ARTICLE TEMPLATE

Templates of generic geographic information for answering where-questions

Ehsan Hamzei, Stephan Winter and Martin Tomko[a]

[a]The University of Melbourne, Parkville, Australia

ARTICLE HISTORY
Compiled December 17, 2021

ABSTRACT
In everyday communication, where-questions are answered by place descriptions. To answer where-questions automatically, computers should be able to generate relevant place descriptions that satisfy inquirers' information needs. Human-generated answers to where-questions constructed based on a few anchor places that characterize the location of inquired places. The challenge for automatically generating such relevant responses stems from *selecting relevant anchor places*. In this paper, we present *templates* that allow to characterize the human-generated answers and to imitate their structure. These templates are patterns of generic geographic information derived and encoded from the largest available machine comprehension dataset, MS MARCO v2.1. In our approach, the toponyms in the questions and answers of the dataset are encoded into sequences of generic information. Next, sequence prediction methods are used to model the relation between the generic information in the questions and their answers. Finally, we evaluate the performance of predicting templates for answers to where-questions.

KEYWORDS
question answering; notion of place; scale; prominence

1. Introduction

Consider the following question and its corresponding answer, taken from the Microsoft Machine Comprehension (MS MARCO) dataset v2.1 (Nguyen et al., 2016):

Question: *Where is Putney Bridge?*
Answer: *Putney Bridge is a bridge crossing of the River Thames in west London.*

This *where*-question is answered using a *place description* – a description that characterizes the location of interest (Putney Bridge) based on a few anchor places (River Thames and London). Place descriptions, however, are not the only way to answer where-questions. Where-questions can also be answered via other representations such as maps or sketches (Church, Neumann, Cherubini, & Oliver, 2010). Invariant to the chosen representation, the answers localize the place in question based on its spatial relationships with chosen anchor places (Couclelis, Golledge, Gale, & Tobler, 1987).

Correspondence concerning this article should be addressed to Ehsan Hamzei, Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia; Email: ehamzei@student.unimelb.edu.au

Hence, answering where-questions poses the following challenges no matter what representation is used:

- Generating informative answers – i.e., the answer should complete the inquirers' gap of knowledge in a way that obvious or already-known responses should be avoided and useful and necessary information are included (Shanon, 1983). In the example, obvious, inadequate or undetailed answers such as *on Earth* or *in the UK* or *over a river* are avoided by the responder.
- Answering the question in a cognitively efficient manner (D. Wilson & Sperber, 2002) – e.g., producing short and straightforward place descriptions (Hamzei, Li, et al., 2019) and personalized map labelling strategies in map visualizations (J. A. Wilson, 2018). In the example, the responder excludes unnecessary information such as the nearby theaters and restaurants to keep the answer as simple and relevant as possible.
- Determining the level of granularity of answers – e.g., a suitable zoom level for maps (Ballatore, 2019) and referring to places of suitable granularity in place descriptions (Hamzei, Winter, & Tomko, 2019). In our example, the name of the roads and streets that are connected to the bridge are neglected in the answer based on the judgement of the responder for the relevant scale level.
- Selecting places that can be assumed to be known by the inquirer – e.g., labelling the places known to inquirers in maps (Suomela, Lakkala, & Salminen, 2009) and referring to them in place descriptions as anchors. In the example, the location of *River Thames* and *London* are assumed to be known to the inquirer.

Where these challenges are met by an answer, the communication succeeds.

Addressing these challenges is a necessary step towards answering where-questions. To understand and imitate human selectivity in choosing anchor places, we investigate and characterize human-generated answers to where-questions. The results of our research are applied for generating answers to where-questions as natural language responses (place descriptions). Selecting relevant anchor places is an essential part of generating place descriptions that succeed in answering where-questions. Moreover, information about anchor places can be used in static maps to be visualized in a proper context frame (Ballatore, 2019).

Current geographic question answering systems are often focused on coordinate retrieval as answers to where-questions (e.g., Luque, Ferrés, Hernando, Mariño, & Rodríguez, 2006; Stadler, Lehmann, Höffner, & Auer, 2012). While coordinates are useful for communication between location-based services to perform spatial analysis or visualization (Jiang & Yao, 2006), it is not necessarily a relevant response to inquirers without a proper map visualization. Yet, a characterization of relevant anchor places to localize a place in question is still missing.

In this paper, we study human-generated answers to where-questions to inform the properties of such answers and to devise and test a method to imitate their structures in machine-generated responses. To achieve these goals, the information in a where-question and its answer is modelled as *an ordered set of places* that are mentioned in their content. Then the properties of places in questions and corresponding answers are derived and further investigated. This model forms a template (i.e., an ordered set of place properties) that enables computers to learn and imitate human answering behaviour. In other words, place properties are utilized to understand why a set of places are chosen as anchors to localize the place in question and how this selectivity can be imitated by computers.

The properties that are used in the templates are generic geographic information

that describe the shared meaning of places in form of generic types from a finite set of categories. Referring to the example above, the place in question is a *bridge* which is localized by referring to the river it goes over and the city it belongs to. Here, the template captures the structure of the answer as relationships between *bridges and rivers*, and *bridges and cities*.

1.1. Background: Geographic Question Answering

Geographic Question Answering (GeoQA) is defined as methods and algorithms that help inquirers to satisfy their information need by deriving answers to their geographic questions. In GeoQA, answering geographic questions can be based on diverse information sources such as textual information (Ferrés & Rodríguez, 2006; Mishra, Mishra, & Agrawal, 2010), geodatabases (Chen et al., 2013), and spatially-enabled knowledge bases (Ferrés & Rodríguez, 2010). GeoQA (and in general QA) architectures typically resolve three tasks: (a) question classification and intent analysis, (b) finding relevant sources, and (c) extracting answers from the sources (Ferrés & Rodríguez, 2006).

The classification of the questions (Hamzei, Li, et al., 2019; Mohasseb, Bader-El-Den, & Cocea, 2018) enables GeoQA to coarsely identify the intent and purpose of asking questions (e.g., localization, or navigation). Next, the questions are translated into formal representations such as database queries or even just a vector representation of extracted keywords (Punjani et al., 2018). Using the representations, the information sources can be searched or queried to look up the possible answers (Zheng, Cheng, Yu, Zou, & Zhao, 2019). Finally, the factoid answers are retrieved from the sources – e.g., a sentence in a Web document, a cell in a selected table, or a node in a graph knowledge base (Sun et al., 2018).

In recent years, several GeoQA studies were conducted for answering geographic questions (Stadler et al., 2012), creating knowledge bases from unstructured data (Mai, Janowicz, He, Liu, & Lao, 2018), and relaxing unanswerable questions (Mai, Yan, Janowicz, & Zhu, 2020). Focusing on answering geographic questions, previous studies provide solutions to retrieve responses from knowledge bases (Stadler et al., 2012) and documents (Buscaldi, Benajiba, Rosso, & Sanchis, 2006; Luque et al., 2006). GeoQA studies are mostly focused on what/which questions about geographic places (e.g., Scheider, Nyamsuren, Kruiger, & Xu, 2020; Vahedi, Kuhn, & Ballatore, 2016). In answering where-questions, the task is either simplified into retrieving stored coordinates (Luque et al., 2006; Stadler et al., 2012), or selecting a part of text without explicit adaptation to the question (Buscaldi et al., 2006).

When answering where-questions, the answer extraction step is particularly challenging. Without a well-designed approach to imitate human answering behavior, the extracted answers can easily be over-specified and consequently uninterpretable for the inquirer, or under-specified and thus obvious and uninformative to the inquirer (Shanon, 1983). Hence, the challenge is to provide relevant answers by selecting proper set of anchor places to localize the place in question.

1.2. Rationale and Research Gap

To enable computers to provide responses with similar qualities to human-generated answers, the responses need to be relevant. An answer is relevant if its positive cognitive effects to inquirers are large and the processing effort to achieve the effect is small (D. Wilson & Sperber, 2002). In other words, answers should be informa-

tive enough and as straightforward as possible. Assuming human-generated answers are relevant responses, machine-generated responses should imitate the selectivity in human-generated answer to provide useful pieces of information and avoid unnecessary ones. Generating such relevant responses is the major prerequisite of intelligent GeoQA as defined by Winter (2009).

Generic information captures shared meaning of geographic places. While generic geographic information is not used in QA, it has been used to investigate and characterize place descriptions (Edwardes & Purves, 2007; Richter, Winter, Richter, & Stirling, 2013), route descriptions (Raubal & Winter, 2002), and regions (Tomko & Purves, 2008).

