Extracting interrogative intents and concepts from geo-analytic questions

Haiqi Xu[1], Ehsan Hamzei[2], Enkhbold Nyamsuren[1], Han Kruiger[1], Stephan Winter[2], Martin Tomko[2], and Simon Scheider[1]

[1] Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, Netherlands
{h.xu1,e.nyamsuren,j.f.kruiger,s.scheider}@uu.nl
[2] The University of Melbourne, 3010 Victoria, Australia
ehamzei@student.unimelb.edu.au,{winter,tomkom}@unimelb.edu.au

Abstract. Understanding syntactic and semantic structure of geographic questions is a necessary step towards true *geographic* question-answering (GeoQA) machines. The empirical basis for the understanding of the capabilities expected from GeoQA systems are *geographic question corpora*. Available corpora in English have been mostly drawn from generic Web search logs or limited user studies, supporting the focus of GeoQA systems on retrieving *factoids*: factual knowledge about particular places and everyday processes. Yet, the majority of questions enquired about in the spatial sciences go beyond simple place facts, with more complex analytical intents informing the questions. In this paper, we introduce a new corpus of *geo-analytic* questions drawn from English textbooks and scientific articles. We analyse and compare this corpus with two general-purpose GeoQA corpora in terms of grammatical complexity and semantic concepts, using a new parsing method that allows us to differentiate and quantify patterns of a question's intent.

Keywords: Geo-analytic questions · Geographic questions · Information extraction · Grammatical parser · Concepts and intents · Geographic question-answering systems.

1 Introduction

Questions about locations of places or events are frequent in Web search. Until recently, only basic geographic questions, such as *"Where is Fiji?"*, were satisfactorily answered by search engines [15,22]. Driven primarily by the need to enable smart assistants, such as Siri (Apple) and Cortana (Microsoft), to answer situated questions, spatial question-answering (QA) has received increased attention in recent years [5]. Researchers have conducted studies about the form of spatial questions used in search engines [11]. Furthermore, geographic question-answering systems (GeoQA) have become a topic of interest in GIScience [21].

Compared to current capabilities of search engines and QA systems, *geo-analytic questions*[3] asked by professionals in Geography and the spatial sciences are

[3] This term vaguely refers to the set of questions used in the context of GIS analysis.

much more sophisticated and require more nuanced answers to be found within Geographic Information Systems (GIS). For example, interpreting and answering the geo-analytic question "What is the best site for a new landfill in the Netherlands?" requires identifying a range of siting criteria with respect to geological, hydrological, and environmental factors, such as "being far from nature conservation areas".

In principle, geo-analytic questions may be answered by QA systems that rely on retrieval of factual knowledge from documents or knowledge bases [6,20,21,25]. However, it is quite unlikely that answers to geo-analytic questions are known a-priori and, therefore, such answers will probably not be accessible through information retrieval. Instead, geo-analytic question answering usually requires generating analytic workflows. Parsing geo-analytic questions and generating corresponding workflows is, however, beyond the capabilities of current QA systems [29]. For a QA system to support this, we need an in-depth understanding of goal concepts and corresponding intentions expressed in the questions, which is a non-trivial task.

Question corpora provide means for analyzing question structures from a syntactic as well as semantic viewpoint [31]. Current efforts focus on the collection and analysis of Web queries [15,22] and directly answerable questions [6,25] for QA systems. However, corpora specific to geo-analytic questions are not available, and it is therefore impossible to assess how the complexity of geo-analytic questions differs from generic questions with a geographic component, as asked e.g. on search engines.

We introduce *GeoAnQu*, a novel geo-analytic question corpus in English that can be used for non-factoid GeoQA (Section 3). We compare this corpus with two existing geographic question corpora, MS MARCO and GeoQuestion201. To our knowledge, this paper is the first to provide such a comparison. In Section 4, we propose a new approach to investigate the syntactic structures and the goal concepts across the questions in all three English corpora. In Section 5, the three corpora are compared with each other at word level, phrase level, and sentence level. In conclusion, we discovered significant differences between these corpora. In a nutshell, geo-analytic questions in GeoAnQu are syntactically and semantically more complex than questions in the two other corpora, in the sense that they contain more phrases and clauses and require understanding of certain concepts on an expert level. It should be noted that this study is focused on the English language. Therefore, the conclusions are not generalizable to other languages.

2 Related work

In this section, we discuss related work, starting with definitions of questions, over corresponding corpora, to studies that analyse geographic questions.

2.1 Geo-analytic questions

Why is geo-analytic QA challenging? And why does the problem require more than what current syntactical question analysis in QA has to offer? Geo-analytic questions are posed (and often operationalized using GIS) primarily to find out about spatial patterns and relations that are not known, or may be less obvious. GIS textbooks (e.g., [23]) provide taxonomies of questions addressed with GIS and rank them from simple to complex, such as: *location*, *condition*, *routing*, *pattern modeling*, *trend modeling*, and *what-if modeling* questions. Location questions are the simplest questions. Condition questions are more complex, e.g, *"What houses are for sale and within 1km from the nearest school in Utrecht?"*. Pattern modeling questions enquire about spatial variation, e.g., *"What is the concentration pattern of ethnic groups in Amsterdam?"*. Furthermore, these classifications are not categorical, i.e., a condition question can also be a location question, or a pattern modeling question can also be a condition question. Answering a geo-analytic question in a GIS workflow may therefore require the decomposition of the question into multiple simple parts that may be answered separately [28].

2.2 Geographic question corpora

Question corpora serve two main purposes: (1) understanding what kind of questions are asked and what the typical answers are; and (2) serving as gold standards for QA system evaluation.

Sanderson and Kohler [27] were among the first to highlight that geographically related queries form a significant subset of search engine queries. They defined geographic queries as any query that mentions at least one of: geographic term (e.g., place name), locator (e.g., postcode), geographic feature (e.g., lake, island). Their analysis revealed that geographic queries constituted 18.6% of 2500 queries randomly sampled from the logs of the internet portal Excite[4] [27]. Similarly, geo-queries with place names constituted 12.7% of a random sample compiled from Yahoo query logs [15]. A similar picture occurs within Microsoft Live search logs, indicating that search engine users have a stable need for geographic information [2].

The Geoquery corpus is an early geographic domain-specific question corpus [32], which contains basic questions about the USA geography (e.g., *"What are the major cities in Kansas?"*). In a more recent effort, Chen et al. [6] collected about 500 questions in five categories based on a survey of 50 students and a provided map instrument (see Table 1). Interestingly, many of these questions currently require analytical operations due to the lack of available factual answers retrievable from text or databases.

