University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Hamzei, E;Li, H;Vasardani, M;Baldwin, T;Winter, S;Tomko, M

Title:
Place questions and human-generated answers: A data analysis approach

Date:
2020-01-01

Citation:
Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S. & Tomko, M. (2020). Place questions and human-generated answers: A data analysis approach. Kyriakidis, P (Ed.) Hadjimitsis, D (Ed.) Skarlatos, D (Ed.) Mansourian, A (Ed.) Geospatial Technologies for Local and Regional Development. Proceedings of the 22nd AGILE Conference on Geographic Information Science, pp.3-19. Springer, Cham. https://doi.org/10.1007/978-3-030-14745-7_1.

Persistent Link:
https://hdl.handle.net/11343/228874

Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S., & Tomko, M. (2019). Place questions and human-generated answers: A data analysis approach. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Mansourian (Eds.), Geospatial Technologies for Local and Regional Development. Proceedings of the 22nd AGILE Conference on Geographic Information Science: Springer. pp 3-19, doi:10.1007/978-3-030-14745-7_1.

Place questions and human-generated answers: A data analysis approach

Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko

Abstract This paper investigates place-related questions submitted to search systems and their human-generated answers. Place-based search is motivated by the need to identify places matching some criteria, to identify them in space or relative to other places, or to characterize the qualities of such places. Human place-related questions have thus far been insufficiently studied and differ strongly from typical keyword queries. They thus challenge today's search engines providing only rudimentary geographic information retrieval support. We undertake an analysis of the patterns in place-based questions using a large-scale dataset of questions/answers, MS MARCO V2.1. The results of this study reveal patterns that can inform the design of conversational search systems and in-situ assistance systems, such as autonomous vehicles.

1 Introduction

In their everyday communication people frequently ask questions about places (Winter & Freksa 2012). This is reflected in the frequency of location-based queries in human-computer interaction, e.g., Web search (Sanderson & Kohler 2004) and question answering systems (Li & Roth 2006). The popularity of conversational bots and assistants (Radlinski & Craswell 2017) requires a shift from retrieving documents as response to keyword-based queries to natural answers to natural language questions.

A better understanding of the nuances expressed in search questions will enable better inference of the intent behind the query, and thus improve the delivery of taylored information in the answer. Due to polysemy of *place* references, their strong contextual dependence (Winter & Freksa 2012), diverse metaphorical uses (Agnew 2011), and complex, often vernacular references (Hollenstein & Purves 2010), the nuanced understanding and answering of place questions are still challenging for

Ehsan Hamzei, e-mail: ehamzei@student.unimelb.edu.au
The University of Melbourne, Parkville, Victoria, Australia

computer-based systems (Sui & Goodchild 2011). Place questions not only concern location (*where-questions*), but also encompass a wide range of informational needs about places from their types and affordances to their qualities. A thorough analysis of questions and answer sets is still missing, yet it is an essential prerequisite for future improvement of question interpretation and answer generation mechanisms.

In this paper, we investigate the structural patterns of place-related questions and their human-generated answers using the MS MARCO V2.1 dataset (Nguyen et al. 2016). By addressing our main research question: *"How place questions and their answers can be modelled based on place-related semantics?"*, we contribute:

- An analysis of the content of place-related questions and answers, by extracting patterns of place-related information through semantic encoding;
- A categorization of place-related questions and human-generated answers of the dataset based on semantic encoding and contextual semantic embedding;
- An investigation of the relations between the categories of questions and corresponding answers.

2 Literature review

Research in geographic information retrieval from Web sources has focused primarily on geographic Web queries (Sanderson & Kohler 2004, Aloteibi & Sanderson 2014) and the geographic context of Web search (Backstrom et al. 2008). Geographical Web queries with *explicit* location (i.e., containing place names – *toponyms*) were analysed first by Sanderson & Kohler (2004). Later, implicit geographic queries relating to generic nouns designating locations known to a subgroup of people (e.g., a family sharing a location known as *home*) have also been identified (Wang et al. 2005).

In contrast, Jones et al. (2008) analyzed the distances between toponyms in the query and the location of the user indicated by their IP address to study the intent of the spatial search. They classified the distances into granular categories: same-city, same-state, same-country and different-country. Recently, Lee et al. (2015) used more accurate user positioning for a finer-grained analysis of search distances and correlated their distribution with the distribution of entities of different types in space. Yet, Web use patterns have since shifted dramatically towards mobile Web use (Lee et al. 2015), voice queries (Guy 2018), and conversational assistants (Radlinski & Craswell 2017). Voice queries are more similar to natural language questions than keyword queries (Guy 2018), and the transition from queries to questions and from result sets with links to automatically generated answers leads to a more natural interaction between humans and machines.