This research hypothesizes that, at least in the English language, generic geographic information can be used to characterize human answering behavior and ultimately to generate templates for answering where-questions. We approach this hypothesis by addressing three sub-hypotheses.

Sub-hypothesis 1 (Characteristics of the answers)**.** Human-generated answers to where-questions have special characteristics that can be described and characterized in terms of generic geographic information such as type, scale, and prominence;

Sub-hypothesis 2 (Relation between where-questions and their answers)**.** There is a strong relationship between generic information in the content of where-questions and their answers which can be used to characterize human answering behavior;

Sub-hypothesis 3 (Generating answers to where-questions)**.** If Hypotheses 1 and 2 hold, the characteristics of human-generated answers and the relation between the questions and their answers can be used to generate templates to answer to where-questions.

To investigate the hypotheses, the following research questions will be addressed:

(1) How can the characterizing patterns of the human-generated answers be derived?
(2) How does generic geographic information in where-questions relate to the generic information in their human-generated answers?
(3) How can the templates be generated to imitate the structure of human-generated answers?

By addressing the research questions, we contribute:

- A generalized approach to investigate human answering behavior to where-questions using generic geographic information;
- An investigation of the human-generated answers to where-question asked in Web search, using patterns of type, scale and prominence of places.

2. Methodology

To investigate the hypotheses, we propose a generalized approach of *specific-generic translation*. Next, a method using specific-generic translation is devised to investigate the QA in interaction of people with a general-purpose search engine. Other QA scenarios (e.g., human-human dialogue, human-social bot interaction) may require different design of specific-generic translations.

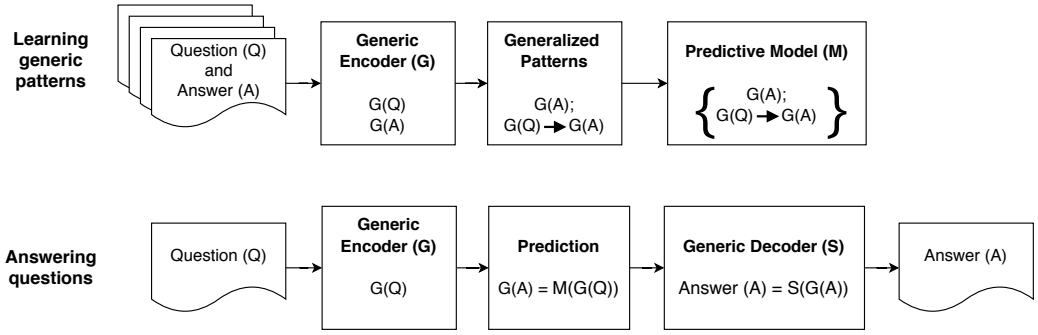

Figure 1. Specific-generic translation approach

2.1. Specific-Generic Translation

Figure 1 shows the proposed approach to derive characterizing patterns in human-generated answers and generating template to answer where-questions. The approach includes two stages: (1) *learning generic patterns* where the objective is to investigate and to characterize human answering behavior into a machine learning model, and (2) *answering questions* where the model is used to generate answers. The novelty of the approach is in the encoding of questions and answers into their generic meaning and to model the relation between questions and answers in their generic forms. Later, the model is used to generate generic forms of answers to where-questions. Finally, the answer is constructed by decoding generic form (e.g., type) of the answer into its specific representation (i.e., toponym).

The specific-generic translation approach involves the following steps:

(1) Selecting a set of generic information classes (e.g., place type, scale, prominence, functions and access) based on the context of QA and availability of data;
(2) Defining a schema for each selected generic information class;
(3) Designing an information extraction approach to encode the questions and answers into generic forms (*Generic Encoder* in Figure 1);
(4) Evaluating how effective each generic class is in capturing the relation between the questions and their human-generated answers (*Generalized Patterns* in Figure 1). The results of evaluation also provide insights about human answering behavior in the context of the QA problem.
(5) Training a predictive model that can learn generalized patterns of human-generated answers (*Predictive Model* in Figure 1);
(6) Defining a decoding approach to map generic forms of answers into specific (toponym) representation (*Generic Decoder* in Figure 1). This step can be followed by additional steps such as natural language generation to be used in real-world applications.

In this paper, we discussed the results of the first five steps for question answering in a Web search scenario in details. The last step is only demonstrated using examples.

2.2. Type, Scale and Prominence (TSP) Encoding

Based on the specific-generic translation, TSP encoding is proposed to investigate where-questions constructed only with toponyms. The generic forms which are used to investigate these questions and their answers are *type*, *scale* and *prominence* of

the toponyms. We first introduce our terminology before discussing the details of the proposed TSP encoding method.

2.2.1. Definitions

The investigated types of where-questions are defined as:

- **Toponym-Based Where-Question (TWQ)**: A toponym-based where-question is a geographical where-question that is generated completely using toponyms. For example, *Where is Beverly Hills?* is a toponym-based where-question, while *Where is Clueless filmed?* (without toponym) and *Where can I buy furniture?* (affordance, buying furniture) do not belong to this type.
- **Simple Where-Question (SWQ)**: Simple where-questions are a sub-class of TWQs that contains only one toponym in their body (e.g., *Where is Beverly Hills?*).
- **Detailed Where-Question (DWQ)**: Detailed where-questions are a sub-class of TWQs with more than one toponym in their content (e.g., *Where is Beverly Hills, California?*). In DWQs, contextual details are provided in the content of the questions that shows what the inquirer already know – e.g., Beverly Hills is located in California.

We use *type*, *scale* and *prominence*, defined as:

- **Type**: A type (e.g., restaurant, mountain) is a reference to a group of places with similar characteristics (e.g., affordance, or physical properties). Type defines similar places and differentiates dissimilar ones, sometimes in a hierarchical or taxonomical manner. Here, the relation between a specific reference to a place (unambiguous toponym) and its type is considered as a one-to-one relation.
- **Scale**: Scale is defined as a finite hierarchically-organized ordinal set of levels grounded in the human understanding and differentiation of places based on their size and their relationships (i.e., containment). The relation between scale and an unambiguous toponym is considered as one-to-one. Due to the specific context of the QA scenario, very fine levels of scale of geographic entities, such as room-level or object-level, can be neglected here, while in everyday human-human communication these levels of scale may have a more important role.
- **Prominence**: Prominence is a measure of how well-known a place is. In this research, prominence is characterized by a finite ordinal set of levels. While prominence of places is subjective and differs from person to person based on their experience, here prominence is considered as an absolute and objective measure to rank places, established through a proxy measure defined later. This approach enables to avoid individual experiential biases and is supported by the evidence of success in day to day communication in which the absolute prominence evaluation is adapted between hearers and speakers.

Type, scale and prominence are used to characterize place descriptions (Edwardes & Purves, 2007; Richter et al., 2013). These geographic concepts can be used to capture different kinds of relationships among places. These relationships can be used to understand the relation between where-questions and their answers. For example, such relationships between rivers and seas (*flows to*), and cities and countries (*part of*) can be captured using place type. Considering the relation between where-questions and their answers, *containment* (different levels) and *nearness* (a same level) can be captured through differences among scale levels. Finally, prominence is a measure to check

whether the answer is interpretable by the inquirers – i.e., more prominent places are expected to be better known by inquires.

Finally, aspects of human-generated answers which are investigated in this paper are defined below:

- **Content**: The content of an answer is a collection of distinct information units that are presented to satisfy the inquirer's information need. Content can be generic (e.g., type) or specific (e.g., toponym). Content is the most important aspect of the answers, in a way that the difference between correct and incorrect responses are completely based on their content.
- **Style**: The style of an answer is the way that the content is presented. Style directly influences the perception of naturalness of the response. Referring to the introductory example, *... Putney Bridge is a bridge crossing of the River Thames in west London* and *... Putney Bridge is a bridge in west London which goes over River Thames* are two different styles of answers (with same content) to the question. Here, the former is preferred by the responder.