The more recent GeoQuestion201 corpus [25] was created as a benchmark for Geographic QA systems and consists of 201 questions. GeoQuestion201 is similar to the Geoquery corpus [6] with respect to questions about relative location and proximity. We discuss more about this corpus in the later sections.

[4] http://www.excite.com

Table 1. Geographic question corpus [6].

Question type	Example question	% corpus
Location	Where is Columbus?	32%
Relative location	Where is Columbus with respect to Cincinnati?	13%
Distance	How far is Columbus from Cleveland?	28%
Proximity entity	Which city is the nearest to Columbus?	12%
Proximity buffer	What cities are within 5 miles from Columbus?	15%

2.3 Analyzing geographic questions

In Geographic Information Retrieval, geographic Web queries were investigated through information parsing, extraction and classification [2,3,11]. Information types such as *place names*, *spatial relationships* and *place types* are widely used [26,27] in queries. Additionally, *activities* and *situations* were proposed as categories to more deeply investigate geographic questions [11]. A strong relationship between information categories and linguistic/grammatical clues such as parts of speech is the main reason for using grammatical parsing in the information extraction task [11].

Determining the type of a question allows for a shallow understanding of the question and the type of answer [8]. Supervised [10], unsupervised [11] and semi-supervised [3] classification methods have been explored. On the broadest level, geographic web queries can be categorized into *spatial* and *non-spatial* [11], as well as into *local* and *global* geographic queries [10].

On a deeper semantic level, the *intent* of a geographic question defines the inquirers' goals or information need [9]. Henrich and Luedecke [13] proposed a taxonomy of information needs for geographic questions, consisting of four classes: (1) *to there*: wayfinding purpose; (2) *about there*: information about a place; (3) *there*: activity inside a place; and (4) *from there*: to get something from a specific place. However, the complex intents of geo-analytic questions are not well captured by the proposed taxonomies. Furthermore, an approach to automatically parse questions and extract their intent is missing. In this paper, we propose a novel automated approach for identification and disambiguation of intents in geo-analytic questions.

3 Question corpora used

In this paper, we investigate three datasets, (1) MS MARCO, (2) GeoQuestion201, and (3) GeoAnQu, a novel corpus of geo-analytic questions. These datasets contain geographic questions asked by different types of inquirers.

MS MARCO v2.1 [22] is the largest available machine-comprehension corpus ($>$ 1 million records). It contains questions sampled from logs of a general-purpose search engine, Bing. The questions are classified as (1) numeric, (2) description,

(3) entity, (4) person and (5) location questions. Here, we focus on a subset of *geographic questions*. Geographic questions are "location questions" related to geographic places identified using the GeoNames gazetteers[5]. This corpus, detailed in [11], contains 36,939 geographic questions out of 56,721 location questions.

GeoQuestion201 [25] contains 201 geospatial questions generated by students enrolled in an Artificial Intelligence course, thus not requiring prior knowledge of GIS and GIS workflows. The students were asked to consider talking to an intelligent assistant to address their geographical information needs [25]. The questions are limited to a few scenarios constructed with geographic concepts including *geographic feature*, *geographic type*, *attributes* and *spatial relationships*.

GeoAnQu is a new dataset introduced in this paper, containing 429 geo-analytic questions compiled from different sources: (1) 100 scientific articles collected in the context of a Master thesis at Utrecht University using Scopus [30]. The articles were filtered via three criteria: in the field of Human Geography, containing GIS analysis, and published in 2009-2018[6]. These articles tackled diverse questions in Human geography, including but not limited to Healthcare and Environmental Planning. In some cases, articles explicitly stated the question, but in most cases, we had to formulate the question based on reading the article; (2) textbooks on GIScience and GIS [1,14,17,23]. These textbooks were searched for questions within GIS tutorial and exercise scenarios. All questions found in these books were included. In some cases (e.g. [1]) we had to reformulate questions when they were not yet made explicit.

4 Method

In this section, we introduce the methods we used for parsing, analyzing and comparing the different corpora. Note that the questions in three corpora are not always idiomatic depending on the language proficiencies of the sources. This is especially true for web queries in MS MARCO. We have preserved the original idiosyncrasies and errors in the MS MARCO and GeoQuestion201 corpora to the extent that it does not influence analysis results.

4.1 Encoding and Parsing

Encoding and parsing are used for extracting semantic information from questions by differentiating (1) the intent of questions and (2) the descriptions and criteria of the intent. The intent defines what can be considered an answer. Descriptions and criteria determine restrictions on a valid answer. For example, *"How many*

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

[6] The collection of articles and questions is made available under the main download link (Sect. 4.4).

refugees live in Germany between 2000 and 2010?" asks for "how many refugees" under the condition that they "live", "in Germany", "between 2000 and 2010". The latter are situation, location and time criteria, respectively.

Fig. 1 shows a flowchart for extracting semantic information from questions. We have used part-of-speech (POS) tagging, named entity recognition, and constituency parsing. As shown in Table 2, the part-of-speech gives already some clue for tagging each part of the question in terms of its semantic type. To further differentiate the noun-phrases into semantic types, we use a pre-trained named entity recognition model [18] to identify *toponyms* and a predefined dictionary to tag *place types*. The dictionary is a list of place types manually selected from all the questions. Constituency parsing [16] is used to capture the relation between extracted semantic types (e.g., *spatial relationships* and *toponyms*). To identify verbs as *activities* or *situations*, vector representations of the verb are derived from the *ELMo* word embedding model [24], and later judged using cosine similarity to handcrafted sets of action and stative verbs. The difference between *entity qualities* and *place qualities* is their link to entities, toponyms or place types. After labeling toponyms, types and entities, the adjectives are labeled based on their associated links in the parse tree. Similarly, to label prepositions that convey *spatial relationships*, the parse tree is used to check direct links between the preposition and a toponym or a place type.

Table 2. Extended semantic encoding.

Semantic type	Part-of-speech	Code	Semantic type	Part-of-speech	Code
where	WH-word	`1`	*toponym*	noun	`n`
what	WH-word	`2`	*place type*	noun	`t`
which	WH-word	`3`	*date*	noun	`d`
when	WH-word	`4`	*entity*	noun	`e`
how	WH-word	`5`	*place quality*	adjective	`q`
how+adj	WH-word	`6`	*entity quality*	adjective	`p`
why	WH-word	`7`	*situation and event*	verb	`s`
yes/no questions	verb	`8`	*activity*	verb	`a`
			spatial relationship	preposition	`r`

To divide a geographic question into *intent* and *criteria*, we use the results of parsing and the types of question words. In some cases, the type of question word unambiguously defines the intent of the question – i.e., *yes/no questions*. However, in other cases, such as *what-questions*, the intent remains unclear from the question word and consequently the answer remains ambiguous.