Place questions vary in focus, targeting a range of aspects of place (Wang et al. 2016). While two types of questions, *'where is'*, and *'how can I get to'* are most common (Spink & Gunar 2001), people also ask questions to get information *about* place names, place types, and activities supported by the places themselves. While in human-human communications place questions are asked with different intent in

mind such as gaining factual knowledge, recommendations, invitations, offers, and opinions (Wang et al. 2016), studies show that in human-computer interaction questions are mostly asked for informational and navigational purposes (Hollink 2008). 'Where is' questions aim at locating places in space, and thus understanding their spatial context. In a fundamental study, Shanon (1983) investigated how people answer such 'where' questions by extending the *room theory* to model human answering of 'where' questions. 'How can I get to' questions are a second dominant type of place questions, seeking procedural instructions to reach an intended destination (Tomko & Winter 2009). Answers to 'where' and 'how can I get to' questions carry descriptions of the environment in form of place or route descriptions, respectively (Winter et al. 2018). In both questions, the intended places can be referred to and described through place names, types, functions, and affordances. Yet, a detailed analysis of structural similarities and the patterns between the types of place questions and their answers has not been conducted thus far.

3 Data

We base our research on MS MARCO V2.1 (Nguyen et al. 2016). The dataset is a representative collection of questions (incl. place-related questions) and their human-generated answers. The current version of MS MARCO V2.1 contains more than one million records[1] of search question to Microsoft Bing, including human-generated answers, retrieved documents, and question types. The dataset identifies five types of questions – (1) LOCATION, (2) NUMERIC, (3) PERSON, (4) DESCRIPTION, and (5) ENTITY (Nguyen et al. 2016), but we focus exclusively on questions and answers which are labeled as LOCATION.

The original dataset has been divided into *train*, *dev*, and *test* sets. Here we use the *train* and *dev* subsets, with in total 56721 place-related questions. In MS MARCO V2.1,[2] the relation between questions and their answers is not necessarily one-to-one (Nguyen et al. 2016). Thus, we find questions with no answer, or more than one answer. Table 1 shows the number of answers per place question in the combined dataset, indicating that around 22% of questions have no answer and about 2% have multiple answers (i.e., due to question ambiguity or insufficient information in the retrieved documents).

4 Method

We first translate place questions/answers into semantic encodings. Next, the questions and answers are clustered to identify common patterns in the data, based on

[1] http://www.msmarco.org

[2] https://github.com/dfcf93/MSMARCOV2/blob/master/README.md

Table 1: Number of answers per location question in the combined dataset

Number of answers (per question)	Count
0	12486
1	42884
2	1345
3	4
4	1
5	1

semantic encodings and contrasted to a second approach based on embeddings. Finally, the relations between the questions and their answers are explored by linking the clusters of the questions and answers. In the following sections, we detail the method for generating the semantic encodings, categorizing the questions and answers, and analyzing the relation between questions and their answers[3].

4.1 Semantic encoding

We propose a semantic encoding schema extending that by Edwardes & Purves (2007). They introduced a mapping from part-of-speech to place semantics to extract elements, qualities, and affordances from generic nouns, adjectives, and verbs, respectively. We extend this semantic representation for the relatively short place questions and answers from MS MARCO, with the following primary elements of semantic encoding: (1) PLACE NAMES (e.g., *MIT*); (2) PLACE TYPES (e.g., *university*); (3) ACTIVITIES (e.g., *to study*); (4) SITUATIONS (e.g., *to live*); (5) QUALITATIVE SPATIAL RELATIONSHIPS (e.g., *near*); and (6) QUALITIES (e.g., *beautiful*). To differentiate types of questions, the WH-WORDS, and other generic OBJECTS are also considered.

Table 2 shows the resulting alphanumeric encoding schema. For example, after the removal of stop words, the question *what is the sunniest place in South Carolina* is translated into the encoding `2qtrn`. This semantic encoding enables to analyze and categorize a large dataset of questions and answers by their structural patterns.

4.2 Information extraction

To extract further place-related semantics and linguistic information from the questions and their answers, we apply a sequence of preprocessing steps, based in part on the Stanford CoreNLP toolkit (Manning et al. 2014).