2.2.2. TSP Sequences

In TSP encoding, we use a sequence representation to model generic/specific information in the questions and answers. A sequence is defined as an ordered set of items (here, references to generic/specific geographic information). We first model questions and their corresponding answers as sequences of toponyms (specific representation). Then, these toponym sequences are encoded into type, scale and prominence sequences by translating each specific toponym into its corresponding generic type, scale and prominence reference. Referring to the introductory example, the specific representations (toponym sequences) and the encoded type sequences (an example of a generic sequence) of question and answer are presented below:

- **Toponym sequences**: [Putney Bridge] [River Thames, London]
- **Type sequences**: [bridge] [river, city]

Here, the *content* refers to the information items in the sequences, and their order defines the *style* in which the information is presented.

3. Implementation

Figure 2 shows the proposed workflow[1] to investigate TWQs and their answers. Here, we detail the dataset, extraction, encoding, generic patterns and prediction. A complete implementation of the proposed TSP encoding approach also includes decoding from generic to specific information. Here, the decoding step is demonstrated through examples, and a fully automated implementation remains out of scope of this paper.

3.1. Data

The questions in MS MARCO v2.1 (Nguyen et al., 2016) are categorized into five categories using tags: (1) *numeric*, (2) *entity*, (3) *location*, (4) *person*, and (5) *description* (Nguyen et al., 2016). Geographic questions can thus be easily extracted using

[1] Additional details of implementation are presented in the supplementary material (Section 1)

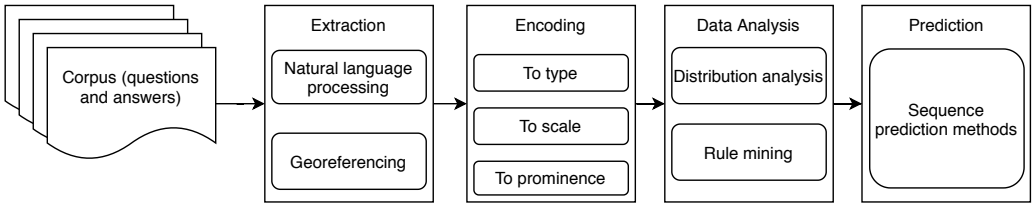

Figure 2. The proposed implementation approach

the predefined *location* tag. The dataset contains over one million records divided into *training*, *development* and *testing* subsets.

Each record in the dataset includes a *question*, *human-generated answer(s)* (except for records in the *test* dataset, where the answers are deliberately excluded), a *tag*, and a *set of documents* retrieved by the Microsoft Bing search engine[2].

The 'location questions' in MS MARCO (56,721 question-answer pairs) include 36,939 geographic questions, and the remainder are questions about fictional, mystic and other non-geographic places (Hamzei, Li, et al., 2019). Among the geographic questions, 13,195 pairs of questions and answers are geographic where-questions (Hamzei, Li, et al., 2019).

There are several reasons to choose MS MARCO for this study considering other available datasets such as SQuAD (Rajpurkar, Zhang, Lopyrev, & Liang, 2016):

- MS MARCO is the largest available QA dataset;
- The questions are labelled and geographic questions can be easily extracted;
- All questions are asked in a specific real-world scenario (i.e., Web search);
- Inquirers pose questions to resolve their information needs while in some datasets such as SQuAD, questions are made from documents. In other words, questions in SQuAD is more about what a document can answer rather than what actual inquirers want to know.
- The answers are provided using an open form strategy. The answerers can utilize suggested documents (one or more) and their own knowledge to answer a question. Hence, the answers are not directly extracted as a single span of a document.

3.2. Extraction

We first extract the questions labelled as *location* and starting with a *where*-token. Next, the text is the toponyms inside the questions and answers are identified using Named Entity Recognition (NER) and gazetteer lookups using both OSM Nominatim and Geonames gazetteers. Here, the Stanford NLP toolkit is used to extract named entities (Finkel, Grenager, & Manning, 2005). In this step, if a compound noun phrase is tagged as location, first the compound noun is checked by gazetteer lookup; if it is not identified, then its constituent simple nouns are checked. If a compound or simple noun phrase is found in both gazetteers, it is stored as a toponym. For the extracted toponyms, we retain only records for which (1) the OSM and Geonames records have same name, and (2) the Geonames' point-based coordinates are inside the region-based representation of their corresponding OSM records.

The toponym disambiguation is then undertaken based on the *minimum spatial*

[2]More information about the dataset can be found in: `https://github.com/dfcf93/MSMARCO`

context heuristic (Leidner, Sinclair, & Webber, 2003). We use bounding boxes to determine which combination of the geographic locations satisfy the minimum spatial extent condition. In cases of duplicate places in GeoNames which lead to the same bounding boxes, the combination with more specific place types is selected. For example, populate place (PPL) is a place type in GeoNames which could be a village, city, state and even country. Hence, administrative divisions (e.g., state) are chosen over the populated places. Finally, if the toponym ambiguity still exists, we use importance value to select the more prominent combination.

More sophisticated heuristics in toponym disambiguation (e.g., Lieberman & Samet, 2012; Wang et al., 2010) are not used due to reliance on significant assumptions – e.g., the relation between place types in the toponyms, city-country relation. These heuristics constrain the relationships between type, scale and prominence of resolved places in the text. This may impact the results of this study and lead to stronger associations based on type, scale and prominence between toponyms in questions and answers. Here, to present fair results, we avoid using these disambiguation methods.

3.3. Encoding

The gazetteers' attributes for the extracted toponyms have been used as proxies to capture type, scale and prominence of the toponyms in questions and answers. Using these proxies, the sequence representations for each question-answer pair are encoded into TSP sequences.

Type: The Geonames type schema[3] has been used without modification to encode generic place types. This schema contains 667 unique types of geographic features, covering both natural and man-made geographic types.

Scale: To capture scale, we have extended the schema from Richter et al. (2013). This schema contains seven levels of granularity: (1) furniture, (2) room, (3) building, (4) street, (5) district, (6) city, and (7) country. We have extended the coarse levels of granularity by adding *county*, *state*, *country*, and *continent*, and removing the *furniture* and *room* levels from the schema. OSM records include an attribute related to the OSM definition of scale (i.e., place_rank[4]), a number between 0 to 30. We convert the extracted gazetteers' records into the appropriate scale level based on a look-up table that maps OSM scale levels into the proposed scale schema (see the supplementary material, Section 1.1).

Prominence: To capture prominence, the *importance* attribute in the extracted OSM Nominatim record is used. The OSM importance value is estimated using Wikipedia importance score (Thalhammer & Rettinger, 2016) with some minor tweaks[5]. The value is defined between 0 and 1, and it is designed to be used for ranking search results. We translate these values into seven finite levels of prominence, derived by *natural breaks* classification (Jenks, 1967) of the frequency spectrum of the values.

3.4. Distribution Analysis and Rule Mining

Distribution analysis and rule mining techniques are used to extract and investigate patterns in the human-generated answers and the relation between the questions and

[3] https://www.geonames.org/export/codes.html

[4] https://wiki.openstreetmap.org/wiki/Nominatim/Development_overview

[5] https://lists.openstreetmap.org/pipermail/geocoding/2013-August/000916.html

their answers. Distributions of type, scale and prominence sequences are used to compare the questions and answers. To derive patterns in the questions and their answers, association rule mining, a-priori algorithm (Agrawal & Srikant, 1994), is used.

The strength of the extracted rules are evaluated using the standard measures – i.e., *support*, *confidence*, and *lift*. Support defines how frequently an association rule is observed in the whole dataset, and confidence determines how often the rule is true. Lift is a measure to evaluate the importance of the rules – i.e., lift greater than one shows positive and strong dependency among the elements of the extracted rule. This part of the method is devised to test the first and second hypotheses.

3.5. Prediction

The input for the prediction is an encoded sequence of TWQs, and the output is the generic sequence of their corresponding answers. The problem can then be formulated as a sequence prediction from concatenated generic sequences for the questions and their answers, where a part of a sequence is known, and the rest is predicted. Table 1 shows the sequence prediction methods which are used in this study. We used and extended an open-source toolkit for sequence analysis (Fournier-Viger et al., 2016) to implement the prediction methods.

These classic methods are divided into probabilistic (Cleary & Witten, 1984; Padmanabhan & Mogul, 1996; Pitkow & Pirolli, 1999) and non-probabilistic categories (Gueniche, Fournier-Viger, Raman, & Tseng, 2015; Gueniche, Fournier-Viger, & Tseng, 2013; Laird & Saul, 1994; Ziv & Lempel, 1978). The probabilistic methods are based a graph representation of conditional probabilities (Cleary & Witten, 1984) or Markov chain's transition probability matrix (Padmanabhan & Mogul, 1996; Pitkow & Pirolli, 1999) of the sequence elements. The non-probabilistic methods compress the sequences in a lossy (Laird & Saul, 1994; Ziv & Lempel, 1978) or lossless approaches (Gueniche et al., 2015, 2013) into tree-based (Gueniche et al., 2015, 2013) or graph-based (Laird & Saul, 1994) data structures (for a review of sequence prediction methods see Tax, Teinemaa, and van Zelst (2020)).