To resolve the issue, first we used grammatical phrases such as WHNP (WH-noun phrase – e.g., "how many workers" or "what measure") which are extracted by *depth-first search* from the parse tree. If the parse tree fails to disambiguate the intent, the following heuristic is used: the first largest entity/type phrase after the question word is the missing part of the intent. An entity/type

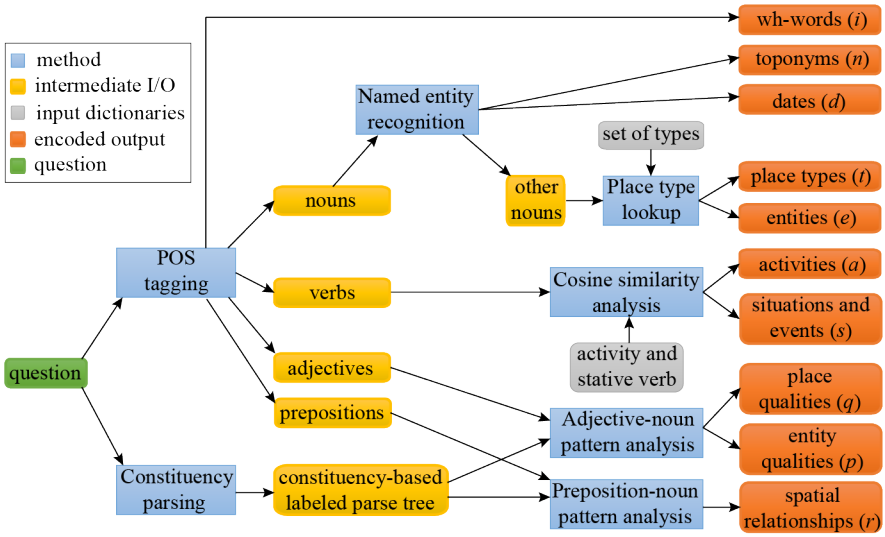


Fig. 1. Encoding and parsing algorithm.

phrase is a noun phrase containing types or entities and their qualities. This assumption is grounded in the *sequential structure of natural language* [4]. Also, entity/type phrases are more ambiguous compared to toponym/date phrases. Hence, these are more likely to be related to the intent of questions. Algorithm 1 shows the proposed approach to detect the intent of questions. Fig. 2 shows an example of the intent extraction.

Algorithm 1 extracting intent of the questions

Require: The parse tree C of question Q
1: **if** the question word (i) in [where, when, yes/no questions, why] **then**
2: **Return** i
3: **end if**
4: Extract the largest WH-phrase (wp) in C
5: **if** wp not equal to i **then**
6: **Return** wp
7: **end if**
8: Find the largest entity/type phrase (ph) after i
9: **Return** concatenation of i and ph

4.2 Analyzing each corpus

To understand the content and basic structures of the three spatial question corpora, we used techniques from statistics and natural language processing

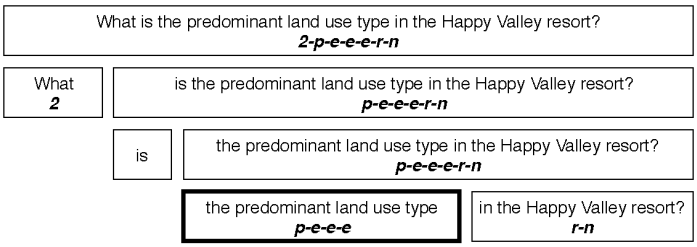


Fig. 2. An example of extracting the intent of a what question using our heuristic.

(NLP). Our analysis aims at discovering general syntactic and semantic patterns of geographic questions, as a basis for generating grammars for translating questions into machine-readable queries.

We generated word clouds [12,19] for semantic types of nouns, adjectives, verbs, and prepositions in all question corpora, to give an overview about the diversity of terms used. In order to quantify frequency, we also created frequency tables for each semantic type that is not a WH-word.

Finally, the syntax of questions was analyzed using *encoding n-grams*. The extended encoding (Table 2) is used as a vocabulary for generating the n-grams, and the frequency of each n-gram is used to define phrase-level concepts – e.g., *rn* is a bigram that defines a geographic extent. To provide detailed information about the syntactic structure of the questions, n-grams are generated on three levels: question, intent, and criteria.

4.3 Comparing three corpora

The three corpora were compared on the word level, phrase level and sentence level. For word-level comparison, frequent words of each encoding type (e.g., quality) are compared across the corpora. The difference in distributions of n-grams reflects the phrase-level differences. For sentence-level comparison, vector representations of question encodings are used to compare the content over three datasets. The encoding vectors are inside a 17-dimensional space (each encoding type in Table 2 is an axis in this space) and constructed by counting the number of types occurring in the content of the question.

4.4 Data and software availability

This article makes use of two third-party data sources. GeoQuestions201 is freely available without a license statement[7]. The MS MARCO corpus is freely available under a proprietary agreement for non-commercial use[8]. All other research data supporting this publication are available under the Creative Commons Attribution

[7] http://geoqa.di.uoa.gr/benchmarkquestions.html
[8] https://microsoft.github.io/msmarco/

4.0 International Public License[9] and can be accessed here[10]. The computational workflow in this publication is executed via multiple script files written in Python and R. All scripts are available under the MIT License[11] and accessible above.

5 Results

We evaluate our parsing model using inter-annotator agreement on MS MARCO, GeoQuestion201, and GeoAnQu in Section 5.1. We then analyse the corpora at the word and phrase level using word clouds, frequency tables, and n-gram tables (Sect. 5.2, 5.3). In Section 5.4, we compare the structural complexity of questions in the three corpora at the sentence level. Finally, we investigate what questions more deeply according to question intents and goals (Sect. 5.5).

5.1 Evaluating parser results

To evaluate the parser performance, we first measured inter-annotator agreement of manual annotations. We randomly selected 100 questions from each corpus. These questions were distributed among four annotators such that each question was annotated by at least three annotators. Annotations with two or more inter-annotator agreements were used for evaluating the parser. Fig. 3 and Fig. 4 show the inter-annotator agreements [7], as well as precision and recall of the parser for encoded classes and corpora, respectively.

