations but exist at different levels of granularity (a **city** may contain a **museum**).
In addition to diverse geographical locations, entities such as events, tours, accommodation and transport are utilised in the representation of catalogues of Tourism entities (Ricci 2002). The Tourism domain also contains many specialised relations between entities that may not be present in broader multi-purpose semantic networks (e.g., Wikipedia and Cyc), such as the specific interaction of airline routes, tourist attractions and seasonal variation in tours. This diversity in content and relational interaction creates a great deal of contextual diversity, and existing tourism semantic networks and ontologies approach the representation of these heterogeneous entities in multiple ways, from the global definition of tourism entities (e.g., The Mondeca Tourism Ontology: http://www.mondeca.com), to facilitating the interoperability of multiple tourism ontologies (e.g., HarmoNET: http://www.harmonet.org, or the Open Travel Alliance Specification). For a detailed discussion of the differing tourism ontologies, see Prantner et al. (2007).

Professionally curated tourism ontologies and semantic networks are used widely by travel agents. The cataloguing of travel destinations, activities, accommodation and travel options allows a travel agent to easily search through options relevant to a given criteria (e.g., flights leaving on a certain date or hotels in a given city). These semantic networks are comprehensive lists of airline schedules, accommodation booking details and tour information, precisely the information needed when booking a holiday, but they do not describe the contextual diversity of the location or related entities (Prantner et al. 2007).

In contrast with industry experts constructing and maintaining an ontology of tourism entities, several studies have examined existing non-industry created resources as the basis for information rich descriptions of tourism entities (e.g., Karoui et al. (2004), Maedche and Staab (2002), Ricca et al. (2010)). Guide websites such
as Lonely Planet\textsuperscript{3} provide detailed descriptions of locations, related accommodation, eating options and sights within the area. These heterogeneous collections of tourism web pages have been adapted to formalised semantic networks, designed to take advantage of the contextual diversity of the entities described, and the domain centric relations (e.g., has-attraction, has-accommodation, is-transport-option, etc.) (Karoui \textit{et al.} 2004). Folksonomic tourism catalogues are able to utilise travellers’ understanding of the domain’s entities to enhance the contextual diversity to include relations and content based on personal experience or understanding (Ding \textit{et al.} 2008). For example, Wikitravel is a Folksonomic Tourism semantic network (wiki) that describes the tourism aspects of a location, relevant hotels, restaurants and attractions, other locations that may be relevant to the current location, and qualities that may make a location desirable to visit (e.g., a strong focus on eco-tourism or sporting events).

\textsuperscript{3}Lonely Planet: http://www.lonelyplanet.com
Professionally curated networks do not contain the contextually diverse descriptive content required to use RACO (or an equivalent measure of semantic relatedness that utilises contextual diversity) or the methods of relation extraction described in Chapter 5, making the use of alternate domain-specific collections of such knowledge necessary (Maedche and Staab 2002). Alternate semantic networks such as Wiki-travel and existing collections of Tourism such as the Lonely Planet website are able to be adapted to provide resources with greater contextual diversity and the facility to utilise the methods constructed in this thesis.

**Application and Computational Use**

Entities that are represented by a large amount of information describing the content of the entity and the manner of its interaction with other domain entities provide a greater facility for the basis of content-based recommender systems. The Film domain was also able to take advantage of collaborative filtering methods to identify relations between entities, however there exists a much greater sparsity of entity ratings in the Tourism domain to form the basis of any collaborative filtering system (Ricci 2002). This is especially true when constructing a tour and in situations where user ratings are based on the tour as a whole: a single divergence in the tour’s structure will negate a rating from even a highly similar tour (Ricci 2002). For this reason, content-based user modelling forms the basis of a large part of the research into travel and tourism recommender systems (Jakkilinki et al. 2007; Ricci 2002).

Multiple systems have utilised an interview style interface in constructing tourism recommender systems that parallels the process taken with a human travel agent by posing a series of questions to the user (e.g., *Where do you want to go? What do you want to do?*) (Jakkilinki et al. 2007; Zins 2003). These approaches make use of
traditional tourism ontologies as used by industry, and less focussed on the contextual interaction of the entities. However, Tomai et al. (2005) identifies that the description of locations is important to the end-user of a travel recommender system in allowing them to plan activities while at a location. Specifically, Tomai et al. state that the use of user profile for the purposes of personalisation requires an explicit analysis of the content of a tourism entity (e.g., a location or activity) in order to test whether or not the entity is relevant to the user profile, rather than the simple construction of a tour based on a stated set of requirements. For example, a user may specify that they wish to travel to a tropical location and stay at a beach resort. Using these criteria, the system would identify the first location that satisfied these requirements regardless of the reason that the use had for providing these criteria. Analysis of the user’s interests may indicate a preference for natural elements rather than for clubs and bars or sporting activities.

