


Initial Analysis of Simple Where-Questions and Human-Generated Answers

Ehsan Hamzei 

The University of Melbourne, Parkville, Victoria, Australia
ehamzei@student.unimelb.edu.au

Stephan Winter 

The University of Melbourne, Parkville, Victoria, Australia
winter@unimelb.edu.au

Martin Tomko 

The University of Melbourne, Parkville, Victoria, Australia
tomkom@unimelb.edu.au

Abstract

Geographic questions are among the most frequently asked questions in Web search and question answering systems. While currently responses to the questions are machine-generated by document/snippet retrieval, in the future these responses will need to become more similar to answers provided by humans. Here, we have analyzed human answering behavior as response to *simple where questions* (i.e., *where* questions formulated only with one toponym) in terms of type, scale, and prominence of the places referred to. We have used the largest available machine comprehension dataset, MS-MARCO v2.1. This study uses an automatic approach for extraction, encoding and analysis of the questions and answers. Here, the distribution analysis are used to describe the relation between questions and their answers. The results of this study can inform the design of automatic question answering systems for generating useful responses to where questions.

2012 ACM Subject Classification Information systems → Question answering; Information systems → Spatial-temporal systems; Information systems → Information extraction

Keywords and phrases question answering, scale, prominence, where-questions

Digital Object Identifier 10.4230/LIPIcs.COSIT.2019.12

Category Short Paper

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

1 Introduction

People frequently ask about geographic information in Web search [7, 13] and question answering systems [10]. Among many types of geographic questions, *where* (localization intention) and *how-to-get-to* (navigation intention) questions are dominant [3]. In everyday communication, these questions can be answered in terms of place and route descriptions, respectively. However, human-generated answers in a human-human dialogue are different from retrieved responses in human-computer interaction [1]. While in human-human question answering, one receives relevant responses with sufficient contextual information, current computer-based tools are not able to deliver answers of similar qualities [4]. In future, tools that provide responses similar to human-generated answers are envisaged instead of just retrieving documents and snippets [8]. For this, human answering behavior should be investigated as a major prerequisite.

As described in relevance theory of communication [15], people's answering behavior is based on the relevance of the answer to the question and to the context of communication. Relevance theory describes human-generated answers as simple, short, selective and cognitively



© Ehsan Hamzei, Stephan Winter, and Martin Tomko;
licensed under Creative Commons License CC-BY

14th International Conference on Spatial Information Theory (COSIT 2019).

Editors: Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart;
Article No. 12; pp. 12:1–12:8



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

informative responses [15]. However, retrieved responses by computer-based tools and human-generated answers differ in both content and structure [6]. The retrieved documents/snippets may contain both relevant and irrelevant information regarding the question, and their structure (the flow of information) is not specifically designed to satisfy the inquirer’s information need.

Geographic questions and specifically *where-questions* have special characteristics compared to other types of questions. The important influence of the inquirer’s location on the relevance of answers has been noted [14]. Similarly, descriptive factors of places – their scale, type, and prominence – have a direct effect on the formation of the answers [14]. A thorough analysis of the relation between the questions and answers in terms of type, scale, and prominence of referred places is, however, still missing. Yet, this is an essential prerequisite for the understanding of the human question answering behavior. Here, we propose an approach to analyze the questions and answers based on type, scale, and prominence of places mentioned in their content. In short, we contribute:

- An encoding representation for the question and answers based type, scale, and prominence;
- Insights on the relation of type, scale, and prominence of places mentioned in the questions and answers by analyzing a large question answering dataset.

2 Data

MS-MARCO v2.1 [11] is a general purpose machine comprehension dataset provided by Microsoft. It contains question-answer pairs of the following types: (1) *numeric*, (2) *entity*, (3) *location* (including geographic questions [2]), (4) *person*, and (5) *description* [11]. Here, we focus on the MS-MARCO *location* records containing *simple where-questions* (questions with a single toponym) and their human-generated answers. Due to the lack of rich contextual information inside the *simple-where questions*, these questions are the base case and likely harder to be answered than *where-questions* with multiple toponyms.