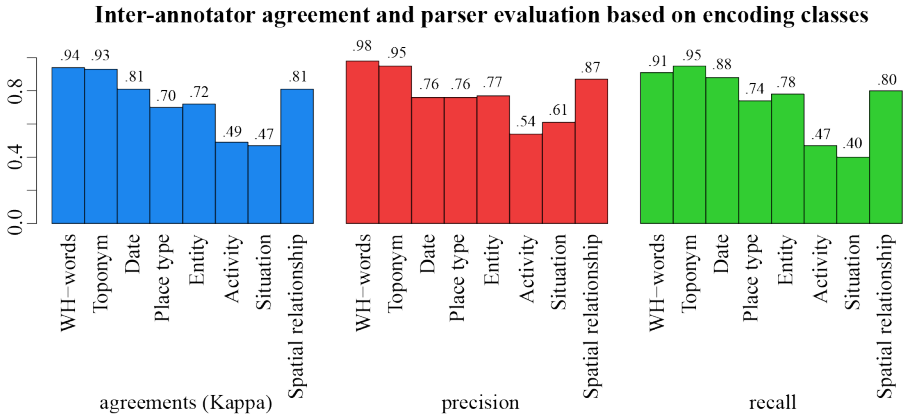


Fig. 3. Inter-annotator agreement. Precision and recall of parser results for each encoding class.

[9] https://creativecommons.org/licenses/by/4.0/
[10] https://figshare.com/s/b3f8b0834ca63b6c5d60
[11] https://opensource.org/licenses/MIT

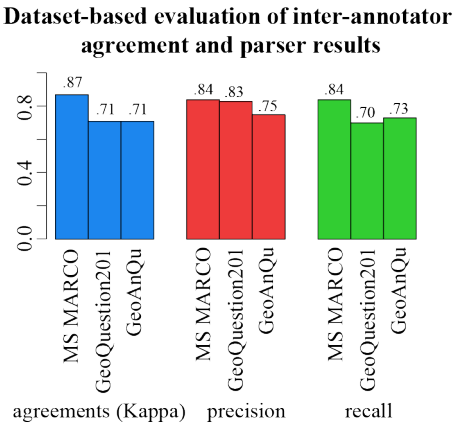


Fig. 4. Inter-annotator agreement. Precision and recall of parser results for each dataset.

While most encodings reach a moderately high agreement score (over 0.7), a considerable disagreement between annotators was observed for activities and situations in Fig. 3, and the parser is also less precise in these cases. Two main reasons are: (1) the sparsity of labelled data can make results extra sensitive to mismatches, and (2) the difficulty of the task caused by vagueness of encoding classes – e.g., whether 'unit' or 'spot' should be encoded as *place type* or *entity*.

According to Fig. 3 and Fig. 4, in some cases the precision and recall of the parser is better than the agreement. However, this does not mean the parser surpasses human performance, as the parser evaluation is based on the agreeing annotations which is only a subset.

5.2 Qualitative and quantitative representation of words

The word cloud graphs of the three question corpora indicate that the GeoAnQu corpus uses very different keywords than GeoQuestion201 and MS MARCO.

Entity words In the GeoAnQu corpus, the word cloud of entities (Fig. 5(a)) primarily consist of *statistical/analytical entities*, such as distribution, pattern, change, and accessibility, *events* such as crime, fire, and hurricane, and *thematic entities* such as segregation (politics), mortality (social science) and tourism (tourism). 78 out of 540 entities occur more than three times and account for 51% of all occurrences. Fig. 5(b) shows that 45% of these occurrences are analytical. Hence, a large fraction of geo-analytic questions ask about statistical and analytical features of events or themes. For example, *"What is the spatial distribution of probabilities of robberies in Salvador, Brazil?"*.

In contrast, GeoQuestion201 (Fig. 6(a)) and MS MARCO (Fig. 6(b)) contain many *spatial relationships*, such as border, north, west and east, which serve to localize places. Furthermore, GeoQuestion201 contains *geometric measurements*

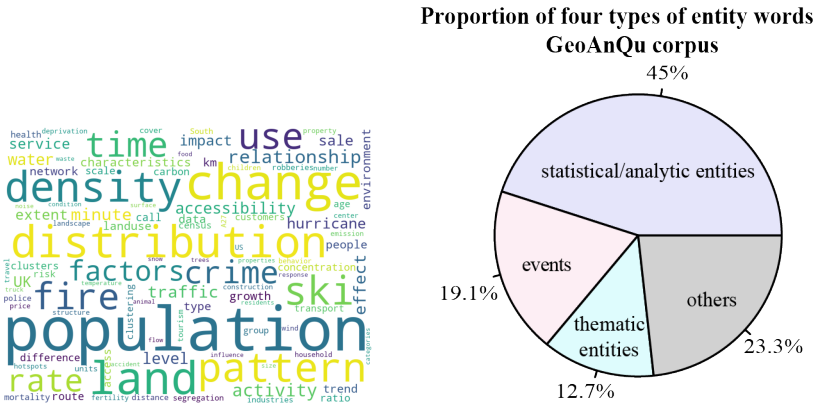


Fig. 5. (a) Word clouds of entity words. (b) Proportions of four types of 78 entity words.

such as length, height, and radius. We conclude that the latter two corpora probably focus more on spatial relationships and attributes of places. For example, *"Is Hampshire north of Berkshire?"* and *"Which rivers in Scotland have more than 100 km length?"*. MS MARCO contains also more words from daily life, such as "address", "capital", and "terminal".

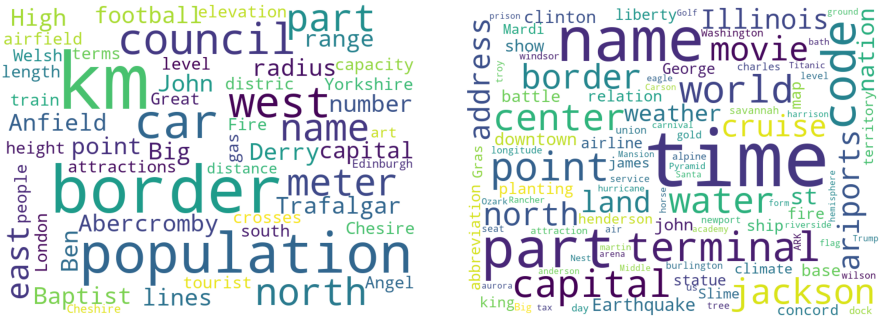


Fig. 6. Word-level analysis of entity words in (a) GeoQuestion201 and (b) MS MARCO

Activity and situation verbs As shown in Fig. 7, the three corpora are also distinct in terms of verbs. In GeoAnQu, 55% of the situation verbs (Fig. 7(a)) are related to the intended analysis, such as "predict", "clustered", "distributed", and "compare" (e.g., "*Where are fire calls highly clustered in Fort Worth?*"). In

contrast, GeoQuestion201 and MS MARCO contain situation verbs (Fig. 7(c) and 7(e)) which reflect simple location questions such as "*Where is Elizabeth Tower located?*". The activity verbs in GeoAnQu (Fig. 7(b)) are related to trend analysis such as "change", "expand" and "reduce", while in MS MARCO, activity verbs (Fig. 7(f)) rather reflect daily processes such as "buy", "eat", "fly".

