Spatial Information Extraction from Text Using Spatio-Ontological Reasoning

Madiha Yousaf
University of Bamberg, Germany
madiha.yousaf@uni-bamberg.de

Diedrich Wolter
University of Bamberg, Germany
diedrich.wolter@uni-bamberg.de

Abstract

This paper is involved with extracting spatial information from text. We seek to geo-reference all spatial entities mentioned in a piece of text. The focus of this paper is to investigate the contribution of spatial and ontological reasoning to spatial interpretation of text. A preliminary study considering descriptions of cities and geographical regions from English Wikipedia suggests that spatial and ontological reasoning can be more effective to resolve ambiguities in text than a classical text understanding pipeline relying on parsing.

2012 ACM Subject Classification Computing methodologies → Information extraction

Keywords and phrases spatial information extraction, geo-referencing, spatial reasoning

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.71

Category Short Paper

Funding This work is supported by the Deutsche Forschungsgemeinschaft (DFG), priority program Volunteered Geographic Information. Financial support is gratefully acknowledged.

1 Introduction

We are involved with a project that aims to develop an automated system capable of interpreting spatial language for resolving place descriptions. While ‘place’ is an inherently complex and elusive concept (cp. [3, 7]), we take a pragmatic approach here: Given a natural language description like “the campground south of Bamberg, near the river”, we seek to identify geographic entities in the OpenStreetMap (OSM) data base[1] that match noun phrases occurring in the sentence that refer to real-world entities, in our example thus identifying a campground, an entity named Bamberg, and a river. Automated interpretation of text is still a challenging task, mainly because of language parsing and the ambiguity of names and human conceptualization. While ambiguity of named entities can be tackled by considering any entity with a matching name found in the database and then applying ranking techniques based on geographic scope [1], there are no easy solutions to tackle failed attempts to parse a piece of text. Ambiguity resolution can be regarded as a task of reasoning since the goal is to identify a single interpretation from a set of candidates which is jointly agreeable with all information given. We are motivated to investigate to which extent reasoning about spatial and ontological properties is also capable of overcoming problems with natural language

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

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 71; pp. 71:1–71:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

parsing. Therefore, we investigate a radical attempt that only performs a very simple analysis of the input text, generating many interpretation candidates. We then apply reasoning to single out the interpretation that is most agreeable. This paper presents our method and a preliminary study that suggests spatio-ontological reasoning is offering powerful means for resolving ambiguous interpretations that can outperform classic interpretation pipelines built around natural language parsing.

2 Approach and Discussion of Related Work

We seek to identify named and unnamed entities in a piece of text. While geo-referencing named entities considers names and spatial relations to other entities [1], dealing with unnamed entities presents a special case that can only exploit spatial constraints and maybe type information. We can thus regard both cases jointly as tasks of ambiguity resolution. One approach is geographic scope resolution which allows potential interpretations to be restricted to within a known scope. Whereas Andogah et al. propose a method based on a set of pre-defined geographic scopes [1], Richter et al. [5] consider granularity effects caused by object types. They state that knowing the finest possible level of granularity with respect to a general ontology of spatial entities is helpful for resolving place descriptions. Both ideas can be integrated by attuning the semantics of relations and queries to focus on results that fit a scope indicated by type and location of geographic entities appearing in the same text. For example, the semantics of "near" can be set according to the granularity of objects and their geographic scope. Exploiting such context information presents a chicken-and-egg problem, though: information obtained by resolving entities is to be employed simultaneously to resolving the entities. This motivates an approach using logic programming techniques since dependencies can be expressed in a declarative manner, abstracting from algorithmic realization. The declarative representation can be regarded as a constraint satisfaction problem (CSP) in which variables correspond to spatial entities. A solution to the CSP is obtained when all variables are geo-referenced by matching them to a spatial database.

Interpretation of a place description can be regarded as a simple cognitive simulation of language interpretation, similar to Tschander et al. [6] who describe an artificial agent capable of following route instruction, using a conceptual-level instruction language. In contrast to that agent, we are mostly interested in interpreting describing statements like "campground south of Bamberg", rather than processing incremental instructions like "take road R123 south". Therefore, we are not incrementally interpreting a place description, but aim to build a single declarative description from a single description.