3 Methodology

In this study, we defined a new representation encoding of the question and answer pairs by capturing their type, scale and prominence sequences, respectively. These sequences consist of values of the factors for the places referred to in the questions, followed by the values for the toponyms in the answers, ordered as they appear in the text. For example, the pair of question and answer *Where is Melbourne? In Victoria, Australia.* is encoded into a *type-sequence*: $\{city, state, country\}$.

Here, we first propose a process to extract, encode, and analyze the question/answer pairs. In the extraction step, toponyms from the questions and answers are extracted using both the Geonames and OpenStreetMap (OSM) Nominatim gazetteers. Next, the records extracted from the gazetteers are encoded to sequences of scale, prominence, and type. Finally, the relation between places in the questions and in their answers are investigated, using distribution analysis of the encoded sequences.

3.1 Extraction

The process starts by first filtering *location questions* that are started with *where* from the corpus. Then, the text is geoparsed for toponyms by matching against the gazetteers. Using parse tree information, noun phrases in the questions and their answers are checked against

the gazetteers starting from compound to simple noun phrases. Due to the characteristics of the extracted question/answer pairs (i.e., short texts, geographic where-questions, and localization information in the answers), every simple/compound noun phrase is considered as a toponym candidate. Finally, the ambiguity of toponyms in the pairs of corresponding questions and answers are resolved using map-based disambiguation techniques proposed in [9]. Consequently, the results of extraction are two gazetteers records (i.e., Geonames and OSM Nominatim records) for each extracted toponym in every pair of question and answer.

3.2 Encoding

To examine the relation between the structures of questions and their answers, three proxies have been defined for type, scale, and prominence of places, respectively. We have used toponym attribute information from gazetteers for this encoding. Sequence representations for each question and answer pairs are then generated based on these encoded values. To reduce the impact of gazetteers data incompleteness, only records which can have all extracted toponyms completely encoded into type, scale and prominence are further analyzed.

For type encoding, the Geonames schema of 667 place types (aka. feature codes) has been used without further changes¹. The feature codes which are mentioned in the content of this paper are described in Appendix A.

A finite set of cognitively meaningful granularity levels is a prerequisite for encoding gazetteers records by scale. We have therefore adapted the seven-level schema from [12], with the granularity levels sequence of (1) furniture, (2) room, (3) building, (4) street, (5) district, (6) city, and (7) country. We have extended the schema to ten levels by adding coarser levels of scale: *county*, *state*, *country*, and *continent*. Nominatim records include an attribute (a number between 0–30) related to the OSM definition of scale (i.e., *place_rank*²). To convert the extracted gazetteers' records into the appropriate scale level, a look-up table linking OSM scale levels into the proposed scale schema has been devised manually.

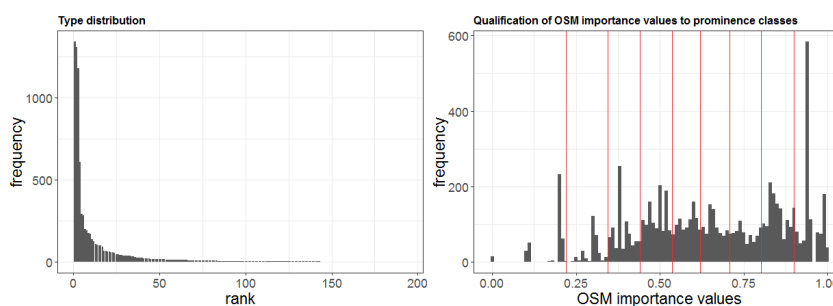
Finally, we have used the *importance* attribute in the extracted Nominatim records as a proxy measure of place prominence. This value is estimated based on different factors, such as the frequency of the place appearances in Wikipedia². The value ranges between 0 and 1, and it is designed to be used for ranking search results. To evaluate the prominence of places in questions and answers, we have classified these value into nine discrete levels of prominence by using the Jenks *natural breaks* method [5].