1. **Tokenization, tagging and dependency parsing:** First, the text is tokenized, and abbreviations are expanded into their canonical forms by using a common place

[3] The implementation is availabe at: `https://github.com/haonan-li/place-qa-AGILE19`

Table 2: Semantic representation encoding

Semantic Type	Part-of-speech	Code	Semantic Type	Part-of-speech	Code
where	WH-word	1	Place name	noun	n
what	WH-word	2	Place type	noun	t
which	WH-word	3	Object	noun	o
when	WH-word	4	Quality	adjective	q
how	WH-word	5	Activity	verb	a
whom	WH-word	6	Situation, and event	verb	s
whose	WH-word	7	Spatial relationship	preposition	r
why	WH-word	8			

name abbreviation table. Next, a part-of-speech tagger and dependency parser are applied to the text.

2. **Noun encoding:** Nouns are encoded in the order of place names (toponyms), place types, and objects. First, all subsequences of a sentence are considered candidate toponyms. These candidates are then matched with the GeoNames gazetteer[4] to extract toponyms. We preference compound place names as better matches than simplex place names. Thus, *North Melbourne* is a compound place name and not an adjective followed by a place name. Next, nouns that are not place names have been considered as candidates for place type and objects. Place types are identified using dictionary lookup, and all nouns that are neither place names, nor place types are encoded as objects. The dictionary of generic place types is constructed by crawling the tag values of the OpenStreetMap (OSM) spatial database,[5] due to the rich diversity of place types that are sourced from a large, global community of active volunteers.
3. **Verb encoding:** We have focused on verbs related to activities or situations (states) and ignored other types of verbs, known as accomplishments and developments (Mourelatos 1978). To differentiate between activities and situations, two sets of dynamic verbs (action and stative verbs) are collected and integrated from online resources [6][7]. Then, a pre-trained contextualized word embedding model ELMo (Peters et al. 2018), trained from large scale bidirectional language models, was used to extract activities and situations by considering contextual information. To compare the semantic similarity between the verbs in sentences with the verbs in the aforementioned verb sets, we first generate an embedding vector for each verb in the verb set. Then, for the extracted verbs from the questions and answers, their embedding vector representations are compared with the vectors in the two sets. Based on the computed Euclidean distances between the vector of the verb

[4] www.geonames.org

[5] https://www.openstreetmap.org

[6] https://www.gingersoftware.com/content/grammar-rules/verbs

[7] https://www.perfect-english-grammar.com/support-files/stative-verbs-list.pdf

in a sentence with the vectors of the verbs in the two sets, the extracted verbs are classified as activities or situations.

4. **Preposition and adjective encoding:** Dependency parsing has been used to encode prepositions and adjectives. For the universal dependencies (De Marneffe et al. 2014) of each sentence, `case` and adjective modifier (`amod`) dependencies for prepositions and adjectives are extracted separately. Prepositions anchored to a place name are encoded as spatial relationships, and adjectives modifying a place type or a place name are considered as qualities of places.

4.3 Question/Answer analysis

After extracting the desired place-related semantics and linguistic information, the questions and answers are translated into semantic encodings. To find the categories of place-related questions and answers, we compute clusters of the semantic representations. In this clustering, we first randomly select 1024 different encodings from the unique set of all semantic encodings, and subsequently re-model the semantic encodings of the questions/answers as 1024-dimensional vectors. The values in these vectors are calculated based on the Jaro similarity (Jaro 1989) between the semantic encodings of the questions/answers and the selected encodings. The Jaro similarity sim_j of two strings s_1, s_2 is given in Eq.1:

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{1}$$

where $|s_i|$ is the length of string s_i, m is the number of matching characters, and t is the number of transpositions. Specifically, two characters from two different strings are considered to match if they are the same and not farther than $\left\lfloor \frac{max(|s_1|,|s_2|)}{2} \right\rfloor - 1$. The number of matching characters (but different sequence order) divided by 2, defines the number of transpositions.

Next, k-means clustering is applied to the questions and answers, separately. To measure whether the semantic representations retain the contextual similarity of the sentences, the ELMo representations for the questions and answers are used with the same similarity measure and clustering technique to provide a second set of clusters. The similarity of the clustering based on the semantic encodings and ELMo-based clusters is then evaluated. The results of clustering are also human-interpreted using the most frequent encodings in each cluster, enabling to derive the categories of place questions and answers.