Tomai et al. also utilise the counseling interaction style used in other tourism recommender systems, but note that this style requires natural language interaction with the system as the user may present desires in the form of, e.g., *I have two days to spend in X, what do you propose I do?* or *Today I want to do some sightseeing in X and then go to the sea.* In coming up with sights or activities that satisfy these requests, the recommender system must communicate these to the user using an explanation of why the activity was chosen, particularly if the interaction with the system takes place with a counseling style as described by Tomai et al. (2005) and Zins (2003). Ricci (2002) also notes that the justification of recommendation is also an important step in the tour recommender systems. This explanation of recommendation is a direct application of the semantic relationship extraction methods described in Chapter 5. For example, if a user wants to go sightseeing in Copenhagen, a recommender system
may recommend that the user see the statue of The Little Mermaid; when asked to provide a rationale for this recommendation, the system may identify that the user has an interest in fairy tales, and that the statue is based on a character from a fairy tale. In this instance, the identification of the content of the statues subject and how it relates to the user’s interests form the basis of the explanation.

The automatic identification of these intricate relationships between Tourism entities is an application that the methods constructed in this thesis are able to contribute to. The domain of Cultural Heritage has been thoroughly explored with regard to the identification of relations between entities (e.g., Cox et al. (1999), Wang et al. (2008), Yatani et al. (2004)), but this application of these approaches to the Tourism domain has thus far not been approached. The methods of relation extraction and semantic relatedness calculation formulated in this thesis are able to utilise existing Tourism semantic networks to produce these relations. The adaptation of these methods to Tourism-specific semantic networks is described in the following section.

Adaptation of Methods

The Tourism semantic networks of Lonely Planet and Wikitravel use alternate methods of organising their content, using a traditional static HTML format and the collaborative Mediawiki framework respectively. However, both of these resources utilise a similar entity representation: describing the content of an entity using a body of text with links to related entities within the text. The structure of the semantic network of each of these differs greatly. In contrast with Wikipedia, Wikitravel uses almost exclusively administrative categories to organise articles, categorising the majority of articles with regard to their level of polish and completion. There is a total of 40 categories used to organise the 25,000+ articles in Wikitravel, with only
one category (the **Huge City Articles** category) used to classify the content of articles. This administrative use of the Mediawiki category system differs greatly from Wikipedia, and does not allow it to be used as the basis for content classification that RACO requires.

Multi-faceted categories subsuming collections of entities are required for RACO’s use, indeed they are the chief reason for its use: to identify the category overlap of these diverse categories. Without such a system of multiple category subsumption, RACO is not able to be used. Conversely, the directory structure used on the Lonely Planet website may be used as the basis for forming a hierarchical classification system. Web pages on the Lonely Planet website are organised using a sub-folder system based on geographical location (e.g., `europe/germany/berlin/reichstag` is the folder organisation for the **Reichstag** entity, and `australia/melbourne/sights/architectural-cultural/young-jackson-s` is the folder organisation for the entity representing Young and Jackson’s pub in Melbourne). This classification scheme provides several layers of subsumption, and several classifications exist at multiple depths (e.g., SIGHTS, RESTAURANTS, etc.), however as with the IMDb genre classification scheme, there only exists a small set of non-geographic classifications for which there is little overlap. For example, restaurants are only classified in terms of the RESTAURANT classification, and are not subsumed by the ENTERTAINMENT-NIGHTLIFE classification. This limited set of categories does provide for overlap between similar entities in different locations, (e.g., a restaurant in Melbourne being related to a restaurant in San Francisco), but does not allow multiple overlap of non-geographic classifications. To properly utilise RACO, entities must be classified in terms of multiple classifications identifying different aspects of the entity’s content.

Another contrast between the two semantic networks is the style of content used
to describe entities: Wikitravel presents information similar to an encyclopedic style (although only describing Tourism-relevant information), whereas Lonely Planet includes traveller reviews and tags in addition to encyclopedic content. The additional use of tags to identify entity content and qualities provides a solution to the problem of little classification overlap when using solely the folder-based hierarchy of Lonely Planet’s website. Tags are selected by users from a predefined collection containing tags such as romance, sport, and fresh produce. As with the use of plot keywords in IMDb, these tags possess the ability to provide a system of classification overlap, similar to the principle of conceptual overlap utilised by RACO. The use of inter-entity links provides the explicit reference to related entities as used by RACO, by determining the overlap of tag clouds from these related articles, a Tourism based equivalent of RACO may be calculated between entities represented in the Lonely Planet collection of Tourism entities. The lack of a comprehensive classification schema does not allow Wikitravel to be used to calculate semantic relatedness using the principle of conceptual overlap as defined in Chapter 4.

It is worth noting that it is entirely feasible to utilise Wikipedia as the basis for representation when computing RACO scores of Tourism entities: Wikipedia contains many Tourism entities (e.g., destination, modes of transport, and even well known hotels), and the existing category hierarchy allows the RACO scores of these entity pairs to be easily calculated. However, when extending to a low level of granularity, Wikipedia does not possess some entities that are contained in Wikitravel or the Lonely Planet website. Entities such as less well-known hotels (e.g., The Nunnery Hostel in Melbourne exists on the Lonely Planet website, but not in Wikipedia) are not as completely represented in Wikipedia due to them not being noteworthy enough to be represented on Wikipedia (Wikipedia 2011g). Where Tourism entities exist in
Wikipedia, it is entirely feasible to use it as the basis of RACO calculations, however, using multiple sources of representation may cause discrepancies between scoring due to the differing focus of entities on different resources (e.g., description of a country’s GDP on Wikipedia versus descriptions of its pristine beaches on Lonely Planet).