Formal ontologies have been claimed to offers adequate means to represent semantic commitments of a spatial language phrase [2]. Likewise, we employ an ontology-like representation to augment the semantic representation of words (the lexicon). However, we have chosen not to employ formal ontology techniques for two reasons. First, truth semantics of classical ontology languages is binary, i.e., entities belong to a certain class or they do not. In the case of spatial entities and concepts, such crisp classification may be hard to achieve and concepts may vary across individuals. Instead, concepts or relations like 'near' may me more adequately represented using a semantic capturing vagueness, e.g., using Fuzzy or probabilistic models. Second, existing ontology languages do not support the spatial domain and manifold spatial relations to the extent required to empower spatial reasoning for computing likely interpretations of a locative phrase.

Since we are involved with natural language texts, application of natural language parsing techniques appears reasonable. Several works make use of different parser and their

	input	Bamberg	is	a	town	north	of	Nuremberg.
1.	**POS tagging**	Bamberg:NE			town:N	north:R		Nuremberg:NE
2.	**named entity resolution**	$\{\text{ID0}, \text{ID1}, \ldots\}$			town:N	north:R		$\{\text{ID8}, \text{ID9}, \ldots\}$
3.	**ontological annotation**	$\{\text{ID0}, \text{ID1}, \ldots\}$			settlement	north:R		$\{\text{ID8}, \text{ID9}, \ldots\}$
4.	**logic program generation**	$(\text{isa}(\text{ID0}, '\textit{settlement}') \wedge \text{northOf}('\textit{settlement}', \text{ID9})) \vee$ $(\text{isa}(\text{ID1}, '\textit{settlement}') \wedge \text{northOf}('\textit{settlement}', \text{ID9})) \vee$ $\ldots \text{northOf}(\text{ID0}, \text{ID8}) \wedge \ldots) \vee \ldots$						

Figure 1 Example of processing steps in generation phase of information extraction (NE: named entity, N: noun, R:relation, ID:denotes reference to objects in OSM database).

modules (like Standford NLTK chunking or the dependency parser[2]) for relating the objects and relations between them. However, as we are only interested in spatial entities which correspond to nouns in the input text and the relations holding between them, a parser which has to take the verb of a sentence as starting point may not be necessary. Moreover, we found that no freely available parser was able to resolve references in the text correctly. A wrongly identified reference can easily inhibit correct interpretation of a sentence. As our experiments discussed further below reveal, wrongly identified references are a common problem. By contrast, an unidentified reference can often be inferred from context. The basic idea of our approach is thus to generate all candidate interpretations of references and then apply reasoning to single out the most likely interpretation.

3 Processing Pipeline

The basic idea of our method is to use a set of logical statements as an intermediate representation that over-generalizes information expressed in a sentence. Then, spatio-ontological reasoning is applied to prune off implausible interpretations. The method can therefore be regarded as exhaustive search consisting of a *generation* and *pruning* phase. Both phases rely on the same sources of information:

- an ontology of geographic entity types
- a geographic data base (OpenStreetMap) providing information about entity names, their type with respect to the ontology, and associated geographical information
- a lexicon comprising all nouns that represent geographic entity types and all spatial relations

In the *generation phase* (see Fig. 1 for an example), we process a sentence as follows:

1. Perform part-of-speech (POS) tagging by applying named entity recognition using the geographic database and checking for occurrence in the lexicon. All recognized words are labeled with their category, all other words are discarded. To handle composite expressions of several nouns, (e.g., "art gallery"), nouns immediately following one another get joined and treated as a single noun. In case of ambiguities at this or any later point, all possible options are stored.
2. For all named entities, possible interpretations from the geographic database are retrieved. For example, in case of Bamberg, we would obtain an OpenStreetMap entity referring to the city of Bamberg, depicted as ID0 in Fig. 1, and to the corresponding district

[2] `https://nlp.stanford.edu/software/`

of Bamberg, ID1, both for Bamberg, Germany and for Bamberg, SC, USA (creating ambiguity in the extracted information).
3. For all nouns ontological type information is obtained from the lexicon. Nouns are then replaced by their ontological type. Every noun is assumed to either represent an unnamed entity (e.g., "**park** in the town of Bamberg") or a type designator for another noun or named entity (e.g., "park in the **town** of Bamberg").
4. Possible interpretations are determined as disjunctions by compiling interpretations of words and references of relations:
   - For every relation, a term is constructed combining any word (noun or named entity) appearing before the relation with any word occurring after the relation. The designator for each relation is retrieved from the lexicon.
   - For every noun an ontological "is-a" relation is generated in reference to any other noun or named entity, e.g., is-a(Bamberg, town).