In addition to categories of the questions and answers, we analyze their contents based on the frequently extracted place-related semantics. We also investigate the geographical distribution of the identified places using GeoNames to provide a deeper understanding of the dataset. In order to disambiguate between candidate toponyms, we combine the place names from the question, with those in the corresponding answers. The set of geospatial groundings that is associated with the minimum total

distance between the toponyms is selected. In the case of a single toponym, this method cannot be applied. Consequently, only the resolved toponyms are used for describing the geographical distributions of the dataset. Finally, the relation between the questions with unresolved toponyms and unanswered questions is investigated.

4.4 Question/Answer relationship

We link categories of questions to categories of answers to investigate the relationship between the content of place questions and their human-generated answers. As a category of place questions can be answered using one or more categories of answers (and vice-versa), we consider generic many-to-many relationships.

The same approach to categorization of questions and answers is also applied to concatenated question-answers. Finally, the result of linking categories of questions and categories of answers is compared with concatenated categories. For questions with multiple answers, multiple concatenated encodings are generated and investigated, while questions with no answers are concatenated with a unique pattern (oo).

5 Results

The results reveal three major groups of question/answers, replicated using both the semantic encoding and embedding approaches. We provide an initial qualitative interpretation of the patterns based on the encoding approach.

5.1 Preliminary analysis

Place questions are constructed with a small number of tokens, and their answers are mostly short descriptions. Using tokenization and sentence segmentation, we find that 95.71% of questions contain less than ten tokens, and 98.17% of their answers are formulated with one, two, or three sentences.

5.1.1 Extracted place-related semantics

Table 3 shows the top-five most frequent values extracted from the questions and answers for each type of place-related semantics. The dataset manifests a geographical bias to places in the USA. We identify similar patterns in geographical scales of place types and place names in the questions and their answers – i.e. related to coarse geographic scales, such as countries, and states. The frequency of activities

Table 3: Top-five frequent place-related semantics extracted from the dataset. Frequency in brackets.

Type	**in Questions**	**in Answers**	**Type**	**in Questions**	**in Answers**
Place name	California (1393)	United States (4845)	Type	County (11702)	City (1714)
	Texas (1391)	California (1482)		State (2291)	State (1653)
	Florida (1148)	Texas (989)		City (1630)	County (1438)
	New York (895)	Florida (961)		Zone (745)	Area (882)
	Illinois (692)	New York (894)		Region (653)	Region (758)
Activity	Buy (340)	Go (64)	Situation	Find (1412)	Find (695)
	Go (296)	Run (62)		Live (746)	Have (405)
	Play (120)	Leave (55)		Have (662)	Live (305)
	Build (88)	Build (53)		Grow (321)	Include (231)
	See (86)	Move (38)		Originate (237)	Originate (125)
Spatial relation	In (3916)	In (10851)	Quality	Largest (242)	Largest (121)
	Near (153)	On (379)		Biggest (106)	Census-designated (68)
	At (142)	At (362)		Highest (97)	Metropolitan (54)
	On (109)	Near (275)		Expensive (56)	Small (46)
	Between (38)	Between (251)		Beautiful	Coastal (36)

and situations in the questions is notably higher than in the answers. In other words, people use activities and situations as criteria to describe the intended places in the questions rather than asking about activities and situations happening in a place –i.e., *where is the place with particular situation/affordance* is more often asked than *what is the affordance/situation of a specific place*. Thus, a set of detailed characteristics of a place may be specified in the question, leading to a simple answer with nominal references. Table 3 shows large differences between the frequencies of spatial relationships extracted from questions, and those extracted from answers. The reason is the use of spatial relationships for localization of places. Finally, the qualities included in questions and answers differ. In questions, qualities are used as criteria for identifying or describing the intended places, with most of the values being superlative adjectives. However, qualities in answers are mostly used to provide additional information about intended places or to describe particular places using combinations of quality and type – e.g., coastal region, or metropolitan area.

5.1.2 Toponym resolution and geographical distribution

Tables 4 and 5 show that 79.2% and 68.1% of extracted places and question/answers records can be disambiguated, respectively. Figure 1 illustrates the geographical distribution of resolved places in the questions and answers. The spatial bias in the

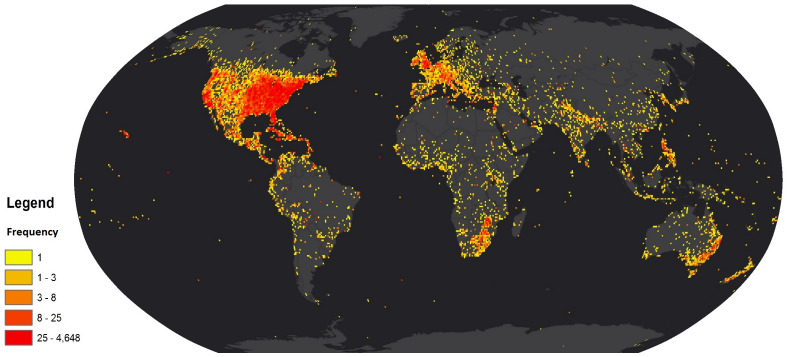
Table 4: Statistics of extraction, and disambiguation process