The use of Wikipedia with methods of relationship identification encounter additional problems when domain specific relationships are desired. The basic principles upon which the extraction methods are built are intended to function with regard to any domain, but restricting the representation to Wikipedia reduces the applicability of identified relationships. Relationships identifying major trading partners in a region may be desirable to some travellers (as could be identified by using Wikipedia based representations of Tourism entities), but smaller minutiae which may be desired by travellers, such as the specific routes taken by tours and how they relate to local cultural events, are less likely to be described using the encyclopedic content of Wikipedia articles (Wikipedia 2011g). Unfortunately, in moving away from Wikipedia as a resource, certain relationship extraction methods become infeasible.

The methods of relation extraction based on the simple textual occurrence of entity names and common third entity links are both able to be used over Wikitravel and Lonely Planet, as entities in both semantic networks contain descriptive text and links to related articles. As with IMDb, the lack of a hierarchical classification scheme does not allow the relationship extraction methods utilising common ancestry and sibling links to be used over Wikitravel. However, Lonely Planet’s website content organisation utilises a hierarchical system of classification based primarily on geographical location. This allows relation extraction methods based on ancestry to be used. Lonely Planet’s geographical classification system differs greatly from
Wikipedia’s category hierarchy, and the principles of ancestry overlap and sibling entities may require further adaptation to function over this new semantic network.

The relationships present in Wikipedia possess a diversity such that the use of pattern-based methods of recognition fail to encompass a large variety of relationship types. When considering the Tourism domain, this diversity remains a desirable attribute: simple pattern-based identification of relationships (e.g., \(X\) is capital city of \(Y\)) limits the amount of discoverable information. For example, being able to identify the significance of a cultural event with respect to local geography may aid a tourist in identifying the best time of year to travel to that destination: a tourist may not want to go hiking when the local tribal ceremony to welcome the seasonal rains is taking place. The methods of relationship extraction proposed in this thesis are ideal for identifying these contextually diverse relationships.

Ultimately, experimentation is required to accurately determine the effectiveness of substituting the Lonely Planet’s website network for Wikipedia category hierarchy in the calculation of semantic relatedness and identification of diverse semantic relations between Tourism entities, but there exist the requirements (i.e., a hierarchical network, textual descriptions of entities, inter-entity links, and tags acting as a secondary classification system) that make it a suitable knowledge base for this task.

6.3 Summary

The primary contribution of this thesis has been the utilisation of Wikipedia as a semantic resource in the calculation of semantic relatedness between entities, and the identification of semantic relationships existing between them. Due to the diversity of content present in the Cultural Heritage domain, available resources (e.g., anno-
tators, document collections), and the benefit to the domain, the Cultural Heritage domain has been used in the development and testing of the methods in this thesis. The ultimate intention of the methods in this thesis is that they be applicable to a wide variety of domains that contain contextually diverse entities. Due to time constraints, the adaptivity of these methods with regard to the domain of application and the semantic network used as the basis of conceptual representation have not been explicitly tested. In this chapter I have examined two alternate domains of application, and several semantic resources that have the functionality to represent collections of contextually diverse entities.

The domains of Film and Tourism are both comprised of contextually diverse entities, whether they be films, gaffers, geographical locations, or carnivals, each contain multiple facets that possess relations to other in-domain entities. Tasks within these domains are able to benefit from the calculation of semantic relatedness between entities for the purpose of identifying entity relatedness in content-based recommender systems, and for identifying the semantic relations that links these entities for use where explanation or transparency of recommendation is required. The domain-specific semantic networks used to represent these domains will identify different semantic relationships to those of a cross-domain semantic network such as Wikipedia as these entities are described in context of other in-domain entities, avoiding potentially non-relevant relationships. Additional resources offered by these domains (e.g., film reviews, hotel reviews, user tags) increase the scope for identifying contextually diverse relationships. Due to the large variety in editorship and structure of user content, non-pattern-based methods of relationship identification are desirable for identifying these relationships.

Cross-domain semantic networks such as Cyc and digital encyclopedias also have
the facility to represent large, broad-scale collections of knowledge, however, these representations contain drawbacks. Cyc is intended as a comprehensive, structured representation of all human knowledge and requires defining conceptual entities in terms of logical is-a predicates and argument frames. This representation requires much work to define complex interactions between entities (such as the measuring of the age of a giant squid similarly to that of a tree), resulting in a network that, while possessing great potential for identifying interesting semantic relations between entities, is too sparsely populated to provide the relations available from analysis of Wikipedia articles. Digital encyclopedias, too, provide an encyclopedic resource similar to that of Wikipedia, but lack the categorisation hierarchy upon which the methods in Chapter 4 and Chapter 5 rely.

Domain-specific wikis provide a valid base for identifying semantic relations between entities provided the limited contextual diversity is not a hindrance to any applications using these as a representational semantic network. The exclusive utilisation of domain specific semantic relations has the potential to provide interesting applications such as the construction of narrative given a set of entities in a semantic network of fictional entities. Ultimately, the functionality of domain-specific wikis must be evaluated respective to their domains of application to provide insight into their suitability for representation.

The use of wikification to augment existing document collections demonstrates the versatility of Wikipedia as a semantic network, as well as the versatility of RACO in that it may be used (in theory) over extra-Wikipedia document collections, rather than only entities existing within Wikipedia.

