Detecting the Geospatialness of Prepositions from Natural Language Text

Mansi Radke
Computer Science and Engineering Department, Visvesvaraya National Institute of Technology, Nagpur, India
mansiaradke@gmail.com

Prarthana Das
Computer Science and Engineering Department, Visvesvaraya National Institute of Technology, Nagpur, India
musiciselixir@gmail.com

Kristin Stock
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand
k.stock@massey.ac.nz

Christopher B. Jones
School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
jonescb2@cardiff.ac.uk

Abstract

There is increasing interest in detecting the presence of geospatial locative expressions that include spatial relation terms such as *near* or *within* <*some distance*>. Being able to do so provides a foundation for interpreting relative descriptions of location and for building corpora that facilitate the development of methods for spatial relation extraction and interpretation. Here we evaluate the use of a spatial role labelling procedure to distinguish geospatial uses of prepositions from other spatial and non-spatial uses and experiment with the use of additional machine learning features to improve the quality of detection of geospatial prepositions. An annotated corpus of nearly 2000 instances of preposition usage was created for training and testing the classifiers.

2012 ACM Subject Classification Computing Methodologies → Artificial Intelligence; Computing Methodologies → Natural Language Processing

Keywords and phrases spatial language, natural language processing, geospatial language

Digital Object Identifier 10.4230/LIPIcs.COSIT.2019.11

Category Short Paper

1 Introduction

Automated recognition and disambiguation of geographic references in text documents has received considerable attention in recent years, often with the motivation of indexing the documents with regard to geographic space. The methods used to date have been dominated by a focus on identifying geographic names, i.e. toponyms, and using these directly as the basis for geographic footprints for text expressions or entire documents. The assumption however is that the references are absolute in the sense that the toponym provides the actual location referred to. While this is a reasonable default assumption, it is very common to refer to locations in an indirect manner using spatial relations, such as *near*, *at*, *close to*, *north of* etc., relative to a reference location. These expressions often take the form of triples of a subject (or located object), the spatial relation and an object (the reference location), as in "St Mary Church near Times Square." While some authors have proposed methods for modelling vague spatial relations such as *near* (e.g. [7, 10, 11]), relatively little work has been done on the basic, initial problem of reliably identifying the presence of relative

14th International Conference on Spatial Information Theory (COSIT 2019).
Editors: Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart;
Article No. 11; pp. 11:1–11:8
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

locational descriptions in natural language texts ([3, 5, 6, 8]). Effective methods for doing this are required as part of the process of extracting and interpreting indirect geographic references and to retrieve other geospatial facts that associate an event or some other object with a reference location, as for example in "Roald Dahl was born in Cardiff". Locational description detection methods are also required for automatic creation of test collections that can be used in developing and evaluating methods for spatial relation extraction and for modelling the use of individual spatial relations, e.g. [9]. In this paper, we present methods for automatic detection of spatial relational terms in sentences, in particular prepositions, that are used specifically in a geospatial sense and we distinguish these from prepositions that have other spatial senses and from prepositions that have no spatial meaning. We are interested in the ability to distinguish between spatial and geospatial senses of prepositions, as this is important for detecting text that can be georeferenced and thus mapped on a geographical scale (in contrast to text that describes a location inside a room, or on a person's body), a goal that is useful in a wide range of application areas.

The approach adopted is here applies the spatial role labelling method of [3]. That work aimed to detect all three components of spatial relational expressions which were referred to as the trajector, i.e, the located object, spatial indicator, i.e. the individual preposition that serves as spatial relation, and the landmark which is the reference location. Here we use their preposition disambiguation method, which was employed as part of a pipeline approach to detection of triples. The method was tested in [3] only for the purpose of detecting generic spatial prepositions, which might or might not be geospatial. Here we train the classifier on sentences containing a preposition that is used either in a geospatial sense, a spatial but not geospatial sense, or in a sense that is not spatial in any respect. We also experiment with modifying the classifier for geospatial prepositions to take account of other evidence that indicates the presence of place names and geographic feature types.

For the purpose of evaluating the approach, we have created a corpus of 1876 instances of preposition usage that have been manually labelled as geospatial, spatial (but not geospatial) and non-spatial. These prepositions occur within 674 sentences.