	Number of Places	Percentage
All	305868	100%
Possible to resolve	242375	79.2%
Ambiguous	63493	20.8%

Table 5: Contingency table of ambiguous/resolved, and unanswered/answered questions

	Answered	Unanswered	Sum (percentage)
Resolved	35184	3444	38628 (68.1%)
Unresolved	9051	9042	18093 (31.9%)
Sum (percentage)	44.235 (78%)	12486 (22%)	56721 (100%)

Fig. 1: Geographical distribution of resolved toponyms in questions and answers



geographical distribution of places is visible. This spatial bias can be an artefact of the monolingual English dataset, population distribution of users of the source search engine (Microsoft Bing), or unknown question sampling from the overall Bing logs used to generate this dataset.

The comparison of resolved/unresolved and answered/unanswered records shows a strong relationship between ambiguous questions and unanswered questions. Table 5 shows the contingency table of this relation, revealing that 72.4% of unanswered questions are ambiguous (9042 records out of 12486 unanswered questions), and 50.0% of ambiguous questions are not answered (9042 records out of 18093 ambiguous question/answer records). Hence, while people can interpret ambiguous toponyms considering other factors, such as saliency of places, and they can answer slightly more than half of the ambiguous questions, spatial ambiguity is one of the major reasons for unanswered questions. These questions may still be interpretable in context, yet this context is lacking in the MS MARCO dataset.

Table 6: Frequent encoding patterns in questions

Pattern	Percentage	Example
2tnn	12.5%	*What is the county for Grand Forks North Dakota*
1n	8.6%	*Where are the Boise Mountains*
1nn	7.8%	*Where is Barton County Kansas*
1o	5.7%	*Where are ores located*
2tn	5.5%	*What is county for Seattle*

5.2 Results of question and answer analysis

We now investigate the structural patterns in place questions and answers separately, using the extracted place-related and linguistic semantic encoding. We identify distinct types of place question/answer pairs using a clustering approach based on the semantic representations with manual interpretation.

5.2.1 Frequent patterns

Tables 6 and 7 present the top-five frequently observed semantic representation patterns. More than 40% of questions and answers can be described using the top-five representations. The top-five encodings for questions show that most of the spatial relationships are implicitly provided, e.g., *Barton County, Kansas* instead of *Barton County in Kansas*. In addition, the top-five questions are semantically different, first in terms of *What* and *Where* questions, and second in the manner the intended places are described. In questions with the semantic encoding `2tnn` and `2tn` the intended places are defined by type and spatial criteria, while questions with encoding `1n` and `1nn` use toponyms as verbal references to intended places. Finally, pattern `1o`, which is related to implicit situations (e.g., *where are orangutans* instead of *where do orangutans live*), or to non-geographical places (e.g., *where is the key button*), is frequently observed.

The most frequent patterns in answers provide evidence that people answer questions concisely. A single place name, which can be a compound or simple noun, is the answer for more than one quarter of the questions. We note that this may be an artifact of the interface used by people to answer the questions (presumably, keyboard). In addition, implicit relationships between places (e.g., `nn` as for *Tigard Oregon*) are also more frequent in answers than explicit spatial relationships. Interestingly, one of the frequent patterns of answers is related to non-geographical places, such as *Hebrew* as an answer for a *where* question about languages.

Table 7: Frequent encoding patterns in answers

Pattern	Percentage	Example
n	26.2%	*Germany* (question: *What country were gummy bears originally made in*)
o	6.2%	*Hebrew* (question: *Where did letter j originate*)
nn	4.2%	*Warren County, United States* (question: *Where is Lacona Iowa*)
rn	2.3%	*In Worcester County* (question: *What county Fitchburg Massachusetts*)
nrnn	2.3%	*Penhook is in Franklin County Virginia* (question: *What county is Penhook Virginia in*)