While this examination of alternate domains of application and alternate representational semantic resources has identified valid fields of enquiry, further experimenta-
tion is required to examine the effectiveness of RACO and the methods of relationship identification proposed in Chapter 5 outside the domain of Cultural Heritage and Wikipedia.
Chapter 7

Conclusions

7.1 Summary and Research Contributions

In the course of this thesis I have addressed the stated research objectives of the construction of a measure of semantic relatedness intended to function over a semantic network of heterogeneous entities, and the problem of relationship identification and extraction among a collection of contextually diverse entities. The identification of these relationships has been performed through analysis of the semantic network which contains these entities as opposed to current pattern based approaches used over the same semantic network and alternate data sources.

In addition to these objectives, I have constructed a gold standard set of human judgements of semantic relatedness over a conceptually diverse collection of entity pairs. This is an important contribution as existing data sets of semantic relatedness scores are designed for tasks of lexical semantic relatedness, specifically word sense disambiguation. The dataset constructed in this thesis utilised a collection of defined entities with which annotators were familiar: museum exhibits. The heterogeneity
of these entities allowed them to be used as the basis for experiments of semantic relatedness utilising the contextual diversity of these entities in their comparison.

In reference to the first stated objective of this thesis: *the identification of a method for calculating the semantic relatedness between entities in a collection of contextually diverse entities*, I have presented a novel measure of semantic relatedness that takes advantage of the categorisation system present in the encyclopedic semantic network Wikipedia (RACO). This approach to determining semantic relatedness is unique from existing measures of semantic relatedness and similarity as it utilises the content available in the description of entities, rather than dictionary definition or frequency distributions as is done by multiple existing methods. This approach allowed an entity to be categorised according to related entities, thus creating a contextual classification of said entity. The performance of RACO was evaluated using the constructed collection of museum exhibit relatedness judgements, and found to exceed the performance of existing document similarity and state-of-the-art measures of semantic relatedness. Existing measures of semantic relatedness are unable to utilise the range of Wikipedia content due to being restricted to the availability of links present in the semantic networks over which they were designed. In using a de facto representation of curators’ thematic clustering of exhibits (the physical distance between exhibits), RACO achieved a higher correlation with visitor judgements of exhibit relatedness. The placement of exhibits in the museum space reflects a layout intended by the curator to highlight a specific quality or context of the collection. Regarded in isolation, individual exhibit pairs may not reflect this curatorial grouping and other contexts become more pronounced. RACO is able to incorporate the definitions of entities across multiple contexts to determine their overall related-
ness, taking into account the diversity of content not utilised in previous measures of semantic relatedness.

The performance of RACO was also evaluated over existing datasets containing word pairs and relatedness judgements. Terms within these datasets were both manually and automatically aligned to Wikipedia articles in order use RACO and another existing Wikipedia based measure of semantic relatedness (the Wikipedia Link Vector Model, WLVM). When compared to the performance of WLVM, RACO did not achieve as high a correlation with the gold-standard judgements provided by human annotators. However, it was shown that the performance of measures of semantic relatedness over aligned data vary greatly depending on whether manual or automated annotation is used. As a final demonstration of the contribution of RACO in identifying the semantic relatedness of diverse entities, I described its use in the identification of appropriate topic model labels. For this task, Wikipedia is desirable as it provides a large and diverse basis from which to draw topic label candidates; it also allows for RACO to be used as a basis for determining the appropriateness of these candidates relative to human judgements.

In reference to the second stated goal of this thesis: *the construction of methods of semantic relationship extraction to identify relationships between entities contained in a semantic network of contextually diverse entities*, I have presented four methods of relationship extraction that use taxonomic links between entities to identify relevant portions of text that describe a relation between pairs of articles. The four separate methods of relationship extraction utilise the semantic network structure and article content of Wikipedia articles in a variety of ways to extract multiple semantic relationships between Wikipedia articles. The approach of using the taxonomic structure of Wikipedia to identify relationships is distinct from existing approaches
that utilise pattern or rule based approaches to identify semantic relations. Rule and pattern based methods of relation extraction are unsuitable for the diversity of content present in Wikipedia as relationships are described using non-uniform structure. In utilising the semantic structure of the network containing the entities, I present an approach to identifying relationships that is not dependent on relying on consistent textual structure to indicate the presence of a relation.

In addition to the formulation of four novel methods of relationships extraction, I present a method for the evaluation of these methods through the manual annotation of extracted relationships. Pilot phases of this evaluation method demonstrated that the annotator agreement was comparable to existing manual annotation tasks over a similar experiment (the manual annotation of the perceived truthfulness of statements). When the annotation task was expanded to evaluate a greater number of relationships, the consistency of annotation lessened and successive iterations of large-scale relationship evaluation yielded mixed results. On the key criteria of relationship validity, however, patterns regarding the overall performance of individual extraction methods were present. The methods of identifying a common ancestor of two entities, and relating both entities to a common third entity were judged as having extracted a high proportion of valid relationships across multiple groups of annotators and for multiple sets of extracted relationships. The results of multiple evaluations indicate that these methods are able to successfully extract relationships between entities with diverse conceptual structure, however, further evaluation is required to identify the consistency of extraction over alternate sets of heterogeneous concepts.