5.3 Phrase analysis

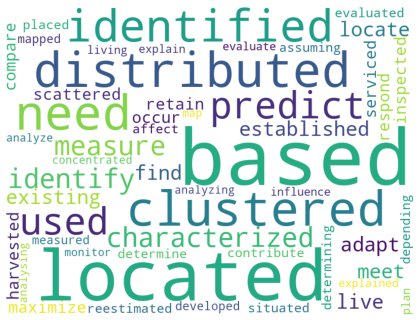
We used n-gram analysis to identify phrase-level patterns. To quantify the influence of each pattern, we computed the percentage of questions that contain a specific pattern. We first investigated *question words* (Fig. 8). Then, we selected patterns according to their semantic role in the question (Fig. 9). The first nine patterns (*question intent patterns*) are used to specify the types of intents of questions, and the last three specify the *spatial extent and further details* of the intent. Table 3 gives examples for each pattern.

Table 3. Example questions for n-grams patterns.

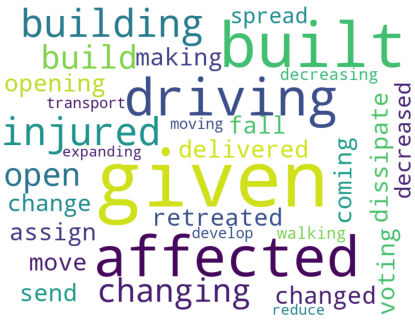
Pattern	Example question
1ee	Where(1) are king(e) Cobras(e) from?
1n	Where(1) is Zlin(n)?
2e	What(2) is the bikeability(e) in the Metro Vancouver region of Canada?
2ee	What(2) is the population(e) density(e) in the Banten province?
2qt	What(2) is the longest(q) bridge(t) in Scotland?
2t	What(2) houses(t) are for sale and within 500m from main roads in Utrecht?
3qt	Which(3) is the oldest(q) bridge(t) of London?
3t	Which(3) counties(t) of Ireland does River Shannon cross?
6qt	How(6) many(q) rivers(t) cross Edinburgh?
rn	Where is Hanover School in(r) Colorado(n)?
rnn	What is the population distribution in(r) Tarrant County(n), Texas(n)?
rtrn	What is the density of trees in(r) parks(t) in(r) Oleander(n)? What is the walkability of(r) each neighborhood(t) in(r) Ghent(n)?

Question words As shown in Fig. 8, *What* and *Where* questions constitute 54% and 19.6% of GeoAnQu respectively. In MS MARCO, *What* (46.7%) and *Where* (49.6%) questions are also prevalent. *Which* (51.5%) and *yes/no* (27.4%) questions are most frequent in GeoQuestion201.

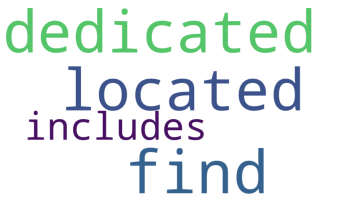
Question intent patterns Question intents consist of *entities*, *place types* and *toponyms*. For example, *"What(*2*) is the bikeability(*o*) in the Metro Vancouver region of Canada?"*, where bikeability is the intent. In GeoAnQu, questions with *entities* as intents constitute 33.7% of the corpus, where *2e* has the largest



(a)

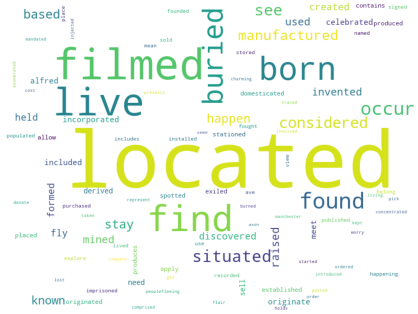


(b)



(c)

(d)



(e)

(f)

Fig. 7. Word-level analysis of activity and situation verbs in (a)(b) GeoAnQu, (c)(d) GeoQuestion201, (e)(f) MS Marco.

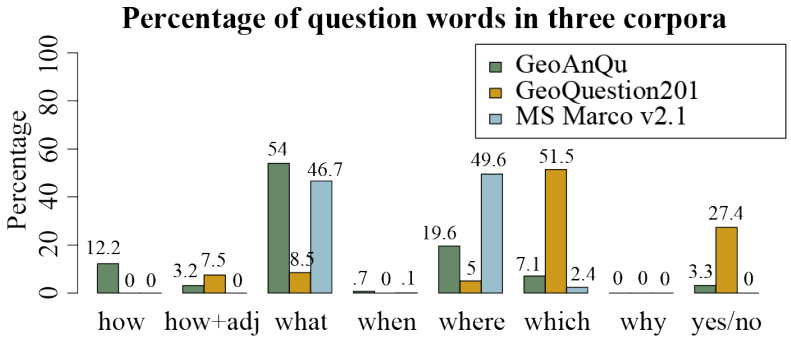


Fig. 8. The prevalence of semantic types within three corpora.

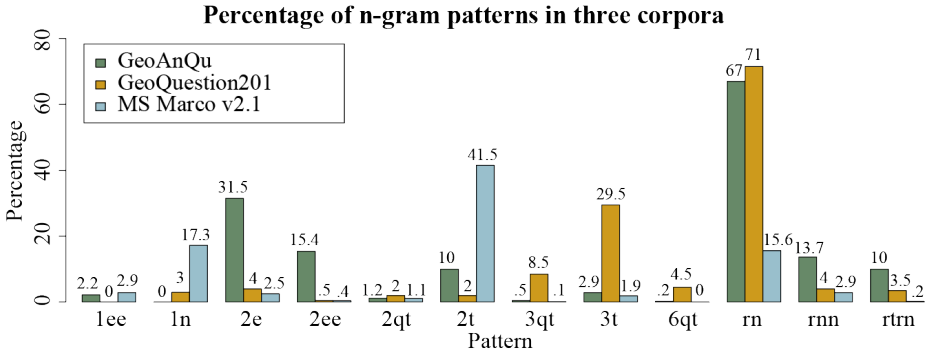


Fig. 9. Phrase analysis of n-grams patterns in three corpora

proportion, followed by *2ee*. *Place types* as intent constitute 14.8% of GeoAnQu. In contrast, *place types* are the dominant kind of intent in GeoQuestion201 (46.5%) and MS MARCO (44.6%), as opposed to 14.8% in the case of GeoAnQu. The bigram *2t* is an important pattern for MS MARCO, and *3t* for GeoQuestion201, both asking for place types. Only MS MARCO has a lot of questions with *toponym* intents.