In the remainder of the paper Section 2 describes related work, Section 3 explains the methodology in detail, while Section 4 gives the details of the data set used and the experiments performed. Section 5 concludes the paper, pointing out some directions for future work.

2 Related work

A method specifically designed to detect whether a preposition has a spatial sense was presented by Kordjamshidi et al. [3] in a paper on spatial role labelling in the context of relation extraction. The paper focused on the three roles of trajector (located object), spatial indicator (spatial relation) and landmark (reference location). Two approaches to spatial role labelling were presented. In the first approach, called the pipeline approach, an input sentence is passed to the first stage of the pipeline which tokenizes the sentence and passes each token to a Part of Speech (POS) tagger. The sentence is also processed by a dependency parser and a semantic role labeller (the LTH software from [1]). If a preposition is identified by the POS tagger, a Naive Bayes classifier is used to make a decision on whether it is used in a spatial sense. The features used by the classifier are based on output from the POS tagger, the dependency parser and the semantic role labeller. For this stage of identifying the spatial sense of a preposition, an F1 score of .88 was achieved for the TPP dataset [4] with 10 fold cross validation. If the preposition is determined to have a spatial sense, then it is passed to a second stage of the pipeline which identifies the trajector and the landmark with respect to the spatial indicator. This second stage uses probabilistic graphical models, in particular a Conditional Random Fields classifier, which again takes a variety of features generated by the initial parsing of the sentence. A triple of the form <Trajector, SpatialIndicator, Landmark> is returned as output by the pipeline. The second approach offered by Kordjamshidi et al. [3] uses joint learning in which all three of trajector, spatial indicator and landmark are detected simultaneously.

A method for detecting just the spatial relation and the reference object of spatial relations was described by Liu [5] where these partial relations were described as degenerate locative expressions (DLE). The approach is analogous to methods of Kordjamshidi et al., though they employed a smaller set of features for machine learning, that did not include dependency relations or semantic roles. An evaluation of the method in [6] obtained an F1 score of .76 when applied fully automatically to their TellUsWhere corpus on which it was trained. Note that no distinction was made in that work between geospatial and other spatial senses of prepositions. The method of [5] to extract DLEs was also exploited in Khan et al. [2] in which locative DLEs which explicitly encode spatial relations, with prepositions such as *near* and *in*, were distinguished from partial DLEs where a preposition such as *to* was not regarded as conveying explicit spatial information. A rule based approach was employed to extend the latter to an explicit spatial DLE when it was used as part of a spatial relation such as *next to.* This technique was part of a procedure to extract spatial triples by matching structures from the Stanford parser, of the form <governor, preposition, dependent>, with locative DLEs that used the same preposition. The governor would then serve as the located object of a spatial triple.

As part of a process of creating a corpus of geospatial sentences, Stock et al. [8] employed a set of language patterns to detect various ways in which geospatial information is described. This included a pattern to recognise when a place name or place type is preceded by a spatial relation which could be a preposition (though other parts of speech were also considered to represent spatial relations). They obtained a precision of 0.66 when applying these methods to detect geospatial expressions. A specialized collection of spatial relational expressions was created by Wallgrun, Klippel and Baldwin [9]. They used search patterns to query the web to find expressions that contained any of the three relations of *near*, *close* and *next to.* Their approach therefore constrained the results to include the specified spatial relation. They also confined the expressions to include specified types of located and reference objects. Our work differs from that in allowing any spatial relation that is classed as a preposition and in using a machine leaning approach to determine the geospatial or other spatial sense of the preposition.

3 Methods

3.1 What is a geospatial sense?

In order to distinguish here between geospatial, other spatial and non-spatial uses of prepositions, we employ a simple definition of a geospatial relation as one in which the preposition has a spatial sense and the reference object to which the preposition applies is a geographic feature, as in a named place or a geographic feature type. The reference object is normally expected to be outdoors. If it is part of a building it is expected to be an exterior part. We impose no constraint on the nature of the located object. If a preposition has a spatial sense but the reference object is not geographic then it is classed as spatial. If the preposition has no spatial interpretation then it is classed as neither geospatial nor spatial.