5.2.2 Categorizations of place-related questions/answers

To identify distinct types of place questions and answers, we used a clustering approach with manual interpretation based on the semantic representations of the results. Two different clustering techniques have been used for finding clusters of questions and answers based on word embeddings, and semantic encodings. We used the Calinski-Harabasz score (Caliński & Harabasz 1974) to find the optimum number of clusters for questions and answers, respectively. The score has been computed for all clusterings in the range of 2-30 clusters. For both clustering approaches and for both questions and answers, the score is optimal for three clusters. Importantly, we find that the results of clustering using semantic encoding and word embedding are highly similar, with a one-to-one relationship between the clusters with a 71.2% to 87.8% similarity using the Dice coefficient. To avoid presenting redundant information, here we report only the results of clustering based on semantic encoding which enables easier interpretability. The results of the clustering for questions and answers are interpreted manually using the most frequent semantic encodings in each cluster.

Tables 8 and 9 present the categories of questions/answers, and their overall percentage. We find that questions differ in terms of intention and formulation. While the first two types, non-spatial and spatial questions, relate to geographical places, the third type of questions is related to non-geographical entities such as virtual places (e.g., *Marvelous Bridge*, a place in the Pokemon World), and places in different types of space (e.g., *liver*, an organ in the space of the human body). In addition, non-spatial and spatial questions differ in the way the intended places are described in the questions. While in non-spatial questions places are described using place types, affordances, and situations, spatial questions contain toponyms as a direct reference to the intended places. In the first and second types of questions, spatial relationships to other places are frequently observed implicitly (e.g., 2nn, and 1nn), as well as explicitly (e.g., 1nrn, and 2trn).

The answers are also classified into three categories (Table 9). There is a noticeable difference between answer patterns relating to geographical and non-geographical places (the first two vs. the third category of answers). The difference between the first two categories of answers derives from how the answer is formulated, using names and implicit relationships, or with spatial relationships which are explicitly

Table 8: Types of questions

ID	Type	Percentage	Frequent Patterns
1	Non-spatial *Place-related questions not aiming at localisation of places*	41.5%	`2tnn`, `2tn`, and `2tno`
2	Spatial *Place-related questions about the location of places*	23.1%	`1n`, `1nn`, and `1nrn`
3	Non-geographical, and ambiguous *Place-related questions about non-geographical places (e.g., fictional places) or ambiguous questions*	35.4%	`1o`, `1os`, and `1oo`

Table 9: Types of answers

ID	Type	Percentage	Frequent Patterns
1	Explicit localization and spatial descriptions	25.8%	`nrnn`, `rnnn`, and `rnn`
2	Implicit localization (place names, and addresses)	42.5%	`n`, `nn`, and `nnn`
3	Non-geographical, and unanswered	31.7%	`oo`, `o`, and `ooo`

mentioned in the answers. Moreover, the category of implicit localization includes notably fewer distinct semantic encodings — only 0.3% of all unique encodings, while the explicit localization category contains most of the patterns, 97% of all unique encodings. In other words, a broad range of simple to complex spatial descriptions are categorized as explicit localization.

5.3 Question/Answer relationships

To investigate the relation between place questions and answers, two different scenarios are taken into account. First, we investigate the relation between types of questions and answers. Next, the concatenated encodings of questions and answers are investigated using a clustering approach and manual interpretation of the frequently observed patterns in each cluster.

5.3.1 From questions to answers

A many-to-many relationship between categories of questions and answers is considered and presented as a contingency table (Table 10). Spatial questions are mostly answered with implicit localization answers and are less ambiguous than non-spatial questions. This is because they are formulated with direct references to the intended places through toponyms. Non-spatial questions are more ambiguous and are mostly answered through complex descriptions or remain even unanswered. For example, *what are the best airports in Southern Utah?* is unanswered in the dataset, and it

Table 10: Contingency table of linking the clusters

Q/A	Explicit localization	Implicit localization	Non-geographical
Non-spatial	9512	7966	6266
Spatial	2463	8400	2332
Non-geographical	2777	7970	9535

cannot be answered without describing what *best* means here. In addition, we observe relationships between non-geographical and ambiguous questions and implicit and explicit localization answers. Several reasons contribute to the observed relationship, such as human interpretation of ambiguous questions, issues in extraction of place-related semantics, which are propagated to the semantic encodings, and similarity of patterns in formulating questions and answers for geographical and non-geographical places impacting on clusterings. There is a strong relationship between non-spatial question and answers with explicit localizations. Three primary reasons for this are:

1. **Answering style:** The content of question (how the intended places are defined) can also be repeated a part of answer (e.g., Question (`2trnn`) : *What beach is closest to Busch Gardens Tampa?* Answer (`nrnn`): *Clearwater beach is the closest to Busch gardens, Tampa.*).
2. **Ambiguous questions:** Ambiguous questions are answered in more detail (partially) related to localization information (e.g., Question (`2tn`): *What city is Olongapo City?* Answer (`nqtrnn`): *Olongapo City is a 1st class highly urbanized city in Central Luzon Philippines*).
3. **Misleading questions:** While the asker's question is semantically a *where-question*, it is formulated as a *what-question* (e.g., Question (`2nn`): *What is Bentonville Arkansas County?* Answer (`nrnn`): *Bentonville is in Benton County Arkansas*). Here, the encoding captures semantics different from the intent of the asker and this therefore affects the categorization of the question.

5.3.2 Concatenated clustering

In a last experiment, the semantic representations of questions and their corresponding answers are concatenated and then clustered. The results of clustering show (Table 11) that there is a direct one-to-one relationship between the categories of questions and the concatenated clusters. Using the Dice similarity coefficient, the similarity of first, second, and third clusters of questions and the first, second, and third clusters of questions concatenated with answers are 88.86%, 75.83%, and 75.98%, respectively.

Table 11: Frequent patterns in concatenated clusters

ID	Frequent Encoding	Example
1	`2tnn-n`, `2tn-n`, and `2tnn-nrnn`	*What county is Roselle NJ? Union County*
2	`1n-n`, `1nn-rnnn`, and `1nn-n`	*Where is Fairview? in Multnomah County Oregon United States*
3	`1o-oo`, `1oo-oo`, and `1no-oo`	*Where is the appendix to the liver? No Answer Present.*

6 Discussion and Conclusions

We analyzed the potential and limitations of MS MARCO V2.1 — a large-scale dataset of place questions and human-generated answers — for place-related studies. We report on our approach using semantic encoding and clustering techniques to derive categories of place questions and answers.

The relations between place questions and answers are studied by linking categories of questions and answers, and concatenated clusters. We manually interpret the most frequent patterns found to provide qualitative insights in place-querying behavior. Using geocoding and disambiguation techniques, we have successfully resolved the ambiguous place references in questions and answers. The geographical representation of places in questions and answers reveals that while the dataset has a global coverage, it is highly biased to North America, Western Europe, and Australia. While this bias may also be in part attributed to the biased coverage of the GeoNames gazetteer used for toponym resolution (Graham & De Sabbata 2015), users of the MS MARCO dataset are advised to be mindful of this coverage bias and its potential impact.

Through extraction of place-related semantics, we have found a semantic bias related to the scale of places. Analyzing the frequent place types shows that while questions to fine-grained places such as *schools, libraries, and airports* occur, most of the frequently asked place types are related to coarse geographic scales. The top-ten most frequent types of places identified in the questions and answers are associated with the *city to country level* scales. The diversity of activities, situations, qualities, and spatial relationships in the dataset shows further potential for studies in geographic information retrieval.

Expressing the patterns in questions and answers through the proposed semantic encoding proved to be useful for categorization. It produces results consistent with those achieved by the state-of-the-art word embedding approach using the pre-trained ELMo model. Yet, the encoding approach allows for manual inspection of the patterns and their qualitative interpretation, thus providing a valuable insight into place-related querying and answer-giving behavior.

Using a data mining approach and interpretation of the frequent patterns in the results, we found three main types of questions in MS MARCO V2.1 – i.e., non-spatial, spatial, and non-geographical and ambiguous questions. Applying the same strategy to the answers similarly reveals three main categories of answers, explicit localization (spatial descriptions), implicit localization (names and addresses), and

non-geographical and unanswered questions. By linking the questions and answers we show that even for spatial questions, people often answer with implicit localizations – e.g., *Where is Killaloe?* is answered with a single toponym, *Ireland*.

Our research reveals a high potential of the MS MARCO dataset for future place-related studies. It is, however, highly recommended to consider appropriate sampling strategies in order to deal with the geographical and semantic biases observed and discussed throughout this paper. Semantically richer representations that can differentiate more complex types of questions and answers may lead to a more nuanced characterization of the dataset and human place-querying behavior. Using sophisticated clustering approaches such *fuzzy C-means* may lead to derive better categorizations, however it remains as a future work of this study. Moreover, the proposed semantic encoding can be extended and used for other purposes such as translating place-related questions into computer-digestible queries.

Acknowledgements The support by the Australian Research Council grant DP170100109 is acknowledged.