The structure of sequence and the relation between prior elements in the sequence to their succeeding elements are trained into a model. The model is then tested on an unseen part of data using K-fold cross validation (K=10). We considered two baseline methods to evaluate the performance of the sequence prediction methods: (1) random sequence generation and (2) most frequent pattern.

The random generation baseline only utilizes the schema of type, scale and prominence without any information about the distributions of values in the answers. The most frequent patterns baseline predicts templates of answers using the schema and the distribution of generic references in the answers. The difference between the prediction performances of random generation and the most frequent patterns shows the impacts of using the distribution of generic values in generating templates of answers (see hypothesis 1). The sequence prediction methods also consider the relation between generic values in the questions to their answers. Consequently, the improvement in generating the templates compared to the most frequent patterns baseline is related to the association between generic values of questions and their answers (hypothesis 2).

In prediction, each generic form of questions is used to predict the same generic form of their answers. In addition, we have devised an approach to predict one of the generic forms of an answer using all generic forms (i.e., type, scale and prominence)

Table 1. Sequence prediction methods

Method	Publication	**Year**
Lampel-Ziv 1978 (LZ78)	(Ziv & Lempel, 1978)	1978
First order Markov Chains (Mark1)	(Cleary & Witten, 1984)	1984
Transition Directed Acyclic Graph (TDAG)	(Laird & Saul, 1994)	1994
Dependency Graph (DG)	(Padmanabhan & Mogul, 1996)	1996
All-k-Order Markov Chains (AKOM)	(Pitkow & Pirolli, 1999)	1999
Compact Prediction Tree (CPT)	(Gueniche et al., 2013)	2013
Compact Prediction Tree Plus (CPT+)	(Gueniche et al., 2015)	2015

of its corresponding question. Algorithm 1 shows the process to use all three type/scale/prominence sequences to predict a generic form of the answers in each generic class. Here, each combination of type, scale and prominence values are mapped to a unique code. Using these codes, a new sequence is generated for each question/answer to capture type, scale and prominence together. Next, these sequences are used to predict the generic form of answers. Finally, a reverse mapping is used to decode these sequences into type, scale and prominence sequences.

Algorithm 1 Training and prediction based on type-scale-prominence together

```
1: procedure TSP_Prediction(type, scale, prominence)
2:     generate a code for each unique combination of type-scale-prominence (TSP)
3:     create encoded sequences based on generated TSP codes
4:     train a model to predict TSP in answers based on TSP in the questions
5:     for every question do
6:         given a question (TSP); predict the answer (TSP)
7:         decode the predicted answer (TSP) to answer (type/scale/prominence)
8:         if multiple predictions are allowed then
9:             avoid counting duplicate decoded values for type/scale/prominence
10:        end if
11:    end for
12: end procedure
```

4. Results

4.1. Extraction and Encoding

The assessment of toponym extraction, finding TWQs, and categorizing the questions into SWQs and DWQs are presented in Table 2. Here, average precision and recall of the extraction results are calculated using manually annotated data (5% of TBWQs and their answers). For the task of finding TWQs in the dataset, the *false negatives* (TWQs that have not been extracted) are not investigated, hence the recall is unknown.

As shown in Table 2, 6,274 TWQs and their answers are found in the dataset. The TWQs are approximately 11.1% of the *location questions* of the dataset. For evaluation, 5% of extracted TWQs (314 questions) are investigated and the precision of extraction is 91.7% – i.e., 288 of 314 extracted questions are TWQs. Using the 288 TWQs, the precision and recall of extracting toponyms and classifying the questions to SWQs and DWQs are presented in Table 2.

Table 2. Extraction evaluation

Extraction	#Extracted	#Investigated	Precision	Recall
TWQs	6274	314 (5%)	91.7% (288 out of 314)	–
SWQs	3285	121 out of 288	89.4%	90.2%
DWQs	2989	167 out of 288	92.7%	92.1%
Toponyms	22307	1133[6]	88.6%	90.8%

Table 3. Encoding results

Encoding	#TWQs	#SWQs	#DWQs
Type sequences	6,274	3,285	2,989
Scale sequences	3,936	1,985	1,951
Prominence sequences	6,051	3,098	2,953

Table 3 shows the number of records that are completely encoded for question-answer pairs in type, scale and prominence sequences. Here, if even the information for one place (which is mentioned either in the question or its answer) is missing, the question and its answer are not used to extract patterns or test the predictability of generating generic form the answer. As shown in the table, the encoding into scale and prominence is not always possible due to incompleteness of attribute information (i.e., *place_rank* and *importance*) in OSM Nominatim.

4.2. Distributions

The distribution of TWQs[7] and their answers based on type, scale and prominence are shown in Figures 3, 4 and 5. Figure 3 shows that the diversity of types in the questions is higher than in the answers. While administrative divisions are more frequent than other generic types in both questions and answers, they are more dominant in the answers.

Figure 4 shows the scale in the answers is systematically one level coarser than in the questions. In addition, the distribution shows that city-level and state-level scales are frequently observed in the questions, while the answers mostly contain references of county and country levels of scale. The results further show that the coarsest level of scale (i.e., continent level) is rarely observed in the answers. This observation shows an answer at the continent level would be under-specified in most cases, and therefore uninformative.

The distributions of prominence levels in questions and answers are similar to the distributions by scale (Figure 5). In the questions, we observe a bi-modal distribution of levels of prominence in the content of questions. The distribution of prominence in the answers, however, shows that higher levels are dominant. In contrast to the distributions by scale, the most prominent level is dominant in the answers. Hence,

[7]A detailed comparison of SWQs and DWQs is presented in the supplementary material (Section 2)

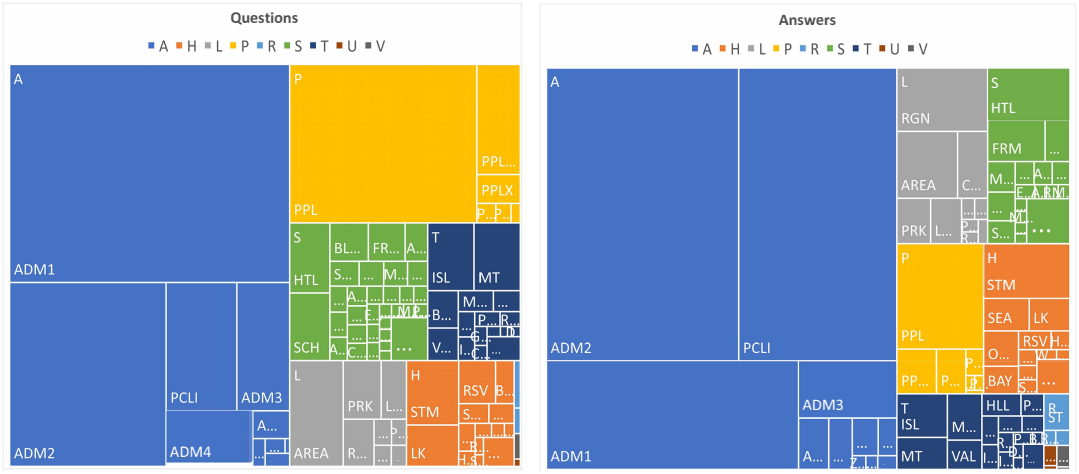

Figure 3. Distribution of place types in the questions and in the answers.