Examples of the kinds of expressions that appear on our corpus include the following, with preposition senses according to our annotation scheme (described above) shown in angular brackets:

- "And now on <non-spatial> a clear morning Graham Little and I are sitting at <geospatial> the bottom of (spatial) the wall fit and ready to go and the wall is plastered with <non-spatial> verglas."
- "In <non-spatial> a minute she had rushed from <geospatial> the house and was running down <geospatial> the garden"

3.2 Classifying prepositions as geospatial or spatial

In this work, we modify the first step of the spatial role labelling pipeline method of [3], i.e. their method for detecting the spatial sense of prepositions, by adding additional features for machine learning. The features used in the original classifier are listed in Table 1. As indicated above these are obtained from a combination of a POS tagger, a dependency parser and a semantic role labeller. The Part-Of-Speech Tagger (POS Tagger) assigns parts of speech to each word, such as noun, verb, adjective, etc. Dependency parsing assigns a syntactic structure to a sentence. The most widely used type of syntactic structure is a parse tree which can be useful in various applications such as grammar checking, but here it plays a critical role in the semantic analysis stage. In natural language processing, semantic role labeling (also called shallow semantic parsing) is a process that assigns labels to words or phrases in a sentence to indicate their semantic role, such as that of an agent, goal, or result. It consists of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. We experiment with using just these features, but we also extend the method to add additional features that indicate whether a place name or a geographic place type is present in the expression that includes the target preposition. The presence of a place name is detected with the Geonames gazetteer, while the presence of a place type is detected with a dictionary of geographic place types. The expat application was used to generate these features (location and gnn patterns).

We used a Naive Bayes multi-class classifier with three output classes of geospatial, spatial but not geospatial, and neither geospatial nor spatial. We also used Naive Bayes binary classifiers for each one of these three classes *vs* the other two classes.

4 Experimental Set Up

4.1 Data set and its Annotation

Our dataset of 674 sentences was derived from two sources. 185 of the sentences came from the source of about 26,000 sentences that were used in the process of creating the Nottingham Corpus of Geospatial Language (NCGL) [8]. These sentences were harvested from the web using the algorithm described in [8], and was thus biased towards retrieving geospatial content, but also included spatial (but non-geospatial) expressions as well as some uses of prepositions that are non-spatial in any sense. The remainder of our collection is a sample of the TPP dataset of sentences produced for the preposition project (see Litkowski and Hargraves [4]). That dataset includes many examples of both spatial and non-spatial uses of prepositions, though relatively few of them have a geographical context.

Many of the sentences include multiple prepositions and so in order to annotate the sense of the individual prepositions we created a distinct instance of a sentence for each preposition

preposition	the preposition itself
preposition	the lemma of the preposition
preposition	the POS tag of the preposition
preposition	the DPRL of the preposition
preposition	the semantic role label of the preposition
preposition	the sense of the preposition if assigned
preposition	the argument of the preposition in the SRL output
head1	the head1 itself
head1	the lemma of head1
head1	the POS tag of the head1
head1	the DPRL of the head1
head1	the semantic role label of the head1
head1	the sence of the head1 if assigned
head1	the argument of the head1 in the SRL output
head2	the head2 itself
head2	the lemma of head2
head2	the POS tag of the head2
head2	the DPRL of the head2
head2	the semantic role label of the head2
head2	the sence of the head2 if assigned
head2	the argument of the head2 in the SRL output

Table 1 Features from [3] used in detecting the sense of a preposition

that it contained (as determined by a POS tagger). We considered a tuple <Sentence, Preposition> as a unique instance. So, if a sentence instance s had two prepositions $p1$ and $p2$, we created two instances from it, namely $< S, p1 >$ and $< S, p2 >$. This resulted in 1876 instances (indicating an average of just under three prepositions per sentence). These preposition-specific instances were then manually annotated as either geospatial, spatial (but not geospatial) or non-spatial.

Annotation was conducted through an iterative process that involved all four authors. In the case of the NCGL sentences, one person annotated all sentences, a subset of 100 of which were then checked by two others followed by a discussion of disagreements. A fourth person then re-annotated all of those sentences taking account of issues raised in the discussions. The TPP sentences were annotated by one person, after which one other checked them and highlighted disagreements. The first annotator then revised annotations to respect the result of this discussion. Finally a further stage of re-annotation of subsets of 100 of each of both groups of sentences was performed resulting in inter-annotator agreements of 0.89 for the larger TPP sourced data set and 0.75 for the NCGL sourced data set.