In the *pruning phase* every conjunctive term generated in the generation phase is processed individually, see also Fig. 2 for illustration. A term gets discarded if

- a single noun occurs simultaneously in a "is-a" and a spatial relation, i.e., it would represent ontological information and an unnamed entity simultaneously,
- a noun or named entity in the input is not contained in at least one relation,
- or the grouping of relations violates word order in the input sentence. We disallow for relational statements $r(w_a, w_b) \wedge r'(w_c, w_d)$ if the position in the sentence (denoted Pos()) is in crossed order, i.e., it violates $\mathrm{Pos}(w_a) < \mathrm{Pos}(w_c) \Rightarrow \mathrm{Pos}(w_b) \leq \mathrm{Pos}(w_d)$. For example, in "Bamberg is a town north of Nuremberg, on the river Regnitz" interpretations containing $\mathrm{isa}('Bamberg','river') \wedge \mathrm{isa}('town','Regnitz')$ get discarded.

After the pruning phase, we search for the conjunctive term which can best be satisfied. This means, for unreferenced nouns a suitable instantiation from the geographic database is searched that agrees with the relations–agreement is measured gradually and summed up. Also, relations between named entities and/or referenced nouns are tested. In case of the example in Fig. 1 we would only find for the entity representing Bamberg in Germany a matching entity Nuremberg such that Bamberg is located north of Nuremberg. The ontological constraint saying Bamberg is a settlement would only be fulfilled for the city of Bamberg, not the administrative region. We thus arrive at the desired interpretation.

4 First Findings

We collected a corpus of place descriptions from English Wikipedia by selecting sentences which present mainly spatial information. We have used the summary part from Wikipedia articles about geographical entities to collect the corpus and have applied our approach to 50 sentences. We also test natural language parsing using the Stanford NTLK parser on the corpus and investigate parser output. One aim of the study is to compare the amount of ambiguities introduced by our over-generalizing method of information extraction to wrongly identified references by the parser. Also, we are interested to learn what kind of spatial and ontological reasoning is required to interpret the output of our approach.

Let us start by considering a first example shown in Fig. 2 showing relations extracted by the parser and by the generation method. For clarity of presentation, no entities were replaced by OpenStreetMap references and no nouns were replaced by ontological types. We write $r(\{n_1, n_2\}, \{n_3, n_4\})$ as shorthand notation to denote that all four interpretations $r(n_1, n_3), r(n_2, n_3), \ldots$ are considered. As can be seen, the parser identifies that the town mentioned is located in Upper Franconia, but it does not make the relation between the

$$\text{“}\overbrace{\text{Bamberg}}^{B}\ \underline{\text{is a}}\ \overbrace{\text{town}}^{T}\ \text{in}\ \overbrace{\text{Upper Franconia}}^{\text{UF}},\ \overbrace{\text{Germany}}^{G},\ \underline{\text{on}}\ \text{the}\ \overbrace{\text{river Regnitz}}^{R}$$
$$\underline{\text{close to}}\ \text{its}\ \underbrace{\text{confluence}}_{C}\ \text{with the}\ \underbrace{\text{river Main}}_{M}.\text{”}$$

parser output	$\text{in}(T, \text{UF}), \text{on}(G, R)$
generation phase	$\text{is-a}(\{B, \text{UF}, G, R, M\}, C), \text{is-a}(\{B, \text{UF}, G, R, M\}, T),$ $\text{in}(\{T, B\}, \{M, C, R, G, \text{UF}\}), \text{on}(\{G, \text{UF}, T, B\}, \{M, C, R\}),$ $\text{close}(\{B, T, \text{UF}, G, R\}, \{C, M\})$
pruning phase	
ontological	$\text{is-a}(\cancel{\{B, \text{UF}, G, R, M\}}, C), \text{is-a}(\{B, \cancel{\text{UF}, G, R, M}\}, T),$
spatial	$\text{in}(\{T, B\}, \{\cancel{M, C, R,}\ G, \text{UF}\}), \text{on}(\{\cancel{G, \text{UF}, T,}\ B\}, \{\cancel{M, C,}\ R\}),$
ordering	$\text{close}(\{B, \cancel{T}, \text{UF}, G, R\}, \{C, M\})$

Figure 2 Example of pruning using spatio-ontological reasoning.

named entity ‘Bamberg’ and town explicit. Also, the parser commits wrongly to claiming Germany would be located on the river Regnitz. By contrast, exhaustive search contains all correct interpretations by construction, but also several statements not following from the input text. Applying ontological reasoning one immediately identifies that only named entity Bamberg is of type town. Spatial reasoning reveals, for example, that Upper Franconia is neither located on river Regnitz nor Main. As our approach is not yet prepared to handle geographic names like “Upper Franconia, Germany”, we miss this important piece of information.