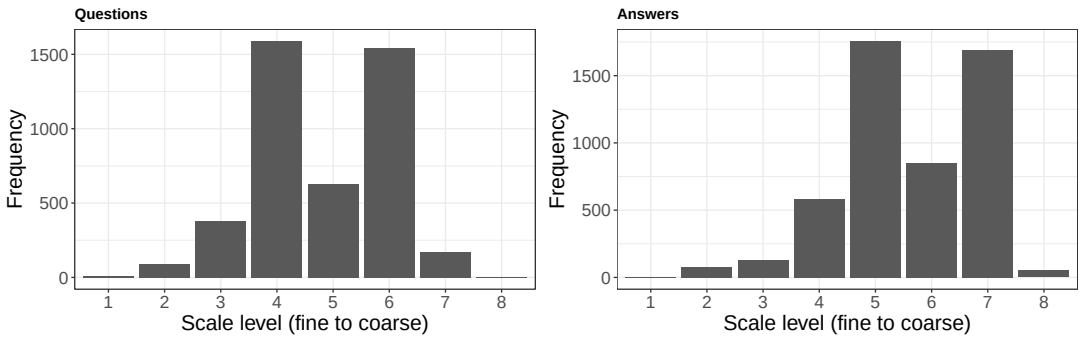

Figure 4. Distribution of levels of scale in all toponym-based where questions and answers.

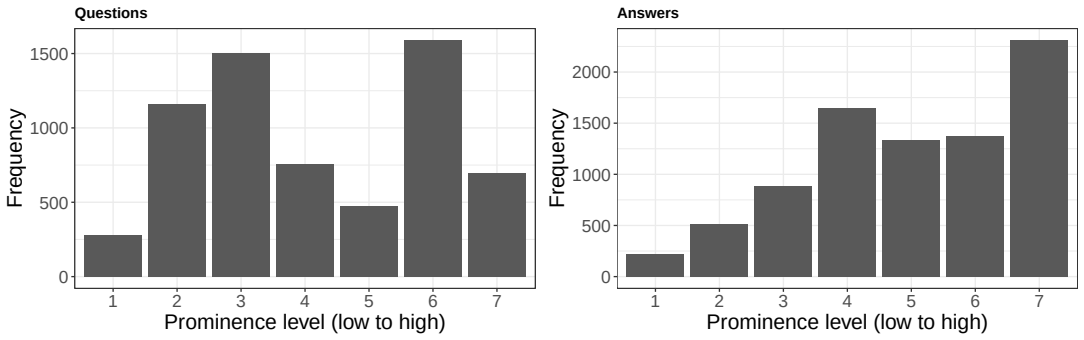

Figure 5. Distribution of prominence levels in the questions versus answers

Table 4. Extracted rules from type sequences

Rank	rule	support	confidence	lift	frequency
	Simple where-questions				
1	{A-ADM2} =>{A-ADM1}	0.15	0.52	1.28	478
2	{Q-ADM1} =>{A-PCLI}	0.08	0.74	1.69	259
3	{A-ADM1,A-ADM2} =>{A-PCLI}	0.08	0.54	1.24	259
4	{Q-ADM2} =>{A-PCLI}	0.06	0.52	1.21	188
5	{Q-PPL,A-ADM2} =>{A-PCLI}	0.04	0.54	1.23	112
	Detailed where-questions				
1	{Q-ADM1} =>{A-ADM2}	0.57	0.76	1.12	1701
2	{Q-ADM1} =>{A-PCLI}	0.38	0.50	1.13	1126
3	{A-PCLI} =>{A-ADM2}	0.35	0.79	1.17	1053
4	{A-ADM2,Q-ADM1} =>{A-PCLI}	0.31	0.54	1.21	916
5	{Q-PPL} =>{A-ADM2}	0.22	0.78	1.15	656

people tend to refer to well-known places in their answers. Unlike with scale, the highest levels of prominence do not necessarily lead to obvious or irrelevant answers[8].

4.3. Extracted Rules

To test Hypotheses 1 and 2 (see Section 1.4), we extract strong rules in the encoded pairs of questions-answers through association rule mining. The association rules extracted from the answers can be used to describe how answers are constructed in detail (Hypothesis 1). The relationship between the content of the questions and their answers can thus also be further investigated (Hypothesis 2).

Tables 4-6 show the top five extracted rules (based on *frequency*/*support*) for type, scale and prominence, respectively. In the tables, the values starting with *Q*- relate to the contents of the questions and the values starting with *A*- to the content of the corresponding answers. As shown in the tables, some rules describe the structure of answers (e.g., {A-ADM1, A-ADM2} =>{A-PCLI}) while the others describe the relationships between questions and answers (e.g., {Q-ADM2} =>{A-PCLI}).

[8]A detailed analysis of sequence distributions is available in supplementary material (Section 3)

Table 5. Extracted rules from scale sequences

Rank	Rule	support	confidence	lift	frequency
	Simple where-questions				
1	{Q-6} =>{A-9}	0.21	0.55	1.01	417
2	{Q-6} =>{A-8}	0.20	0.54	1.24	404
3	{A-7} =>{A-9}	0.16	0.56	1.73	307
4	{Q-6} =>{A-7}	0.15	0.54	1.37	295
5	{A-7} =>{A-8}	0.15	0.54	1.25	293
	Detailed where-questions				
1	{Q-8} =>{A-7}	0.65	0.80	1.08	1277
2	{Q-6} =>{Q-8}	0.49	0.81	0.99	952
3	{Q-6} =>{A-7}	0.48	0.80	1.07	940
4	{A-9} =>{Q-8}	0.45	0.87	1.07	887
5	{A-7,Q-6} =>{Q-8}	0.42	0.88	1.07	823

Table 6. Extracted rules from prominence sequences

Rank	Rule	support	confidence	lift	frequency
	Simple where-questions				
1	{A-4} =>{A-7}	0.14	0.54	1.09	425
2	{A-5} =>{A-7}	0.13	0.50	1.02	417
3	{Q-3} =>{A-7}	0.12	0.52	1.05	382
4	{Q-6} =>{A-7}	0.08	0.58	1.18	260
5	{Q-4} =>{A-7}	0.08	0.53	1.07	250
	Detailed where-questions				
1	{Q-6} =>{A-4}	0.32	0.56	1.19	957
2	{Q-6} =>{A-7}	0.30	0.51	1.06	884
3	{A-4} =>{A-7}	0.25	0.54	1.11	742
4	{Q-3} =>{Q-6}	0.24	0.54	0.93	695
5	{Q-3} =>{A-7}	0.22	0.51	1.04	650

Table 4 shows the dominant role of administrative divisions in the human-generated answers. Association rules extracted based on the scale (Table 5) show the *greater-than* and *between* levels of the answers to SWQs and DWQs. The top five patterns of answers are mostly constructed with references to the highest level of prominence (A-7). This shows the major impact of prominence in human answering behavior to where-questions – i.e., people refer to prominent places in answering where-questions.

Tables 4, 5 and 6 show that stronger association rules with higher support are extracted from DWQs in comparison to SWQs. The rules show strong associations between antecedent and consequent parts of the extracted rules with lift value greater than one. The results show that stronger rules with higher confidence and support are extracted using *scale* in comparison to *type* and *prominence*.

The tables only present the extracted rules with highest frequency and support. These tables show how a small set of generic rules describes a large proportion of data in the MS MARCO dataset. Sorting rules by confidence or lift will change the order of the rules. For example, the maximum lift (equal to *8.93*) in the extracted rules belongs to {Q-6, Q-9} =>{A-8} for detailed-where questions using scale. The frequency of this rule is *43*, and it describes the relevant scale level (between minimum and maximum levels of the question) for detailed where-questions. The maximum confidence is 0.93 for detailed where-questions encoded by type. This association rule is {Q-PPLA2, Q-ADM1} =>{A-ADM2} with a frequency *109*. This rule describes that *populated places* in detailed where-questions are mostly localized by referring to *counties* they belong to.

4.4. Predicting the Generic Form of Answers

We test the predictability of the generic sequence of an answer given the generic sequence of the corresponding question. We investigate different prediction scenarios, including (1) the same generic class prediction (e.g., predicting type sequence of answers using type sequence of questions), and (2) prediction of one generic class using all generic classes (e.g., predicting type sequence of answers using type/scale/prominence sequences of questions, see Algorithm 1).

We assess the prediction accuracy of *content* and *content-and-style* of the answers (defined in Section 2.2.1). Referring to the introductory example, if type sequence of the answer is predicted as [river, city] then it is captured as a correct prediction for both content and content-and-style. The other permutation of this sequence (i.e., [city, river]) is considered as a correct prediction of content and incorrect prediction for content-and-style. Evidently, any other type sequence is an incorrect prediction for both content and content-and-style scenarios.