As an example of inter-annotator disagreement, consider the following sentence. "After 50m, you will reach a road with wide verges where you turn left toward Lambley." The first annotator marked *after* as non-spatial in sense. The second annotator noted that here *after* is used to represent the geospatial arrangement of different locations, and the latter sense was adopted for the final data set. In another example, in the phrase "Republic of China", the preposition *of* was marked spatial by one annotator, as "China" is a geographical place name, while the other annotator considered it as non spatial since "Republic of China" is an administrative entity. We adopted this latter annotation for the final data set.

4.2 Experiments performed

Before we present our results, we mention the balance of the classes in the datatset used. Out of the total preposition instances (1877), the number of instances marked as non-spatial was 770, the number of instances marked as spatial was 773, and the number of instances marked as geospatial was 334.

Table 2 Features used in experiments

Kord	All features used for preposition sense detection in [3]
Kord-Geo	The features from Kord plus the number of placenames and the number of geographic feature types found in the head words of the preposition
Kord-Geo-S	The features from Kord plus the number of place names and the number of geographic feature types found within the entire sentence in which the preposition occurs
Kord-Geo-All	The features from Kord-Geo-S plus the sum of the numbers of place names and a binary value of true if either a place name or a geographic feature type is present
Geo-Baseline-S	The number of place names and the number of geographic feature types found within the entire sentence in which the preposition occurs

Table 3 Results for 3-class classifier predicting geospatial, spatial (but not geospatial) or neither

	Geospatial			Spatial			Neither		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Kord	0.442	0.578	0.501	0.747	0.744	0.745	0.763	0.664	0.710
Kord-Geo	0.514	0.614	0.559	0.751	0.762	0.757	0.772	0.696	0.732
Kord-Geo-S	0.566	0.638	0.600	0.732	0.802	0.765	0.783	0.665	0.719
Kord-Geo-All	0.600	0.692	**0.643**	0.749	0.797	**0.772**	0.796	0.692	**0.740**

Table 4 Results for three 2-class classifiers predicting geospatial, spatial (but not geospatial) and neither

	Geospatial			Spatial			Neither			Spatial or Geospatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Kord	0.370	0.647	0.471	0.696	0.790	0.740	0.762	0.751	0.756	0.828	0.836	**0.832**
Kord-Geo	0.423	0.680	0.521	0.704	0.798	0.748	0.760	0.755	0.757	0.830	0.835	**0.832**
Kord-Geo-S	0.480	0.704	0.570	0.688	0.846	**0.759**	0.755	0.753	0.754	0.829	0.830	0.829
Kord-Geo-All	0.542	0.728	**0.621**	0.672	0.837	0.745	0.750	0.771	**0.761**	0.838	0.821	0.829
Geo-Baseline-S	0.625	0.419	0.502	0.494	0.889	0.635	0.422	0.326	0.368	0.595	0.689	0.639

Several experiments were conducted with a Naive Bayes classifier to evaluate the methods described above (note that the original method from [3] uses this classifier for determining the sense of a preposition). In the first experiment (Table 3) a multi-class Naive Bayes classifier was used to predict each of the three classes of geospatial, spatial (but not geospatial) and neither. There were several versions of the classifier that use different combinations of features (summarised in Table 2). One of these (Kord) just uses the features from [3] described above. It resulted in an F1 value of 0.50 for the geospatial class and better values of 0.745 for spatial and 0.710 for neither. This was extended by adding the two features of the number of place names and number of geographical features detected in the head words of the preposition that is being tested (Kord-Geo). Note that the head words are among the features generated by the procedure used in [3]. They correspond to the subject and object of the preposition. A further variation (Kord-GeoS) records these latter numbers at the sentence level, which was found to improve upon the performance when only observing head words (though note that the quality of performance will depend upon the performance of the script to detect place names and geo-feature types). Experiments to employ features consisting of a binary value to record whether a place name or geo-feature were present and, separately, of a value that is the sum of the numbers of place names and geo-feature types, did not improve on sentence level performance and are not listed here. However, combining these latter data items with those in Kord-Geo-S did provide an improvement (referred to as feature set Kord-Geo-All) with an F1 for Geospatial of 0.643.