The methods presented in this thesis utilise the domain of Cultural Heritage to provide context for evaluation and application, however, these methods are intended
for use over alternate collections of heterogeneous entities with sufficient contextual diversity. The method used for description of these domains also contains potential for diversity from the semantic network used throughout this thesis, Wikipedia. The form of description used for the domain can influence the type of relationship extracted, allowing entities to be described with relation to one another using an in-domain relationship. For example, using a tourism specific semantic network (e.g., Wikitravel, or an adapted version of the Lonely Planet website) one is potentially able to identify relationships between two countries in terms of what flights connect them, whereas an encyclopedic network such as Wikipedia is able to extract relationships based around trade agreements or common membership of international organisations. Furthermore, the existence of methods such as Wikification allows the construction of specialised semantic networks based on existing document collections that directly reference the Wikipedia semantic network and concepts within this network. RACO and the relationship extraction methods of common third article reference and textual search only require that Wikipedia exists in order to function, not that the articles being compared exist in Wikipedia. As such, the methods constructed in this thesis can potentially be used over domain-specific document collections that contain sufficient contextual diversity to describe the content of an entity and the placement of it within the context of other entities.

7.2 Future Work

The approach of this thesis has dealt with the determination of strength of relatedness between diverse conceptual entities, and the identification of specific relationships between these entities. This research has a number of natural progressions, both as an
extension of the use of these methods in the application domain of Cultural Heritage used in this thesis, and in the adaptation of these methods to alternate domains of heterogeneous entities.

Recommender systems have been mentioned throughout this thesis as a direct application of the methods formulated for semantic relatedness calculation and relationship extraction, and are a logical next step in regard to user evaluation and method application. Deploying extracted relationships and pair-wise relatedness scores as part of a larger, content-based recommender system would provide insight into how end-users perceived the validity of scores (via recommendation) and relationships (via explanation of recommendations) identified by the system. With reference to Cultural Heritage, the ability to identify museum exhibits relevant to an existing collection of a visitor’s interests may allow for the construction of personalised tours within a museum environment in the form a pre-constructed tour based around a visitor’s interests, or dynamically as they tour a Cultural Heritage site. The integration of semantic relatedness calculation with recommender systems necessitates an interface with user modelling and the collection of visitor interests. In this regard the semantic relatedness of museum exhibits can be used as a basis for a recommender system, providing a natural progression into a direct user application for the RACO measure presented in this thesis. Dynamic dissemination of recommendations as a visitor tours a Cultural Heritage site would entail the additional engagement of a mobile device and the ability to track the visitor. Ultimately the integration with recommender systems is a potential utilisation of the semantic relatedness paradigm with the benefit that RACO is able to utilise myriad connections between entities with differing conceptual structure.

Another advantage is provided with the application of relationship extraction
in systems that utilise recommendation transparency. Explicit description of why an exhibit has been placed in a visitor’s tour, or has been recommended based on previous exhibits seen may facilitate the visitor’s trust and confidence in the system. The linking together of exhibits with a sequential description can potentially be used to aid the visitor’s understanding of the collection as a whole. Using this property, the construction of museum tours focussed on a visitor’s interest model could be used to incorporate explicit explanations as to how each exhibit in the tour refers to the visitor’s interest. The use of extracted relationships using the methods in this thesis can be used in further research on the ability of explanation of exhibit connections to facilitate learning and uptake of information as previously described in Chapter 2.

An application of the research presented in this thesis, touched on only briefly in Chapter 6, is in the construction of a procedural narrative. Specifically the ability to identify the points of relatedness of a group of entities and the ability to describe the logical progression from one of those entities to the next. In the context of the Cultural Heritage domain, this would entail the explanation of how one museum exhibit relates to another and how that exhibit in turn relates to a third, and so on. Throughout this trail of exhibits, reference to the central connecting theme of the group of exhibits would provide a central touchstone for the collection of exhibits contained in the visitor’s tour, potentially allowing the visitor to understand the interconnectedness of a collection of exhibits as a whole. Ultimately, such an implementation of the methods constructed in this thesis would need to be trialled to determine the validity of this hypothesis, but it is a research direction that is able to combine the two primary elements of this thesis in a manner which, as I am aware, has not been explored previously.
7.3 Summary

In summary, I have presented an approach to determining the semantic relatedness of contextually diverse heterogeneous entities. The use of a diversely structured collaborative encyclopedia, Wikipedia, allows for comparison of diversely structured entities, and I have used a contextual categorisation approach to calculating the relatedness of entities. I have presented a collection of methods for extracting semantic relationships between diversely structured entities by utilising the taxonomic structure of Wikipedia. This method resolves the issue of non-standard descriptions of conceptual links between articles by utilising the semantic network that organises these articles. Finally I have presented a discussion of how the methods constructed in the course of this thesis could be applicable to domains other than the domain of Cultural Heritage used throughout this thesis as a resource for diverse entities and evaluation.
Appendix A