Each prediction scenario is applied over all questions, SWQs and DWQs to investigate the impacts of *question types* on the prediction accuracy. Each scenario is tested using all six sequence prediction methods and is compared with the two baseline approaches (i.e., random generation and most frequent patterns). Only the best prediction performances among the six sequence prediction methods are presented. The *best performance* is the maximum prediction accuracy achieved by one of the methods for a prediction scenario. We also test prediction accuracy when multiple predictions are allowed – i.e., *top-k predictions* for k from one to five. In top-k predictions, k unique sequences are predicted for each answer and if one of the sequences matches with the generic form of the answer then the prediction is successful.

Table 7 shows the best performances in predicting type sequences of answers. The

Table 7. Prediction accuracy for type sequences

#Predictions (k)	Content		Content and Style	
	Type → Type	TSP → Type	Type → Type	TSP → Type
All questions				
1	45.2	55.7	29.0	40.7
2	68.9	77.1	44.6	60.5
3	80.2	83.3	57.8	73.3
4	83.6	84.7	64.0	76.1
5	84.4	85.5	68.3	77.4
Simple where-questions				
1	39.5	47.5	14.2	27.4
2	60.8	69.4	32.7	48.5
3	73.2	75.8	48.2	63.4
4	77.2	77.5	58.1	66.2
5	78.5	78.2	63.1	67.0
Detailed where-questions				
1	59.1	67.3	47.1	59.6
2	80.4	88.7	61.3	76.3
3	84.0	91.2	65.9	84.4
4	88.0	91.3	73.6	86.4
5	88.5	92.1	75.6	87.1

prediction accuracy based of TSP sequences is noticeably higher than that of predictions using only type sequences. This shows a complementary role of scale and prominence in predicting type sequence of the answers.

Contrasting DWQs and SWQs shows that extra details in DWQs are useful for prediction of the generic form of answers. In addition, we observe how subjectivity in style of answers and flexibility of language to convey information lead to noticeable less accuracy in prediction of content-and-style of answers in comparison to prediction of content. This observation is related to the flexibility of natural language, in which the same meaning can be presented in different ways. Finally, the number of predictions (k in the table) shows that the accuracy dramatically increases in the case of multiple predictions.

Tables 8 and 9 show that compared to type sequence prediction, the TSP sequences contribute less effectively in predicting the prominence and scale sequences – i.e., only slightly improve the prediction accuracy. When considering multiple predictions, TSP sequences lead to worse results than prominence sequences or scale sequences alone. This can be explained by overfitting to specific patterns in the training dataset. Here, overfitting is observed because the schema of types is more than 20 times larger than the scale and prominence schemas. Hence, using type in prediction of scale or prominence leads to very detailed patterns that are not generalizable enough and decrease the prediction accuracy on unseen data. Finally, scale is the most predictable, and prominence is the least predictable generic class. Similar to the observations based on type prediction performances, DWQs are more predictable than SWQs based on scale and prominence.

Table 8. Prediction accuracy for scale sequences

#Predictions (k)	Content		Content and Style	
	Scale → Scale	TSP → Scale	Scale → Scale	TSP → Scale
All questions				
1	55.0	56.7	38.2	42.2
2	79.4	79.2	61.0	62.8
3	91.6	86.1	79.0	76.0
4	96.3	88.7	92.0	81.9
5	98.0	89.3	96.0	83.5
Simple where-questions				
1	48.5	49.5	20.4	28.6
2	79.6	71.8	49.1	49.8
3	89.9	78.3	71.9	67.0
4	95.6	81.8	90.3	74.0
5	97.5	82.6	94.9	75.5
Detailed where-questions				
1	69.6	68.2	59.8	60.6
2	88.4	89.6	78.4	77.3
3	95.8	93.3	88.6	87.1
4	97.5	95.2	94.8	92.1
5	98.6	95.2	97.0	92.7

Table 9. Prediction accuracy of prominence sequences

#Predictions (k)	Content		Content and Style	
	Prominence → Prominence	TSP → Prominence	Prominence → Prominence	TSP → Prominence
All questions				
1	50.8	53.0	19.9	30.7
2	74.1	73.4	39.2	49.1
3	85.0	81.9	61.6	66.4
4	92.1	86.7	79.2	77.1
5	96.1	88.6	89.2	81.8
Simple where-questions				
1	45.4	45.6	14.3	19.5
2	75.4	69.4	34.9	39.1
3	84.7	77.0	54.5	56.9
4	91.3	80.5	73.7	68.2
5	95.6	81.9	87.9	72.7
Detailed where-questions				
1	53.3	58.2	26.9	43.8
2	75.1	80.0	50.9	60.4
3	86.0	88.9	70.9	78.4
4	93.1	93.0	82.5	87.0
5	96.8	95.4	91.9	92.6

Table 10. Accuracy improvement using sequence prediction compared to the baselines

Prediction Scenario	Random	Most Frequent Pattern(s)
Type → Type	+48.9%	+18.3%
Scale → Scale	+58.1%	+27.6%
Prominence → Prominence	+39.2%	+30.4%
TSP → Type	+61.6%	+31.0%
TSP → Scale	+54.1%	+23.6%
TSP → Prominence	+42.3%	+33.5%
Overall	+50.7%	+27.4%

Table 11. RMSE of sequence prediction methods

Prediction Scenario	LZ78	Mark1	TDAG	DG	AKOM	CPT	CPT+
Type	7.4%	15.2%	**21.8%**	13.4%	17.3%	**7.1%**	12.9%
Scale	9.8%	12.5%	**17.9%**	10.6%	14.3%	**5.7%**	11.8%
Prominence	8.7%	13.3%	**19.2%**	9.9%	15.2%	**4.9%**	11.5%
Content	8.9%	15.5%	**22.7%**	9.1%	17.4%	**1.9%**	10.2%
Content and Style	8.6%	11.7%	**16.1%**	13.2%	13.7%	**8.2%**	13.8%

Table 10 shows the improvement of accuracy in best prediction performances compared to two baselines – i.e., random generator, and most frequent pattern(s). The minimum improvement is +18.3% in prediction of type sequences of answers using type sequences of questions in comparison to the most frequent pattern(s). This observation shows that strong patterns exist in the distributions of answers and consequently, the baseline method performs well in prediction of type sequences of answers. The strongest improvement is +61.6% when comparing the best predictive performance of type sequences using type/scale/prominence sequences together, compared to the random baseline. This is because of the large number of distinct types in type schema that lead to false predictions for the random baseline. The accuracy improvements illustrate the strong relationship between the generic content of questions and generic content of their answers.

To compare the sequence prediction methods, we used the difference between the prediction accuracy of each method to the best performance achieved by all methods for each prediction scenario. Table 11 shows the root mean square error (RMSE) for each sequence prediction method. The RMSE shows how well-performed a method is in comparison to other methods. If the RMSE of a method is lower than others, the prediction accuracy of the method is higher than the others. The prediction scenarios in Table 11 are simplified groups of actual predictions. For example, prediction scenario of scale is related to predicting scale sequences of answers using (1) scale sequence of questions or (2) type/scale/prominence sequences of questions.

As shown in Table 11, in all scenarios the **CPT** method is the *best* performing method and **TDAG** performs *worst* based on the RMSE values. The results suggest that **CPT** is the best method to construct predictive models to predict the generic form of answers.

5. Demonstration: From Generic to Specific

Translating generic encoding of answer to specific form (e.g., type sequence to toponym sequence) is the last phase in the proposed approach. Our approach to the generic-to-specific translation problem is grounded in the following assumption: *places mentioned in the questions have relationships to places referred to in their answers, and these relations can be found in a knowledge base.* In addition, the specific form of questions and generic form of answers are available through encoding and prediction, respectively. Based on this assumption and the available information, the specific form of answer can be derived using a SPARQL query template (Query 1). While the *structure* of a suitable knowledge base for this purpose has been studied before by Chen, Vasardani, Winter, and Tomko (2018), no such knowledge base is yet available with the definitions of type, scale and prominence as used in this study. Hence, the translation is only demonstrated here using the introductory example[9].