In addition to above example, some candidate interpretations generated by exhaustive search that are not valid interpretations of the input text take more effort to reject. In case of *The Historical Museum of Bamberg is a museum located in the Alte Hofhaltung next to the city cathedral*, the interpretation in(’museum’, ’city cathedral’) cannot easily be rejected if the geographic database also includes a museum in the city cathedral. If, during search for the most likely interpretation, the unintended reference is accepted, then order constraints inhibit any further connection to the named spatial entity “Alte Hofhaltung” (Old Court). So in this case we are relying on preferring the larger set of jointly possible interpretations that involves in(’museum’, ’Alte Hofhaltung’) and next_to(’Alte Hofhaltung’, ’city cathedral’) over just in(’museum’, ’city cathedral’). We have tested our corpus and in initial testing we have found out that in 50% sentences it is providing us information that is not present in the sentence but generated by the algorithm. In many examples, these facts are not incorrect like in(’Bamberg’, ’Germany’) from the example in Fig. 2. While these unintended but correct interpretation candidates did not inhibit a correct manual interpretation, it remains an open question whether this will hold for automated interpretation on a larger corpus.

Now looking at the parser outputs, we can clearly see that it provides us with limited information. In particular relations from complex language constructs are missing. In case of the output of Fig. 2, the relations apply to different entities which inhibits any chaining by means of reasoning. All in all, the parser is not able to provide a densely connected set of facts that would make spatial or ontological reasoning effective. Carrying out spatial and ontological reasoning manually and comparing residual errors after processing the output of exhaustive search with facts extracted from the parser, we cannot rule out all ambiguous interpretations in 25% of the sentences, but we are facing wrong outputs from the parser in 50% of the cases.

5 Conclusion and Next Steps

This paper outlines an approach to information extraction from text which does not rely on natural language parsing, but employs a simple part-of-speech tagging and applies spatial and ontological reasoning for interpretation. Making spatio-ontological reasoning an explicit step in the interpretation also enables consideration of contextual dependencies. Clearly, exhaustive search does not tackle the fundamental problem of language understanding, but it relies on the assumption that the largest set of statements that can be matched to a geographic database corresponds to the intended interpretation. While our approach is unable to deal with negation or complex language structures, it may indeed be sufficient for typical descriptive texts. In a manual comparison using sentences from English Wikipedia that describe geographic entities we see that reasoning is able to prune off most invalid interpretations, whereas natural language parsing results in some wrong commitments one is unable to recognize in a later processing step.

Before embarking on a comprehensive study to analyze the new method, a comprehensive lexicon and knowledge base have to be prepared and reasoning methods to be automated. Information required to build lexicon and knowledge base are readily available using data sources such as WordNet[4] and OpenStreetMap, yet these need to be linked on a semantic level. We are currently working on implementing the automated reasoning method using these sources in order to arrive at a spatial interpretation of the constraints. To make the approach efficient, a query strategy will be required to avoid costly queries by serializing queries and by focusing on reasonable candidate locations.

References

1 Geoffrey Andogah, Gosse Bouma, John Nerbonne, and Elwin Koster. Placename ambiguity resolution. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 4–10, 2008.

2 John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174:1027–1071, 2010.

3 Brandon Bennett and Pragya Agarwal. Semantic categories underlying the meaning of 'place'. In Stephan Winter, Matt Duckham, Lars Kulik, and Ben Kuipers, editors, *Spatial Information Theory: Proceedings of the 8th International Conference, (COSIT-2007), Melbourne, Australia*, volume 4736 of *Lecture Notes in Computer Science*, pages 78–95, Berlin, Heidelberg, 2007. Springer-Verlag.

4 Christiane Fellbaum, editor. *Wordnet: An Electronic Lexical Database.* MIT Press, 1998.

5 Daniela Richter, Stephan Winter, Kai-Florian Richter, and Lesly Stirling. Granuality of locations referred to by place descriptions. In *Proceedings of Workshops and Posters at the 13th International Conference of Spatial Information Theory (COSIT 2017)*, computer environments and urban systems, pages 88–99. Elsevier, 2012.

6 Ladina B. Tschander, Hedda R. Schmidtke, Carola Eschenbach, Christopher Habel, and Lars Kulik. A geometric agent following route instructions. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl F. Wender, editors, *Spatial Cognition III*, volume 2685 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 99–111, Berlin, 2003. Springer.

7 Stephan Winter and Christian Freksa. Approaching the notion of place by contrast. *Journal of Spatial and Information Science (JOSIS)*, 5:31–50, 2012.