Museum Exhibit Definitions

This appendix contains a list of the museum exhibits used in experimentation in Chapter 4. The exhibit codes correspond to the exhibit codes used in Figure 4.2 and Figure 4.3. Also contained are the titles of the exhibits used in experimentation, and the galleries of which they were members at the time of data collection. Note that some exhibits did not occur as part of galleries and were placed around the museum as “island” exhibits; these exhibits are labelled with “NA” in the gallery column. A second table contains the Wikipedia articles to which exhibits were aligned in order to represent their content. The “Aligned Wikipedia Article” is the suffix of the Wikipedia article’s url (and hence the title of the Wikipedia article), for example, the Phar Lap article is represented by the Wikipedia article at the url http://en.wikipedia.org/wiki/Phar_Lap.
<table>
<thead>
<tr>
<th>Exhibit Code</th>
<th>Exhibit Title</th>
<th>Gallery</th>
</tr>
</thead>
<tbody>
<tr>
<td>exphar</td>
<td>Phar Lap</td>
<td>Melbourne Gallery</td>
</tr>
<tr>
<td>excsir</td>
<td>CSIRAC</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>examge</td>
<td>Amethyst Geode</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exgisq</td>
<td>Giant Squid</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exanth</td>
<td>Ant Colony</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exmine</td>
<td>Quartz Reef Gold Mine Model</td>
<td>NA</td>
</tr>
<tr>
<td>exgold</td>
<td>Gold Nuggets</td>
<td>NA</td>
</tr>
<tr>
<td>exmout</td>
<td>Insect Mouthpart models</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>extara</td>
<td>Tarantula</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exmicr</td>
<td>Electron Microscope</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>exlava</td>
<td>Hardened Lava Flow</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exbtlta</td>
<td>Model of Beetle Anatomy</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exdipr</td>
<td>Diprotodon Fossil</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>exhadr</td>
<td>Hadrosaur in-ground Fossil</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>ex3prt</td>
<td>3-part sauropod bone</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>extril</td>
<td>Small Triolbite fossil</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>edxino</td>
<td>Large Dinosaur Skeletons</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>exgori</td>
<td>Gorilla Diorama</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>exbird</td>
<td>Bird Family Tree (songbirds)</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>extang</td>
<td>Tanganyikan Aquarium</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>exgene</td>
<td>The Human Genome</td>
<td>Evolution Gallery</td>
</tr>
<tr>
<td>exhumm</td>
<td>Hummingbirds</td>
<td>NA</td>
</tr>
<tr>
<td>excora</td>
<td>Coral</td>
<td>Science &amp; Life Gallery</td>
</tr>
<tr>
<td>exanat</td>
<td>Human Anatomy model</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>exskel</td>
<td>Human Skeleton</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>extjeb</td>
<td>Mummy and Sarcophagus</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>excall</td>
<td>Steel Callipers</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>exfrze</td>
<td>Greek Frieze</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>exgurn</td>
<td>Greek Urn</td>
<td>Mind &amp; Body Gallery</td>
</tr>
<tr>
<td>exarma</td>
<td>Giant Armadillo</td>
<td>NA</td>
</tr>
<tr>
<td>exnest</td>
<td>Bowerbird Nest</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>exfern</td>
<td>Tree Ferns</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>exgond</td>
<td>The Break-up of Gondwanaland</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>excrek</td>
<td>Bush Creek</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>excrys</td>
<td>Crystal Cavity</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>exmidd</td>
<td>Beach Midden Cutaway</td>
<td>Kalaya Meeting Place</td>
</tr>
<tr>
<td>exblkt</td>
<td>Blackened Trees</td>
<td>Forest Gallery</td>
</tr>
<tr>
<td>exsail</td>
<td>Canoe Sail</td>
<td>Te Pasifika Gallery</td>
</tr>
<tr>
<td>extree</td>
<td>Tree section</td>
<td>NA</td>
</tr>
<tr>
<td>exguns</td>
<td>Indian Musket and Blunderbuss</td>
<td>NA</td>
</tr>
<tr>
<td>exmeda</td>
<td>Medals and Awards</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table A.1: The exhibits codes, exhibit names, and respective gallery memberships of exhibit entities used in experiments in this thesis.
### Table A.2: The Wikipedia article titles used to represent exhibits used in experiments in this thesis. Exhibit codes correspond to the definitions in Table A.1. The Wikipedia article titles correspond to the article titles used in the September 9th, 2009 Wikipedia dump (20090909).