3.3 Distribution analysis

To investigate the relation between the questions and answers, we conducted distribution analysis of the encoded question/answer pairs. In distribution analysis, overall and sequence distributions are investigated and discussed. Overall distributions for questions and answers reveal the differences between places mentioned in questions, and places referred to in the answers. Sequence distributions show the distributions of values in each position of the encoded sequences (e.g., type sequences). The sequence distributions are used to investigate formation of the human-generated answers, in addition to their relations to the corresponding questions.

¹ <https://www.geonames.org/export/codes.html>

² https://wiki.openstreetmap.org/wiki/Nominatim/Development_overview



■ **Figure 1** Distributions of type and prominence of toponyms in the questions and answers.

4 Results and discussion

4.1 Extraction and encoding results

In the extraction process, 3238 simple where questions (from 31204 where questions) are found. Due to incompleteness of data in gazetteers in some cases the encoding into type, scale, and prominence cannot be done. Hence, during the encoding to type, scale, and prominence the number of records decreases by *22.5%* (2511 records out of 3238), *50.1%* (1587 records out of 3238), and *22.5%* (2511 records out of 3238).

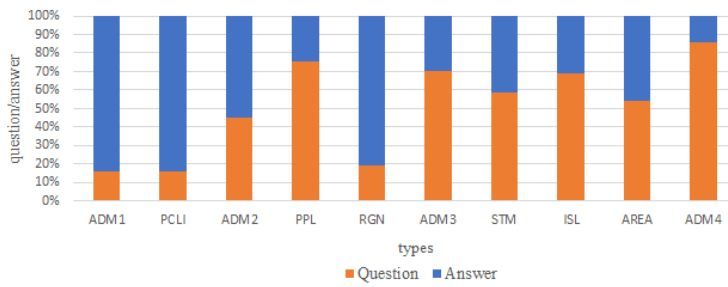
After encoding the data, we find that only 185 unique place types out of 667 are referred to in the questions and answers. The frequency of these types forms a heavy long-tail distribution (Figure 1), where *81.6%* (i.e., 6072 out of 8218) of the extracted types belong to twenty unique types. This shows the reliance of people on few fundamental place types in the interpretation and answering of a large number of Web-based where-questions. In other words, the types in the corpus are biased in a way that a few types (e.g., states) are frequently observed, and a relatively large number of types (e.g., bridges) are found rarely in the dataset.

As shown in Figure 1, the *importance of places* extracted from Nominatim records are biased to medium and high values, which can be related to the geographic information people seek when they submit questions to search engines. The vertical lines in Figure 1 show the class breaks after classification of the continuous quantitative *importance* values into the nine levels of prominence.

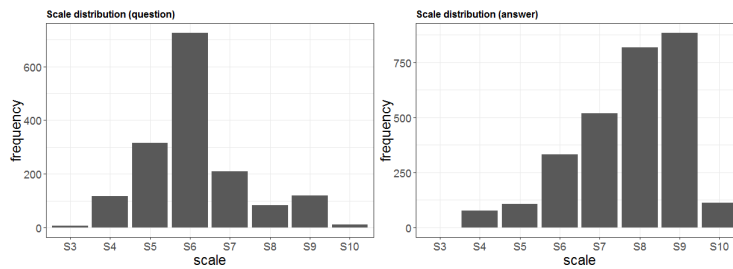
4.2 Overall distribution analysis results

Figure 2 shows the distribution of the ten most-frequent place types in the questions and answers. Some types, such as *ADM1* (first-order administrative divisions), *PCLI* (independent political entities), and *RGN* (regions) are mostly used to formulate answers, while types such as *ADM3* (third-order administrative divisions), *ADM4* (fourth-order administrative divisions), *STM* (streams and rivers) and *PPL* (populated places, incl. villages and cities) are more frequently referred to in the questions. In other words, the distributions of type in the questions and their answers are systematically different. While lower-levels administrative divisions (e.g., *ADM1*) are frequently observed in human-generated answers, natural places (e.g., streams) and higher-levels administrative divisions (e.g., *ADM4*) are most frequently mentioned in the questions.