References

Agnew, J. A. (2011), Space and place, *in* J. Agnew & D. N. Livingstone, eds, 'The SAGE Handbook of Geographical Knowledge', SAGE Publications Ltd, London, pp. 316–330.

Aloteibi, S. & Sanderson, M. (2014), 'Analyzing geographic query reformulation: An exploratory study', *J. Am. Soc. Inf. Sci. Technol.* **65**(1), 13–24.

Backstrom, L., Kleinberg, J., Kumar, R. & Novak, J. (2008), Spatial variation in search engine queries, *in* 'Proceedings of the 17th International Conference on World Wide Web', WWW '08, ACM, New York, NY, USA, pp. 357–366.

Caliński, T. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics* **3**(1), 1–27.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. & Manning, C. D. (2014), Universal Stanford dependencies: A cross-linguistic typology, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation', Vol. 14, pp. 4585–4592.

Edwardes, A. J. & Purves, R. S. (2007), Eliciting concepts of place for text-based image retrieval, *in* 'Proceedings of the 4th ACM Workshop on Geographical Information Retrieval', pp. 15–18.

Graham, M. & De Sabbata, S. (2015), 'Mapping information wealth and poverty: the geography of gazetteers', *Environment and Planning A* **47**(6), 1254–1264.

Guy, I. (2018), 'The characteristics of voice search: Comparing spoken with typed-in mobile web search queries', *ACM Transactions on Information Systems (TOIS)* **36**(3), 30:1–30:28.

Hollenstein, L. & Purves, R. (2010), 'Exploring place through user-generated content: Using Flickr tags to describe city cores', *Journal of Spatial Information Science* **1**(1), 21–48.

Hollink, J. M. V. (2008), Effects of goal-oriented search suggestions, *in* 'BNAIC 2008 Belgian-Dutch Conference on Artificial Intelligence', p. 177.

Jaro, M. A. (1989), 'Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida', *Journal of the American Statistical Association* **84**(406), 414–420.

Jones, R., Zhang, W. V., Rey, B., Jhala, P. & Stipp, E. (2008), 'Geographic intention and modification in web search', *International Journal of Geographical Information Science* **22**(3), 229–246.

Lee, C.-J., Craswell, N. & Murdock, V. (2015), Inter-category variation in location search, *in* 'The 38th International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, 2767797, pp. 863–866.

Li, X. & Roth, D. (2006), 'Learning question classifiers: the role of semantic information', *Natural Language Engineering* **12**(3), 229–249.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014), The Stanford CoreNLP natural language processing toolkit, *in* 'Association for Computational Linguistics (ACL) 2014 System Demonstrations', pp. 55–60.

Mourelatos, A. P. (1978), 'Events, processes, and states', *Linguistics and Philosophy* **2**(3), 415–434.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R. & Deng, L. (2016), 'MS MARCO: A human generated machine reading comprehension dataset', arXiv preprint arXiv:1611.09268.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018), Deep contextualized word representations, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', pp. 2227–2237.

Radlinski, F. & Craswell, N. (2017), A theoretical framework for conversational search, *in* 'Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval', pp. 117–126.

Sanderson, M. & Kohler, J. (2004), Analyzing geographic queries, *in* 'SIGIR Workshop on Geographic Information Retrieval', Vol. 2, pp. 8–10.

Shanon, B. (1983), 'Answers to where-questions', *Discourse Processes* **6**(4), 319–352.

Spink, A. & Gunar, O. (2001), 'E-commerce web queries: Excite and ask jeeves study', *First Monday* **6**(7).

Sui, D. & Goodchild, M. (2011), 'The convergence of gis and social media: challenges for giscience', *International Journal of Geographical Information Science* **25**(11), 1737–1748.

Tomko, M. & Winter, S. (2009), 'Pragmatic construction of destination descriptions for urban environments', *Spatial Cognition and Computation* **9**(1), 1–29.

Wang, L., Chen, L., Dong, M., Hussain, I., Pan, Z. & Chen, G. (2016), 'Understanding user behavior of asking location-based questions on microblogs', *International Journal of Human-Computer Interaction* **32**(7), 544–556.

Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y. & Li, Y. (2005), Detecting dominant locations from search queries, *in* 'Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 424–431.

Winter, S. & Freksa, C. (2012), 'Approaching the notion of place by contrast', *Journal of Spatial Information Science* **5**(1), 31–50.

Winter, S., Hamzei, E., Van De Weghe, N. & Ooms, K. (2018), A graph representation for verbal indoor route descriptions, *in* 'Spatial Cognition', Springer.