Spatial extent patterns The bigram *rn* and the trigram *rnn* both represent spatial extents in geographic questions, especially for the corpora GeoAnQu and GeoQuestion201. The second *n* of the trigram pattern always represents a place name that contains the first *n*. The 4-gram pattern *rtrn* specifies even finer spatial extents, such as "*in parks in Oleander*", and spatial units within the extent, such as "*of each neighborhood in Ghent*".

5.4 Multidimensional comparison of corpora

Fig. 10 represents the structural complexity of the three corpora. To balance the distribution of GeoQuestion201 and GeoAnQu, we randomly sampled 200 questions from each corpus and compared them in Fig. 10.

Most of the questions in MS MARCO and some of the GeoQuestion201 questions are clustered while GeoAnQu questions are more distributed. The clusters show that the questions are very similar in terms of structural encoding. Hence, geo-analytical questions in GeoAnQu are more diverse in structure.

5.5 Investigating *What* questions in GeoAnQu corpus

The prevalence of *What* questions in GeoAnQu warrants a deeper investigation of its syntactic and semantic structures. Syntactically, most *What* questions follow one of two forms defined by the position of the intent phrase.

In *Form 1*, the intent phrase follows the auxiliary: "*What is/are the* [intent phrase] ...". An example is "*What* $^{(Aux)}$*is the* $^{(Int)}$*predominant land use type in the Happy Valley resort?*". In *Form 2*, the intent phrase precedes the auxiliary: "*What* [intent phrase] *is/are* ...". Example: "*What* $^{(Int)}$*areas* $^{(Aux)}$*do have too few roads to handle the traffic in Oleander?*". We used regular expressions to parse the two forms based on the encoding according to Table 2.

Due to the nature of the inquiry, place type intents do not reveal much about analytic contents, whereas entity intents are more insightful in this respect (Fig. 11). Most frequent entity intents are *pattern*, *relationship*, *distribution*, *density*, *effect*. They reveal geo-analytic components of questions and expected answers. For example, in the question "*What is the spatial probability distribution of robberies in buses in Salvador, Brazil?*", the intent phrase "*spatial probability distribution*" requests distribution data. Furthermore, entity intents also indicate what analytic methods can generate the answer. For example, *accessibility* measures imply distance estimation methods. *Spatial patterns* and *distributions* can be measured with auto-correlations and clustering methods.

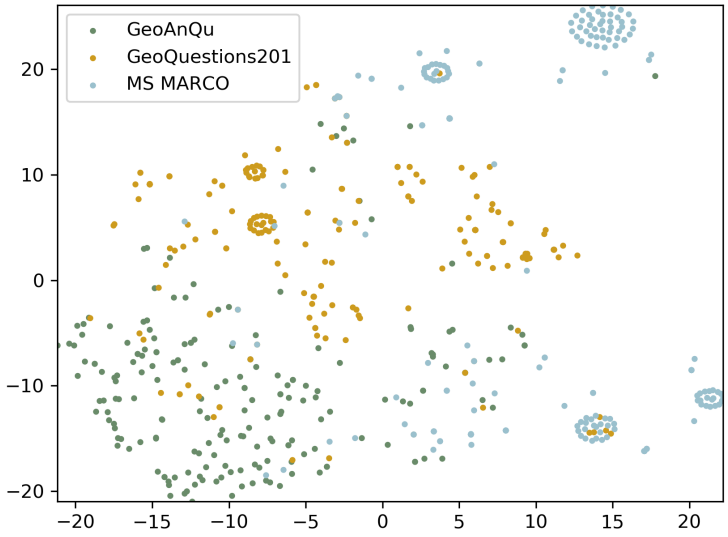


Fig. 10. Comparing question content of the three corpora. We used *t-Distributed Stochastic Neighbor Embedding* (t-SNE) to reduce the dimension of the word vectors to 2D for visual inspection.

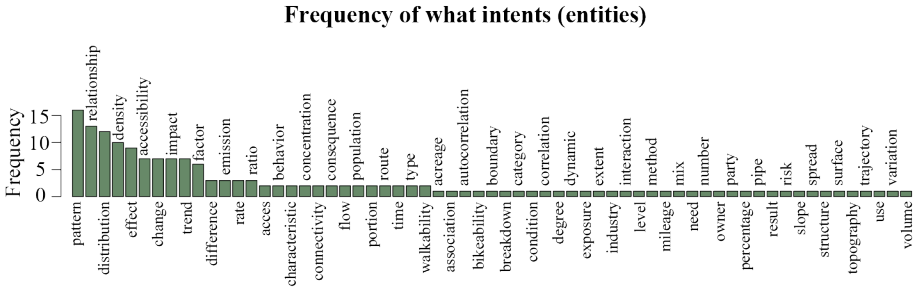


Fig. 11. Words used as entity intent in What questions.

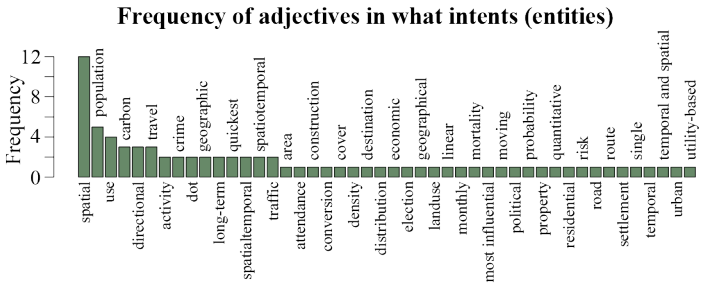


Fig. 12. Adjectives modifying the intent.

In many cases, the intent word is modified by one or more adjectives. For example, in the question "*What are the noise mitigation zones around each runway in Schiphol airport?*", the intent word "*zones*" is modified by two adjectives "*noise*" and "*mitigation*". These adjectives provide thematic information by clarifying the kinds of zones relevant to this particular case. Intent adjectives often carry thematic information. Fig. 12 shows a distribution of adjectives modifying entity intents in *What* questions of the GeoAnQu corpus. Most words, such as *population*, *(land) use*, and *carbon* are thematic, hinting at datasets that may be relevant to answering the question. For example, given a question "*What is the population density ...*", the adjective "*population*" may map to a population census dataset and the intent "*density*" may map to some density estimation method. However, adjectives may denote also other semantic information, as revealed by the intent-adjective co-occurrence matrix in Fig. 13. Adjectives and intents co-occurring frequently, such as *dot density*, *directional trend*, and *spatial distribution*, clarify the analytical measure necessary. In other patterns, such as *quickest route* and *best site*, adjectives rather hint at criteria according to which objects should be selected as an answer.