<table>
<thead>
<tr>
<th>Exhibit Code</th>
<th>Aligned Wikipedia Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>exphar</td>
<td>Phar_Lap</td>
</tr>
<tr>
<td>excsir</td>
<td>CSIRAC</td>
</tr>
<tr>
<td>examge</td>
<td>Geode</td>
</tr>
<tr>
<td>exgisq</td>
<td>Giant_squid</td>
</tr>
<tr>
<td>examth</td>
<td>Ant_colony</td>
</tr>
<tr>
<td>examine</td>
<td>Quartz_reef_mining</td>
</tr>
<tr>
<td>exgold</td>
<td>Gold_nugget</td>
</tr>
<tr>
<td>examout</td>
<td>Mouthparts</td>
</tr>
<tr>
<td>extara</td>
<td>Tarantula</td>
</tr>
<tr>
<td>examicr</td>
<td>Electron_microscope</td>
</tr>
<tr>
<td>exlava</td>
<td>Lava</td>
</tr>
<tr>
<td>exbtlb</td>
<td>Beetle</td>
</tr>
<tr>
<td>exdipr</td>
<td>Diprotodon</td>
</tr>
<tr>
<td>exhadr</td>
<td>Hadrosaurid</td>
</tr>
<tr>
<td>ex3prt</td>
<td>Sauropoda</td>
</tr>
<tr>
<td>extril</td>
<td>Trilobite</td>
</tr>
<tr>
<td>edxino</td>
<td>Dinosaur</td>
</tr>
<tr>
<td>exgori</td>
<td>Gorilla</td>
</tr>
<tr>
<td>exbird</td>
<td>Petroicidae</td>
</tr>
<tr>
<td>extang</td>
<td>Lake_Tanganyika</td>
</tr>
<tr>
<td>exgene</td>
<td>Human_genome</td>
</tr>
<tr>
<td>exhumm</td>
<td>Hummingbird</td>
</tr>
<tr>
<td>excora</td>
<td>Coral</td>
</tr>
<tr>
<td>exanat</td>
<td>Human_anatomy</td>
</tr>
<tr>
<td>exskel</td>
<td>Human_skeleton</td>
</tr>
<tr>
<td>extjeb</td>
<td>Mummy</td>
</tr>
<tr>
<td>excall</td>
<td>Caliper</td>
</tr>
<tr>
<td>exfrze</td>
<td>Frieze</td>
</tr>
<tr>
<td>exgurn</td>
<td>Urn</td>
</tr>
<tr>
<td>exarma</td>
<td>Glyptodon</td>
</tr>
<tr>
<td>exnest</td>
<td>Bowerbird</td>
</tr>
<tr>
<td>exfern</td>
<td>Cyatheales</td>
</tr>
<tr>
<td>exgond</td>
<td>Gondwana</td>
</tr>
<tr>
<td>excrek</td>
<td>Stream</td>
</tr>
<tr>
<td>excrys</td>
<td>Crystal</td>
</tr>
<tr>
<td>exmidd</td>
<td>Midden</td>
</tr>
<tr>
<td>exblkt</td>
<td>Bushfire</td>
</tr>
<tr>
<td>exsail</td>
<td>Sail</td>
</tr>
<tr>
<td>extree</td>
<td>Dendroclimatology</td>
</tr>
<tr>
<td>exguns</td>
<td>Blunderbuss</td>
</tr>
<tr>
<td>exmeda</td>
<td>Medal</td>
</tr>
</tbody>
</table>
Appendix B

Annotation Interfaces

This Appendix provides the details of the surveys and annotation interfaces used to collect user annotations and data evaluation. For each interface, the instructions given to users, a description of the experiment and an image of the layout is given.

B.1 Museum Exhibit Relatedness Survey

The museum visitor survey was designed to collect relatedness judgements between museum exhibits at Melbourne Museum. Full details of the deployment and procedure used to gather the ratings are given in Section 4.2. The survey began by asking the museum member to answer a series of questions based on their museum visiting habits, and the recency of their last visit. The profiling questions given are shown in Figure B.1.

Following these questions, the user was presented with the set of instructions given in Figure B.2.

After these instructions, the user was presented with 15 exhibit pairs using the layout presented in Figure 4.1.
Figure B.1: The set of profiling questions given to museum members on starting the exhibit relatedness survey. These were used for used in user modelling experiments approached seperately to this thesis.

The rest of this survey consists of 15 questions, all of the same form: 
You will be shown two exhibits from Melbourne Museum, and asked if you recall seeing these exhibits.
You will then be asked to score the degree to which you believe the exhibits are related, on a scale of zero to four.
We encourage you to provide a brief indication of the conceptual basis for your rating in an optional field on each page. The reason for the exhibits being related is entirely up to you.

Figure B.2: Instructions given to museum members taking part in the exhibit relatedness survey.
B.2 Relationship Evaluation Pilot Annotation 2

Due to unforeseen circumstances, the design of the first iteration of the pilot relationship evaluation task was overwritten and replaced with the second iteration of the interface. The first iteration of this pilot interface is described in Section 5.3.1, but there are no images of the layout of this iteration of the interface. The second iteration of the annotation interface incorporated feedback from annotators that performed the task using the first pilot annotation interface. The major criticism provided by annotators was that the criteria of \textit{validity}, \textit{clarity} and \textit{specificity} did not apply to all relationships presented, and thus did not make sense in all cases. The annotators’ suggested solution was to increase the number of criteria to make the choice of relationship classification broader (rather than restricted to solely \textit{clarity} and \textit{specificity} as in the first interface). Annotators were presented with the following set of instructions before commencing the evaluation task. The presentation format used to display the relationships and the style of question used is presented in Figure 5.4. The data used for this annotation exercise was comprised of 5 article pairs, each of which contained 5 relationships, for a total of 25 relationships to be evaluated by each annotator. The details of the deployment of both iterations of the pilot annotation interface are given in Section 5.3.1.

B.3 Mechanical Turk Relationship Evaluation 1

Each Mechanical Turk Human Intelligence Task (HIT) is treated as a single task, a user is not obliged to complete more than one (unless stipulated by the instructions). Each HIT in a series must be of the same number of questions, and possess the same format. The form of all HITs in the first iteration of the Mechanical Turk relationship
Relatedness Annotation Task.