The type distribution is strongly related to the scale of the referents (Figure 3). While most of the questions are asked using place references at the *city* level of scale, they are answered at the *country*, *state*, and *county* levels. People are more searching for geographical-

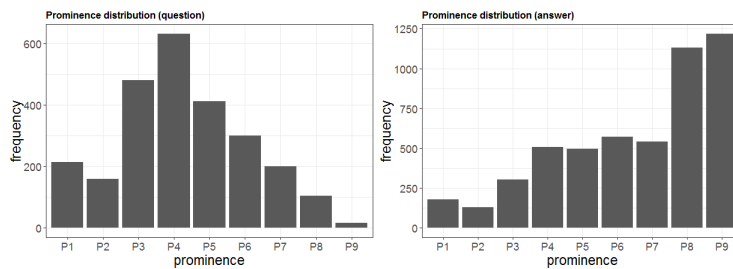


■ **Figure 2** Distribution of types in the questions and answers, for the top ten most frequent types.

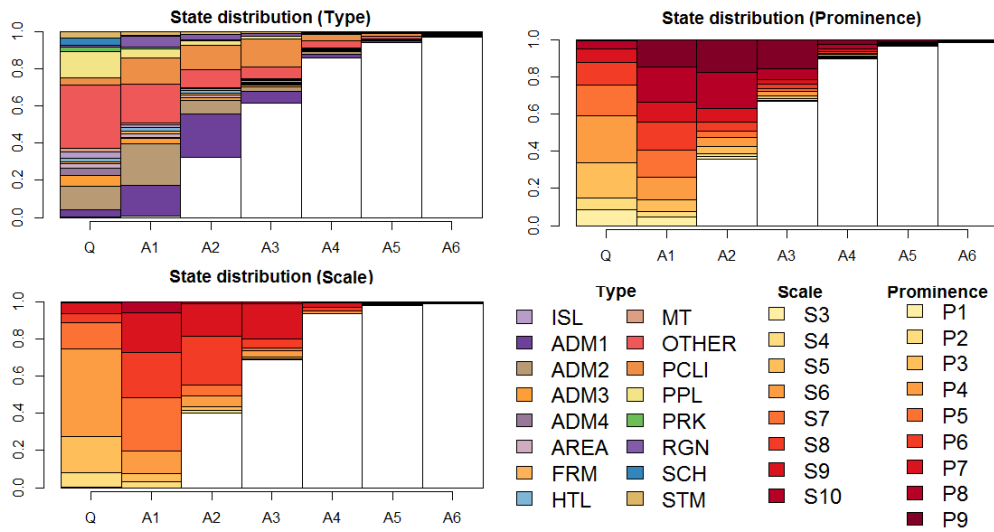


■ **Figure 3** Scale distribution in the questions and answers.

scale places at the district and city levels of scale, while the answers to these questions are related to coarser levels such as country and state levels. We also note the lack of questions relating to fine-grained scale places. Similarly, Figure 4 shows the prominence distribution, centered around mid-range values for questions and biased to high-levels in answers. Two differences are, however, noticeable when comparing the distributions of scale and prominence. First, the coarsest level of scale (Level 10) is far less frequent than the highest level of prominence. The reason is that simple where question answers using continent level places references would be uninformative (i.e., of *low relevance*), while this is not the case for prominence (i.e., more prominent references are more relevant due to lower cognitive processing effort). Second, the overall distributions are similar in terms of skew (questions have positive skew, and answers have negative skew), however, their kurtosis is different (in both questions and answers, the distribution of scale is steeper than prominence). These patterns reveal that while scale and prominence may seem generally correlated, they capture distinct characteristics of places, with complex non-linear mapping between them. Evidently, the observed results are directly affected by the proxies used to capture type, scale and prominence.