We have used DBPedia and Geonames as sources to demonstrate how SPARQL queries can be used to find specific forms of answers. Considering the information stored in DBPedia and Geonames, this demonstration is limited to type sequences of the answers because the prominence and scale are not available in the place ontology of these knowledge bases. Even the type schema used in DBPedia is different from the Geonames' type schema, and consequently in the following example, mapping to the DBPedia type schema is done manually.

```
PREFIX [KNOWLEDGE BASE]

SELECT distinct ?question ?answer
WHERE {
    VALUES ?question [SPECIFIC] .
    ?answer a [GENERIC] .
    {?question ?r ?answer} UNION {?answer ?r ?question} .
}
```

Query 1 SPARQL template

Referring to the introductory example, the where-question and its answer is modelled as follows:

- specific representation (question): [Putney Bridge];
- TSP encoding (question): type sequence [BDG], scale sequence [4], prominence sequence [3];
- TSP encoding (answer): type sequence [STM, ADM2], scale sequence [6, 6], prominence sequence [6, 7];
- specific representation (answer): [River Thames, London]

The SPARQL queries for finding the specific forms of answers are presented in Queries 2 and 3 using DBPedia and Geonames ontologies. The results of these queries are shown in Table 12. Using DBPedia, the generic forms are correctly translated into River Thames and London. However, the generic to specific translation using Geonames is partially successful. In Geonames, places are conceptualized as points and it supports only containment. This example shows that point-based conceptualization of places is not sufficient for generic to specific translation and more diverse support of spatial relationships can be useful to find the correct specific forms.

[9]More examples are provided in the supplementary material (Section 4)

Table 12. SPARQL results to find specific form of the answer

Knowledge Base	Q1	A1	A2
DBPedia	Putney Bridge	**London**	**River Thames**
Geonames	Putney Bridger	**London**	–

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT distinct ?q1 ?a1 ?a2 WHERE {
  VALUES ?q1 {<http://dbpedia.org/resource/Putney_Bridge>}

  ?a1 a dbo:PopulatedPlace .
  {?a1 ?r1 ?q1} UNION {?q1 ?r1 ?a1} .

  ?a2 a dbo:River .
  {?a2 ?r2 ?q1} UNION {?q1 ?r2 ?a2} .
}
```

Query 2 SPARQL query of the example (DBPedia)

```
PREFIX gn: <http://www.geonames.org/ontology#>

SELECT distinct ?q1 ?a1 ?a2 WHERE {
  VALUES ?q1 {<http://sws.geonames.org/6619925/>}

  ?a1 gn:featureCode gn:A.ADM2 .
  {?a1 ?r1 ?q1} UNION {?q1 ?r1 ?a1} .