Welcome to this sentence annotation task, thank you for choosing to take part. The purpose of this task is to grade a number of textual explanations for Wikipedia article relatedness.

The task consists of a series of Wikipedia article pairs. Each article pair will be described briefly at the top of the page, followed by a list of reasons that express potential relationships between these two articles.

For each of the reasons presented, you will be asked if this reason describes a valid relationship between the two articles (in the form of a yes/no radio button choice). For each reason, you will also be asked to select adjectives that describe the relationship’s qualities. For example, if the relationship given is a broad (e.g., ”Both are living”), or specific one (e.g., ”X can be used as a replacement for Y when building motorcycle engines”), or if the relationship is overly obvious or obscure.

Broad-Specific and Obvious-Obscure are presented as scales with a median option. Select this option if you think that the relationship contains an appropriate level of specificity or obscurity (i.e., the relationship is ”just right”, or not at the extreme ends of the scale).

If you believe that a label does not apply to the relationship at all, then select the NA option. Optionally, you may also label a relationship as ’interesting’ or ’tenuous’.

Please enter all relevant labels for a relationship whether the relationship is valid or not.

This task consists of 6 article pairs.

Figure B.3: Instructions given to University of Melbourne postgraduates performing the second annotation task.

evaluation task took the form of the purpose of the task (instructions), a pair of Wikipedia article blurbs, and three relationships to be evaluated. The instructions in Figure B.4 were presented at the top of every HIT. An example of an instance of a HIT in this version of the interface can be seen in Figure 5.5. The details of the deployment of this annotation interface are given in Section 5.3.2.
Appendix B: Annotation Interfaces

Instructions and guidelines:
The two blurbs below describe articles that exist in Wikipedia. The blurbs present a broad overview of the content of the article. Your task is to decide whether or not the reason presented below possesses a relationship between the two topics represented by these articles.

Figure B.4: The Instructions given for the first iteration of the Mechanical Turk Relationship Evaluation Task.

Instructions and guidelines:
In this task you will evaluate pieces of text ("reasons") that describe a potential relationship between two Wikipedia articles.

- The two blurbs below are snippets of Wikipedia articles.
- Read and ensure that you understand the subject the blurb is describing before proceeding with rating the reasons.
- Decide if the relationship presented in each reason describes a valid relationship between the articles.
- Make sure you select a response to each question; if a question is not answered, the HIT will be rejected.

Figure B.5: The Instructions given for the second iteration of the Mechanical Turk Relationship Evaluation Task.

B.4 Mechanical Turk Relationship Evaluation 2

Again, each HIT is treated as an individual task, the following instructions were presented on every set of relationships presented to the Turker. Note that there are more instructions than the previous iteration interface. This explicit statement was made to eliminate any confusion experienced by Turkers, and to clearly state the circumstances in which their work would or would not be accepted. An example of the format in which the article blurbs and relationship evaluation questions appeared
Figure B.6: The presentation used to present relationships to Turkers in the second Mechanical Turk annotation interface. Two article blurbs appear at the top of the form, followed by 5 potential relationships between the two articles described by these blurbs.

is given in Figure B.6. The details of the deployment of this annotation experiment are given in Section 5.3.2.

B.5 Museum Staff Relationship Evaluation

Staff at Melbourne Museum were emailed out a request to take part in the evaluation task. If they chose to take part, they were taken to a web page designed to closely reflect the format of evaluation used in the second iteration of the Mechanical Turk interface: 5 article pairs presented on individual web pages, containing 5 potential
**Relationship Evaluation Task** Welcome, and thank you for taking part in this evaluation experiment, which relates to Kubadji project between The University of Melbourne/Monash University and Melbourne Museum http://hum.csse.unimelb.edu.au/kubadji.

The purpose of this task is to determine the validity of relationships between items on display (or formerly on display) at Melbourne Museum. In this, we represent each item by its most closely-matching Wikipedia article: for example, the Amethyst Geode in the Rocks and Minerals exhibit is represented by the Wikipedia article “Geode.”

The evaluation task contains a series of 5 item pairs (e.g., Geode and Crystal), which are presented as two two-sentence blurbs describing the content of both items. The two blurbs are followed by 5 potential relationships between the two items. Your task is to decide if each relationship is a valid relationship between the two articles.

This task is designed to be brief, and evaluating the 5 sets of relationships should take less than 10 minutes to complete.

Figure B.7: Instructions given to Melbourne Museum staff members performing the relationship evaluation task.

The layout of the relationships and the article blurbs, and the format of question is identical to that used in Figure B.6. For details on the distribution and results of this evaluation task, see Section 5.3.3.
Bibliography


Bennett, James, and Stan Lanning. 2007. The Netflix Prize. In Proceedings of the KDD Cup and Workshop, 3–6, San Jose, United States.


Joint Conference on Artificial Intelligence Workshop on Knowledge Reasoning for Answering Questions, 61–70, Edinburgh, Scotland.


Author/s: GRIESER, KARL

Title: Computing relationships and relatedness between contextually diverse entities

Date: 2011


Persistent Link: http://hdl.handle.net/11343/37066

File Description: Computing relationships and relatedness between contextually diverse entities

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.