■ **Figure 4** Prominence distribution in the questions and answers.



■ **Figure 5** Sequence distribution of type, scale, and prominence.

4.3 Sequence distribution analysis results

Figure 5 shows the sequence distributions of type, scale, and prominence of places. In Figure 5 only the most frequent types are visualized in the sequence, and the rest are presented as *OTHER*. As the data contain only few answers with more than six toponyms (with a long tail distributions up to a maximum of 13 toponyms), we have focused only on the first six toponyms (capturing 94.3% of the question-answer pairs). Most of the answers contain less than three toponyms. Figure 5 also reveals the differences between questions and answers in terms of type, scale, and prominence. Answers are formulated such that they start with lower values and end with higher values of both scale and prominence (fine to coarse, less to more prominent). In answers, certain type sequences are dominant: *ADM2* (e.g., Los Angeles County), *ADM1* (e.g., California), and *PCLI* (e.g., United States) are the most popular types in the first, second, and third positions of the answer-sequences, respectively. In general, the sequences of places which are mentioned in the answers are starting with less-known values in terms of type, scale, and prominence (i.e., low levels of scale and prominence, and particular types of places such as *ADM2*, and *ADM3*), and continue to well-known places in terms of these factors (i.e., higher levels of scale and prominence, and specific types of places such countries and political entities). In Appendix B, the patterns in scale and prominence sequences are investigated in more detail.

5 Conclusion

This paper presents a preliminary investigation of the relation of simple where questions and their human-generated answers. Type, scale, and prominence have been used as factors to investigate the human answering behavior of the simple where questions. We have proposed an approach for extracting, encoding and analyzing MS-MARCO question/answer records into type, scale, and prominence sequences. Later, we have discussed the relation based on overall and sequence distributions of these factors in the questions and their answers.

The results of this study show that human-generated answers to the questions follow a specific pattern starting from less-known values of type, scale, and prominence and continue to well-known places. This study reveals that type, scale, and prominence of places mentioned in questions has a direct relationship to formation of their answers. In summary, we have

shown that type, scale, and prominence are important factors which can be used to describe human answering behavior. Consequently, these factors can be used for mimicking human answering behavior to provide synthetic responses similar to human-generated answers.

This study shows the preliminary results of analyzing question answering data using type, scale, and prominence encoding. In future research, more research is needed to utilize and extend the proposed encoding approach to extract association rules from question answering datasets and to predict the structure of answers based on the encoding representation of the questions. In addition, the results of this study are limited to the context of Web search questions. Future work in other question/answering scenarios, especially contextualized human-human dialogue, lead to better understanding of human answering behaviour.

References

- 1 Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. *Comparing Several Aspects of Human-Computer and Human-Human Dialogues*, pages 133–159. Springer Netherlands, Dordrecht, 2003. doi:10.1007/978-94-010-0019-2_7.
- 2 Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. Place questions and human-generated answers: A data analysis approach. In *Geospatial Technologies for Local and Regional Development*, pages 1–16, 2019.
- 3 Andreas Henrich and Volker Luedecke. Characteristics of Geographic Information Needs. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR '07*, pages 1–6, New York, NY, USA, 2007. ACM. doi:10.1145/1316948.1316950.
- 4 Lynette Hirschman and Robert Gaizauskas. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300, 2001.
- 5 G. F. Jenks. The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography*, 7:186–190, 1967.
- 6 Karen Sparck Jones. *Retrieving information or answering questions?* London : British Library. (British Library Annual Research Lecture ; 8), 1990.
- 7 Emilia Kacprzak, Laura Koesten, Jeni Tension, and Elena Simperl. Characterising Dataset Search Queries. In *Companion Proceedings of the The Web Conference 2018, WWW '18*, pages 1485–1488, Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. doi:10.1145/3184558.3191597.
- 8 Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011. doi:10.1016/j.ins.2011.07.047.
- 9 Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding Spatial Named Entities for Information Extraction and Question Answering. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1, HLT-NAACL-GEOREF '03*, pages 31–38, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi:10.3115/1119394.1119399.
- 10 Xin Li and Dan Roth. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249, 2006.
- 11 Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016.
- 12 Daniela Richter, Stephan Winter, Kai-Florian Richter, and Lesley Stirling. Granularity of locations referred to by place descriptions. *Computers, Environment and Urban Systems*, 41:88–99, 2013. doi:10.1016/j.compenvurbsys.2013.03.005.
- 13 Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *SIGIR Workshop on Geographic Information Retrieval*, volume 2, pages 8–10, 2004.
- 14 Benny Shanon. Answers to where-questions. *Discourse Processes*, 6(4):319–352, 1983.
- 15 Deirdre Wilson and Dan Sperber. Relevance theory. In *Handbook of Pragmatics*. Blackwell, 2002.