In GeoAnQu, *What* questions with *Form 1* and entity intents are the most analytic questions. So far, we have justified our focus on *What* questions with prevalence of these questions in the GeoAnQu corpus. However, it is likely that *What* questions are also more universal than other types of questions. In other words, it may be possible to formulate the most other types of questions in terms of *What* questions. For example, questions of type "*How many ...*" may be reformulated as "*What is the number ...*", and *Where* questions may be restated as "*What is the location*". In this way, *What* questions may offer more utility in expressing complex analytical inquiries than other question types. This utility may explain why *What* questions of Form 1 are so much more prevalent in geo-analytic context.

So far, we only focused on intent phrases of geo-analytic *What* questions. Yet, these questions usually have a richer semantic structure. Fig. 14 suggests a more comprehensive schema of *What* questions on the syntactic and semantic level. Syntactically, a what-question consists of an intent phrase that is followed by zero or more entities that are not part of the intent phrase. The question usually ends with a statement about a relevant place and, optionally, a date. On the semantic level, these constituents reveal what type of answer is expected and what method and datasets may be used. As discussed earlier, parts of the intent phrase may encode analytic measures, selection criteria or thematic content. The entities following the intent phrase further specify the thematic content, such as *urban growth*. These entities can be quite complex, and, therefore, necessitate further analysis beyond this work. Finally, places and dates specify spatial and temporal extents to which the analysis should be limited. The schema allows geo-analytic *What* questions to express a wide variety of geo-analytic problems, and it enables decomposition into meaningful parts for machine processing and automated question answering.

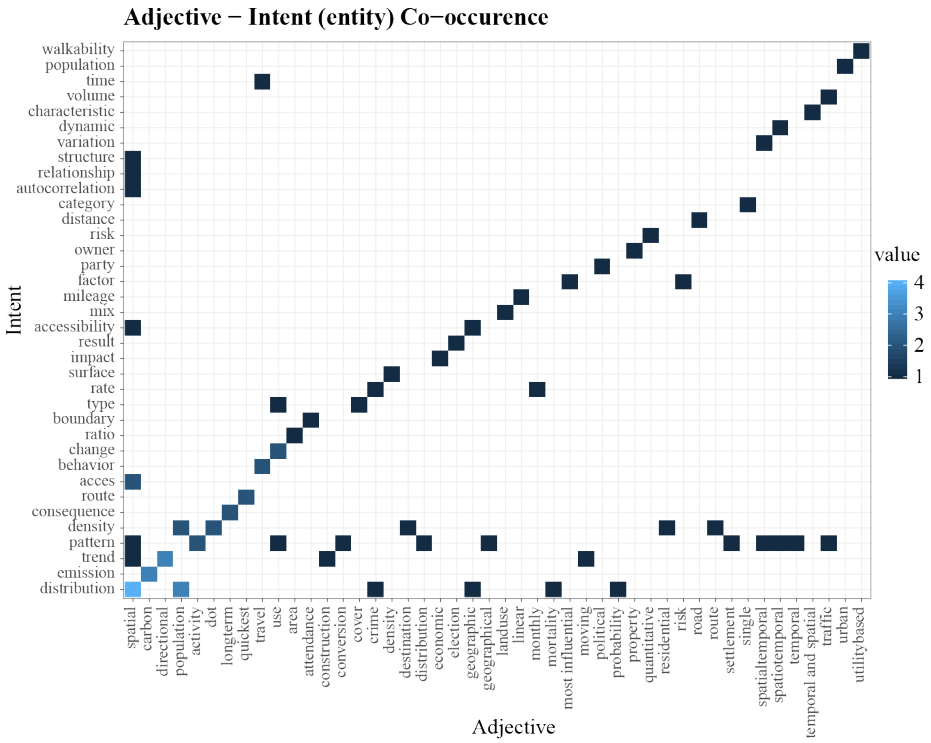


Fig. 13. Co-occurrences of intent words and adjectives.

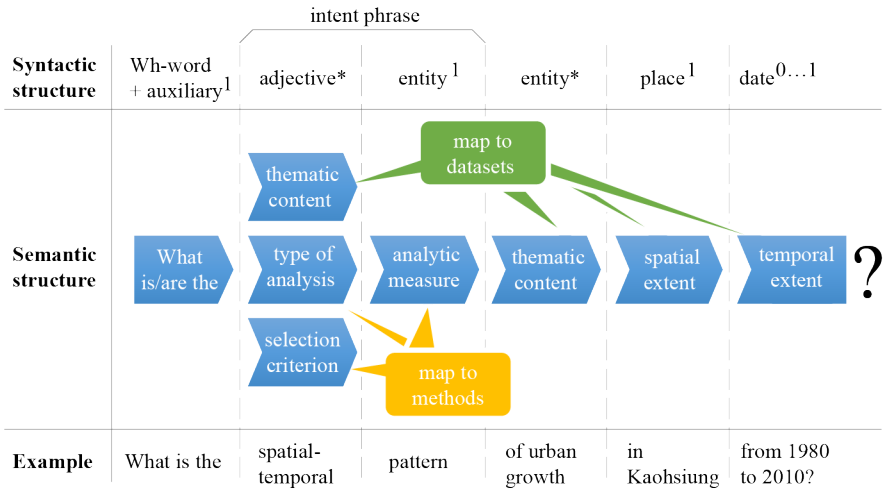


Fig. 14. Structural schema of geo-analytic *What* questions.

6 Discussion and conclusion

This paper introduces a new corpus of geo-analytic questions, which can be used to investigate the questions usually asked in the GIS domain, and for training parsers of geographic QA systems. We have proposed a novel approach to parse and encode the components of geographic questions. The approach was tested on three question corpora: MS MARCO, GeoQuestion201, and GeoAnQu. Inter-annotator agreement and parsing accuracy are high over three corpora, except for activities and situations. However, the latter were not in focus here. We have analyzed each corpus at the word and phrase levels, based on generating word clouds and by quantifying syntactic patterns. Additionally, sentence-level comparison reveal substantial semantic and syntactic differences between these corpora. GeoAnQu contains an abundance of what questions primarily asking for analytical entities, events and thematic entities, which are further specified using adjectives and dependent entities, as well as spatial and temporal extents. In contrast, MS MARCO questions ask more about particular places and place types, their qualities and relationships instead, and reflect everyday processes as opposed to information processes.