  ?a2 gn:featureCode gn:H.STM .
  {?a2 ?r2 ?q1} UNION {?q1 ?r2 ?a2} .
}
```

Query 3 SPARQL query of the example (Geonames)

6. Discussion

The results of the proposed method shows how generic information can be used to characterize and imitate human answering behavior to generate templates for answering the questions. While the results are limited to the human-search engine interaction, the proposed methodology (specific-generic translation) is flexibly defined to be applicable to other QA scenarios as well.

We have used type, scale and prominence as generic classes to investigate MS MARCO dataset. We have compared their potentials in describing human answering behavior and their performances in predicting the generic forms of the answers. As a result, two major observations are reported.

First, while strong patterns for each generic class have been observed, we find that *scale* is the most predictive class. This is because where-questions are a specific subset of spatial questions and scale directly captures inherent spatial aspect of places. Meanwhile, in the notions of type and prominence, other aspects of places contribute as well – e.g., functional and physical aspects. In addition, scale is a generic class that captures hierarchical relationships between places, and previous studies show that these relationships are the basis for answering where-questions (Shanon, 1983). Moreover,

we have observed that type is performing better than prominence in both characterizing and predicting human-generated answers. This observation is highly influenced by the proxies used to capture type and prominence.

Second, when comparing SWQs and DWQs, our investigation shows that the generic templates to answer to DWQs, compared to SWQs, can be generated more accurately. We find stronger rules and patterns in the answers to DWQs than in answers to SWQs. This is because DWQs contain richer details which helps narrowing down the list of possible relevant answers. To illustrate this point, two examples are provided (1) *Where in California is Beverly Hills?* and (2) *Where is Beverly Hills?* In the first question, the list of possible relevant answers is narrowed down to *Los Angeles County* because the inquirer already knows it is in *California*. For the latter, respondents are free to subjectively guess the state of the inquirer's geographic knowledge and provide answers such as *Los Angeles County*, *California*, and *United States*.

Theoretical limitations: The instruction for specific-generic approach is devised in a flexible manner to be usable in different GeoQA scenarios. However, utilizing the approach needs a careful design (e.g., selecting appropriate list of generic classes) to fit for a particular scenario. The proposed TSP encoding is limited to the QA scenario of general Web search and may not be suitable for other QA scenarios such as human interaction with autonomous cars. In short, the theoretical limitations of this study are:

(1) The generic-to-specific translation approach is only focused on where-questions, and other types of geographic questions are neglected.
(2) The proposed approach is focused only on the questions and their relationship with the answers, when no other contextual information about inquirers is available.
(3) The approach is designed with an exclusive focus on toponyms while qualitative spatial relationships have an important role in answering where questions.
(4) The additional impacts of qualitative spatial relationships (e.g., *in southern part of*) as modifiers of scale are neglected in the TSP encoding.

Results limitations: There are some limitations to the implementation presented in this study:

(1) The biases of the MS MARCO dataset directly influence our results. The data are extracted from the Microsoft Bing search engine, and hence the results are necessarily biased to the questions asked by users of this search engine. In addition, the sampling approach used when extracting MS MARCO questions from the MS Bing query logs may have a direct and unquantifiable impact on the generality of the results.
(2) The results are influenced by the geographic biases and incompleteness of data in Geonames and OSM Nominatim. The bias and incompleteness of gazetteers are well-documented by Acheson, Sabbata, and Purves (2017).
(3) The bias in the proxies that have been used to capture the TSP encoding also have an impact on the results.

Despite these limitations, the identified patterns align well with everyday experience and provide a grounding for answering where-questions.

7. Conclusions

Generating responses with a similar quality to human-generated answers is a challenge to current search engines and QA systems. In particular, where-questions are hard to answer because the responses can sometimes be either vague or obvious to the inquirers. To avoid generating ambiguous or obvious responses or retrieving unnecessary information as a part of the answers, a proper set of anchor places must be identified to localize the place in question. The assumption that answers to where-questions can be found completely, without any further modification, inside a textual document or as a node or its properties in a knowledge base may not hold in general. Consequently, we introduced here an approach to generate templates to answer where-questions based on relevant pieces of information.

The approach is based on the automatic extraction of patterns of generic geographic forms from human-generated QA. These captured in predictive models and are used to generate templates of answers similar human-generated responses. Three generic classes (i.e., type, scale and prominence) are used to investigate the properties of the anchor places in human-generated answers. We have used questions and answers from MS MARCO v2.1, an extensive dataset constructed from questions submitted to a general-purpose search engine.

Using distribution analysis and rule mining techniques, we have identified the characteristics and recurrent patterns in the questions and their answers (Hypotheses 1 and 2). We have then applied sequence prediction methods to generate the generic forms for answers based on the generic forms of the corresponding questions (Hypothesis 3). We have also briefly sketched an approach how such generic forms may help with the generation of the appropriate answers, based on the information available in the knowledge bases.

The results show that the prediction of answer structures based on *scale* is more precise, compared to predictions relying on *type* and *prominence*. The rules extracted based on scale have higher support and confidence than the rules extracted from type or prominence. We also observe how the type of questions (i.e., SWQs vs. DWQs) influence the strength of the extracted rules and lead to noticeable differences in prediction performances. Finally, we compared different sequence prediction methods and find that CPT (Gueniche et al., 2013) is the best performing approach in all scenarios. However, the results of this study are limited to human interaction with a general purpose search engine. Consequently, an important future direction of this research is to investigate other corpora of QA related to different scenarios – e.g., human-human dialogue.

We have also observed that the neglect of qualitative spatial relationships in our encoding and prediction mechanism may present a a major theoretical shortcoming of the proposed specific-generic translation. Consequently, developing a more sophisticated encoding is necessary to extract a deeper understanding of the human answering behavior of where-questions.

Developing an automatic approach to decode generic forms of answers into specific representations (i.e., toponyms) is a necessary step to complete the process of the specific-generic translation approach. Available information in documents or knowledge bases can be used to derive the specific representations. Another important future direction is to investigate how the proposed approach can be combined with current personalization methods, in order to adapt answers to specific inquirers and their context. Finally, investigation of other types of where-questions (i.e., where-questions with generic references) and their human-generated answers using specific-generic translation remains as a future work.

8. Data and Codes Availability Statement

This study makes use of a third-party data source, MS MARCO v2.1 (Nguyen et al., 2016). The dataset is freely available under a proprietary agreement for non-commercial use[10]. The computational workflow of this publication is implemented in Java and R. The implementation is available under the MIT License[11] and accessible in an anonymous FigShare repository (only for review): `https://figshare.com/s/a317393764c7f443cdcd`.

References

Acheson, E., Sabbata, S. D., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, *64*, 309 - 320.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487–499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ballatore, A. (2019). A context frame for interactive maps. In *AGILE Conference on Geographical Information Science: Short Papers* (p. 1—5).

Buscaldi, D., Benajiba, Y., Rosso, P., & Sanchis, E. (2006). The UPV at QA@ CLEF 2006. In *CLEF (Working Notes)*.

Chen, H., Vasardani, M., Winter, S., & Tomko, M. (2018, jun). A Graph Database Model for Knowledge Extracted from Place Descriptions. *ISPRS International Journal of Geo-Information*, *7*(6), 221.

Chen, W., Fosler-Lussier, E., Xiao, N., Raje, S., Ramnath, R., & Sui, D. (2013). A synergistic framework for geographic question answering. In *Proceedings of IEEE 7th International Conference on Semantic Computing* (p. 94-99).

Church, K., Neumann, J., Cherubini, M., & Oliver, N. (2010). The "map trap"? an evaluation of map versus text-based interfaces for location-based mobile search services. In *Proceedings of the 19th international conference on world wide web* (p. 261–270). New York, NY, USA: Association for Computing Machinery.

Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE transactions on Communications*, *32*(4), 396–402.

Couclelis, H., Golledge, R., Gale, N., & Tobler, W. (1987). Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology*, *7*(2), 99 - 122.

Edwardes, A. J., & Purves, R. S. (2007). A Theoretical Grounding for Semantic Descriptions of Place [Conference Proceedings]. In J. M. Ware & G. E. Taylor (Eds.), *Web and Wireless Geographical Information Systems* (pp. 106–120). Springer Berlin Heidelberg.

Ferrés, D., & Rodríguez, H. (2006). Experiments adapting an open-domain question answering system to the geographical domain using scope-based resources. In *Proceedings of the Workshop on Multilingual Question Answering* (pp. 69–76). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ferrés, D., & Rodríguez, H. (2010). TALP at GikiCLEF 2009. In C. Peters et al. (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments* (pp. 322–325). Berlin, Heidelberg: Springer Berlin Heidelberg.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)* (pp. 363–370).

Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., ... Tresp, V. (2016). The SPMF Open-Source Data Mining Library Version 2. In *Machine Learning*

[10] `http://www.msmarco.org/dataset.aspx`

[11] `https://opensource.org/licenses/MIT`

and Knowledge Discovery in Databases (pp. 36–40). Cham: Springer International Publishing.

Gueniche, T., Fournier-Viger, P., Raman, R., & Tseng, V. S. (2015). CPT+: Decreasing the time/space complexity of the compact prediction tree. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 625–636).

Gueniche, T., Fournier-Viger, P., & Tseng, V. S. (2013). Compact prediction tree: A lossless model for accurate sequence prediction. In *International Conference on Advanced Data Mining and Applications* (pp. 177–188).

Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S., & Tomko, M. (2019). Place questions and human-generated answers: A data analysis approach. In *Geospatial Technologies for Local and Regional Development* (pp. 1–16).

Hamzei, E., Winter, S., & Tomko, M. (2019). Initial analysis of simple where-questions and human-generated answers. In *Proceedings of Short Papers at the 14th International Conference on Spatial Information Theory.* Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, *7*, 186–190.

Jiang, B., & Yao, X. (2006). Location-based services and gis in perspective. *Computers, Environment and Urban Systems*, *30*(6), 712 - 725. (Location Based Services)

Laird, P., & Saul, R. (1994). Discrete sequence prediction and its applications. *Machine Learning*, *15*(1), 43–68.

Leidner, J. L., Sinclair, G., & Webber, B. (2003). 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* (pp. 31–38). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lieberman, M. D., & Samet, H. (2012). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 731–740). New York, NY, USA: ACM.

Luque, J., Ferrés, D., Hernando, J., Mariño, J. B., & Rodríguez, H. (2006). GEOVAQA: a voice activated geographical question answering system. In *Iv jornadas en tecnologia del habla* (pp. 309–314).

Mai, G., Janowicz, K., He, C., Liu, S., & Lao, N. (2018). POIReviewQA: A Semantically Enriched POI Retrieval and Question Answering Dataset. In *Proceedings of the 12th Workshop on Geographic Information Retrieval.* New York, NY, USA: Association for Computing Machinery.

Mai, G., Yan, B., Janowicz, K., & Zhu, R. (2020). Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Mansourian (Eds.), *Geospatial Technologies for Local and Regional Development* (pp. 21–39). Cham: Springer International Publishing.

Mishra, A., Mishra, N., & Agrawal, A. (2010, Nov). Context-aware restricted geographical domain question answering system. In *Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks* (p. 548-553).

Mohasseb, A., Bader-El-Den, M., & Cocea, M. (2018, nov). Question categorization and classification using grammar based approach. *Information Processing and Management*, *54*(6), 1228–1243.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016).* Barcelona, Spain.

Padmanabhan, V. N., & Mogul, J. C. (1996). Using predictive prefetching to improve world wide web latency. *ACM SIGCOMM Computer Communication Review*, *26*(3), 22–36.

Pitkow, J., & Pirolli, P. (1999). Mining longest repeating subsequences to predict world

wide web surfing. In *Proceedings of the 2Nd Conference on USENIX Symposium on Internet Technologies and Systems - Volume 2* (pp. 13–13). Berkeley, CA, USA: USENIX Association.

Punjani, D., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., ... Stamoulis, G. (2018). Template-Based Question Answering over Linked Geospatial Data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval* (pp. 7:1—-7:10). New York, NY, USA: ACM.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

Raubal, M., & Winter, S. (2002). Enriching wayfinding instructions with local landmarks. In M. J. Egenhofer & D. M. Mark (Eds.), *Geographic Information Science* (pp. 243–259). Berlin, Heidelberg.

Richter, D., Winter, S., Richter, K.-F., & Stirling, L. (2013). Granularity of locations referred to by place descriptions [Journal Article]. *Computers, Environment and Urban Systems*, *41*, 88-99.

Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020). Geo-analytical question-answering with gis. *International Journal of Digital Earth*, *0*(0), 1-14.

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

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012, October). Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, *3*(4), 333–354.

Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., & Cohen, W. (2018, October-November). Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4231–4242). Brussels, Belgium: Association for Computational Linguistics.

Suomela, R., Lakkala, H., & Salminen, I. (2009, January 20). *Displaying a map having a close known location.* Google Patents. (US Patent 7,480,567)

Tax, N., Teinemaa, I., & van Zelst, S. J. (2020). An interdisciplinary comparison of sequence modeling methods for next-element prediction. *Software and Systems Modeling*.

Thalhammer, A., & Rettinger, A. (2016). Pagerank on wikipedia: towards general importance scores for entities. In *European Semantic Web Conference* (pp. 227–240).

Tomko, M., & Purves, R. S. (2008). Categorical prominence and the characteristic description of regions. In *Proceedings of the Semantic Web Meets Geospatial Applications Workshop, held in conjunction with AGILE 2008.* Girona, Spain.

Vahedi, B., Kuhn, W., & Ballatore, A. (2016). Question-based spatial computing—a case study. In T. Sarjakoski, M. Y. Santos, & L. T. Sarjakoski (Eds.), *Geospatial Data in a Changing World* (pp. 37–50). Cham: Springer International Publishing.

Wang, X., Zhang, Y., Chen, M., Lin, X., Yu, H., & Liu, Y. (2010, June). An evidence-based approach for toponym disambiguation. In *Proceedings of 18th International Conference on Geoinformatics* (p. 1-7).

Wilson, D., & Sperber, D. (2002). Relevance theory. In *Handbook of Pragmatics.* Blackwell.

Wilson, J. A. (2018, May 17). *Systems and methods for presenting personalized map labels.* Google Patents. (US Patent App. 15/811,376)

Winter, S. (2009). Spatial intelligence: ready for a challenge? *Spatial Cognition & Computation*, *9*(2), 138–151.

Zheng, W., Cheng, H., Yu, J. X., Zou, L., & Zhao, K. (2019, may). Interactive natural language question answering over knowledge graphs. *Information Sciences*, *481*, 141–159.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, *24*(5), 530–536.