In addition to the three class classifiers we implemented several 2-class classifiers (see Table (4) with target classes of geospatial (*vs* spatial or neither), spatial *vs* (geospatial or neither) and neither (*vs* geospatial or spatial). Just as with the 3-class classifiers we used either just Kordjamshidi features (Kord), and place name and geographic features from the preposition's head words (Kord-Geo) and from the whole sentence in which the preposition occurred (Kord-GeoS). We also tested the method using Kord-Geo-All features, which gave the best 2-class performance for geospatial sense with an F1 of 0.621 but this did not improve on the result from the 3-class classifier. Output from the 2-class classifiers also included the complement of the Neither class, i.e. detection of prepositions that are either used in a spatial or a geospatial sense, which is equivalent to preposition classification task in [3]. We obtained an F1 value of 0.832 when using just the original features from [3].

As a baseline (Geo-Baseline-S) we implemented a Naive Bayes method for detecting whether a preposition has a geospatial sense, that uses, as machine learning features, just the presence of a place name and the presence of a geographic feature type. This was conducted at the preposition specific level, in which their presence was recorded only in the head words of the preposition, and at the level of whether they occurred anywhere in the sentence. The latter approach gave the better performance with an F1 of 0.502.

5 Conclusions and future directions

In this paper we have experimented with a method for detecting the geospatial nature of prepositions in sentences using a machine learning approach that was developed in [3] for generic spatial role labelling. Using a corpus of sentences annotated as either geospatial, spatial (but not geospatial) or neither geospatial nor spatial, we found that, when trained on this corpus, the original method was not able to detect geospatial prepositions with an F1 value greater than 0.50. However, it detected the spatial (but not geospatial) class with F1 of .745 and it detected prepositions that are used with either a geospatial or a spatial sense with an F1 of 0.832. We have adapted the method in an effort to improve its performance for detecting geospatial sense by adding features (for machine learning) that record whether a place name or a geospatial feature type is present in the head words that serve as subject and object of the preposition or, alternatively, whether they are present in the entire sentence. Using the sentence level features provided better performance with an F1 of 0.643 for geospatial sense. It also resulted in an improvement in detection of the spatial (but not geospatial) class with an F1 of 0.772. It may be noted that a classifier using only the presence of a place name or geographic feature type in the sentence provided better performance than the basic spatial role labelling method.

In future work we will investigate methods to make further improvements to the performance of the methods presented here. In particular we will address a limitation of the current method with regard to detection of place names and feature types by using a richer gazetteer and extending the dictionary of geographical feature types.

References

References

1 Richard Johansson and Pierre Nugues. LTH: Semantic Structure Extraction using Nonprojective Dependency Trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230. Association for Computational Linguistics, 2007.

2 Arbaz Khan, Maria Vasardani, and Stephan Winter. Extracting Spatial Information From Place Descriptions. In *COMP@ SIGSPATIAL*, page 62, 2013.

3 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36, December 2011.

4 Ken Litkowski and Orin Hargraves. SemEval-2007 Task 06: Word-sense Disambiguation of Prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 24–29, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

5 Fei Liu. *Automatic identification of locative expressions from informal text.* Masters by Coursework & Shorter thesis, University of Melbourne, Melbourne, Australia, 2013. URL: `http://minerva-access.unimelb.edu.au/handle/11343/38520`.

6 Fei Liu, Maria Vasardani, and Timothy Baldwin. Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. In *Proceedings of the 4th International Workshop on Location and the Web*, LocWeb '14, pages 9–16, New York, NY, USA, 2014. ACM.

7 V.B. Robinson. Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, 113(1):133–145, 2000.

8 Kristin Stock, Robert C. Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone. Creating a Corpus of Geospatial Natural Language. In *Proceedings of the 11th International Conference on Spatial Information Theory - Volume 8116*, pages 279–298. Springer-Verlag New York, Inc., September 2013.

9 Jan Oliver Wallgrün, Alexander Klippel, and Timothy Baldwin. Building a Corpus of Spatial Relational Expressions Extracted from Web Documents. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, GIR '14, pages 6:1–6:8, New York, NY, USA, 2014. ACM.

10 M. Worboys. Nearness relations in environmental space. *International Journal of Geographic Information Science*, 15(7):633–651, 2001.

11 Xiaobai Yao and Jean-Claude Thill. How far is too far? – a statistical approach to context-contingent proximity modeling. *Transactions in GIS*, 9(2):157–178, 2005.