Our parsing approach can be improved in the future. For instance, in GeoQuestion201, nouns such as England and Wales were not always recognized as place names, but regarded as entities. This is because the Named Entity Recognition model was trained on proper English documents, while questions of GeoQuestion201 included simple grammar errors which lead to misclassifications. Moreover, the parser only considers the prepositions before a place name and annotates them as spatial relationships. For the phrase such as "trajectory of a hurricane", "neighborhood with low crime rate", the prepositions here connect the conditions ("*a hurricane*", "*low crime rate*") with the entities ("*trajectory*", "*neighborhood*"). It would be useful to distinguish these prepositions. To reduce the limitation of the constituency parse tree, we might use the dependency parser instead. The latter not only tokenizes text but also defines the relationships between tokens, and so might yield a better parsing performance.

In the future, we plan to use the phrase-level patterns found from the syntactic analysis to extract the *geo-information concepts* within geo-analytic questions, including analytic entities such as *patterns*, *densities* and *trend*, as well as related thematic entities such as *spatial objects* and *fields*. These could form a syntactic basis for a geo-analytic grammar that would allow automatizing the formulation of questions and their translation into analytic workflows.

Funding

This work was supported by the European Research Council (Grant No.803498) and the Australian Research Council (Grant No.DP170100109).

References

1. Allen, D.: GIS Tutorial 2: Spatial Analysis Workbook. Esri Press, Redlands (2013)

2. Aloteibi, S., Sanderson, M.: Analyzing geographic query reformulation: An exploratory study. Journal of the Association for Information Science and Technology **65**(1), 13–24 (2014)
3. Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving automatic query classification via semi-supervised learning. In: Fifth IEEE International Conference on Data Mining (ICDM'05). pp. 42–49. IEEE Press, New York (2005)
4. Bogaerts, L., Szmalec, A., Hachmann, W.M., Page, M.P., Duyck, W.: Linking memory and language: Evidence for a serial-order learning impairment in dyslexia. Research in Developmental Disabilities **43-44**, 106–122 (2015)
5. Canbek, N.G., Mutlu, M.E.: On the track of artificial intelligence: Learning with intelligent personal assistants. Journal of Human Sciences **13**(1), 592–601 (2016)
6. Chen, W., Fosler-Lussier, E., Xiao, N., Raje, S., Ramnath, R., Sui, D.: A synergistic framework for geographic question answering. In: 2013 IEEE Seventh International Conference on Semantic Computing. pp. 94–99. IEEE Press, New York (2013)
7. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
8. Ferrés Domènech, D.: Knowledge-based and data-driven approaches for geographical information access. Ph.D. thesis, Universitat Politècnica de Catalunya (2017)
9. González-Caro, C., Baeza-Yates, R.: A multi-faceted approach to query intent classification. In: String Processing and Information Retrieval. pp. 368–379. Springer, Heidelberg (2011)
10. Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the 12th International Conference on Information and Knowledge Management. pp. 325–333. ACM, New York (2003)
11. Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S., Tomko, M.: Place questions and human-generated answers: A data analysis approach. In: Proceedings of the 22nd AGILE Conference on Geographic Information Science. pp. 3–19. Springer International Publishing, Cham (2019)
12. Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: Text analytics based on word clouds. In: 47th Hawaii International Conference on System Sciences. pp. 1833–1842. IEEE Computer Society, Washington, D.C. (2014)
13. Henrich, A., Luedecke, V.: Characteristics of geographic information needs. In: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. pp. 1–6. ACM, New York (2007)
14. Heywood, I., Cornelius, S., Carver, S.: An Introduction to Geographical Information Systems. Pearson Education Limited, Harlow (2011)
15. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic intention and modification in web search. International Journal of Geographical Information Science **22**(3), 229–246 (2008)
16. Joshi, V., Peters, M., Hopkins, M.: Extending a parser to distant domains using a few dozen partially annotated examples. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 1190–1199. Association for Computational Linguistics, Melbourne (2018)
17. Kraak, M., Ormeling, F.: Cartography: Visualization of Spatial Data. CRC Press, Boca Raton (2013)
18. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego (2016)
19. Lohmann, S., Heimerl, F., Bopp, F., Burch, M., Ertl, T.: Concentri cloud: Word cloud visualization for multiple text documents. In: 19th International Conference on Information Visualisation. pp. 114–120. IEEE, New York (2015)
20. Mai, G., Janowicz, K., Yan, B., Scheider, S.: Deeply integrating linked data with geographic information systems. Transactions in GIS **23**(3), 579–600 (2019)
21. Mai, G., Yan, B., Janowicz, K., Zhu, R.: Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In: The Annual International Conference on Geographic Information Science. pp. 21–39. Springer International Publishing, Cham (2019)
22. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: 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 (2016)
23. O'Looney, J.: Beyond Maps: GIS and Decision Making in Local Government. ESRI Press, Redlands (2000)
24. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2227–2237 (2018)
25. Punjani, D., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., Beris, T., Bilidas, D., Ioannidis, T., Karalis, N., et al.: Template-based question answering over linked geospatial data. In: Proceedings of the 12th Workshop on Geographic Information Retrieval. p. 7. ACM, New York (2018)
26. Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B.: The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. International Journal of Geographical Information Science **21**(7), 717–745 (2007)
27. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: SIGIR Workshop on Geographic Information Retrieval. pp. 8–10 (2004)
28. Scheider, S., Ballatore, A.: Semantic typing of linked geoprocessing workflows. International Journal of Digital Earth **11**(1), 113–138 (2018)
29. Scheider, S., Nyamsuren, E., Kruiger, H., Xu, H.: Geo-analytical question-answering with gis. International Journal of Digital Earth pp. 1–14 (2020)
30. Wielemann, J.: The semantic structure of spatial questions in Human Geography. Master's thesis, Utrecht University (2019), https://dspace.library.uu.nl/bitstream/handle/1874/384695/Thesis_Report_Joris_Wieleman.pdf
31. Yin, Z., Zhang, C., Goldberg, D.W., Prasad, S.: An NLP-based question answering framework for spatio-temporal analysis and visualization. In: Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis. pp. 61–65. ACM, New York (2019)
32. Zelle, J.M., Mooney, R.J.: Learning to parse database queries using inductive logic programming. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. pp. 1050–1055. AAAI Press (1996)