A Feature codes

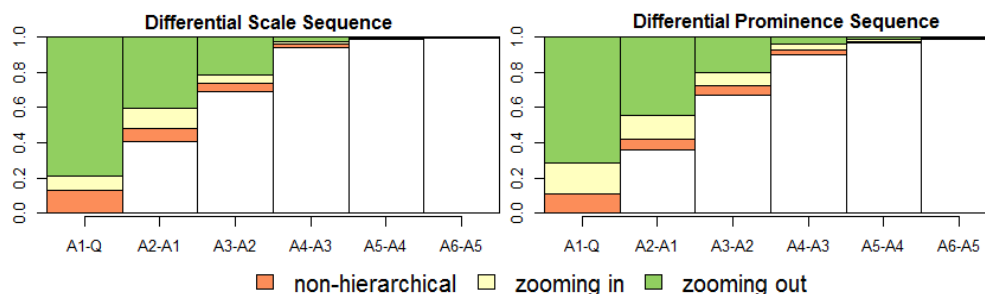
Table 1 below shows the types which are mentioned in paper. The complete list can be found in the Geonames website.

■ **Table 1** Feature codes used in the paper (extracted from Geonames documentation).

Code	Description	Example
ADM1	first-order administrative division (states, and provinces)	Oklahoma
ADM2	second-order administrative division (counties)	Brevard County
ADM3	third-order administrative division (cities)	City of Alhambra
ADM4	fourth-order administrative division (towns)	Newburgh
AREA	a part of land without homogeneous character/boundaries	Theresienwiese
FRM	a part of land dedicated to agricultural purposes	Branksome
HTL	hotels	The Carriage House
MT	mountains	Eagles Nest
PCLI	independent political entity	Paraguay
PPL	diverse type of populated places (e.g., cities, and villages)	El Granada
PRK	parks and recreational places	Franklin Square Park
RGN	an area with particular cultural character	Central Africa
SCH	schools and universities	Stuyvesant High School
STM	streams	Withlacoochee River

B Differential scale and prominence sequences

Figure 6 shows the hierarchical (i.e., zooming in, zooming out), and non-hierarchical patterns in scale and prominence sequences using differential sequences. The differential sequences are created by subtracting values from their previous values in the scale and prominence sequences. Due to the fact that scale and prominence are ordinal values, the subtraction values are not valid, and consequently using *sign* function the values are translated into meaningful values – i.e., **0** (equal), **+** (greater than) and **-** (less than). Here, **0** values show the *non-hierarchical* pattern because the scale or prominence levels are not changed. The **+** values show the *zooming out* pattern, because the level of scale or prominence is increased compared to its previous level in the sequence. The **-** values show the *zooming in* pattern with same rationale. Figure 6 supports the discussion made in the paper, section 4.2, that values in the scale and prominence sequences are hierarchically structured starting with lower values (levels) followed by higher ones.



■ **Figure 6** Sequence distribution of differential scale and prominence sequences.