Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Disambiguating spatial prepositions: the case of geo-spatial sense detection

Abstract

Spatial relations in natural language are frequently expressed through prepositions. Thus, in the locative expressions "New York in the United States" and "the house on the river" the prepositions "in" and "on" respectively serve to communicate the relationships in space between the subject and object of the preposition. Automatic detection of the use of prepositions in a spatial and in particular a geo-spatial sense that refers to geographic context is of interest in supporting automated methods for determining the actual geographic location referred to by locative expressions. This work focuses on disambiguation of prepositions in natural language, with the goal of distinguishing whether a preposition is used in a specifically geo-spatial sense. We conduct machine learning experiments that demonstrate the clear benefit for geo-spatial sense detection of using transformer model deep learning methods when compared with a variety of methods, that include Naive Bayes, Support Vector Machine (SVM) and Random Forest classifiers with hand crafted linguistic features, and a bag of words approach with a meta-classifier that adds geo-spatial features. The best performance was obtained with the BERT-based XLNet transfomer model, with a best precision of 0.96 and and an F1 score of 0.94 when evaluated on a corpus of natural language expressions that were annotated for this task. We also conducted experiments to detect generic spatial sense, in which the best the best F1 score, of 0.95, was again obtained with XLNet.

KEYWORDS:

Preposition disambiguation, geo-spatial sense, generic spatial sense, transfer learning, BERT-based models, geo-spatial corpus, geo-referencing, locative expressions

1 | INTRODUCTION

Natural language texts contain a great deal of geo-spatial information which, if extracted and geo-referenced to locations on the Earth's surface, constitute a massive potential source of data that could be exploited in geographical information systems. Descriptions of locations typically include relationships between some entity or event and a geographical place, where the spatial relationship is very commonly expressed with a preposition [Herskovits, 1987], though parts of speech such as verbs can also be employed [Dittrich et al., 2015]. For example in the locative expression "Otaki Gorge Road near Otaki Forks", the preposition "near" indicates a spatial proximity relationship between the named places "Otaki Gorge Road" and "Otaki Forks" . In spatial relational expressions of this form, "Otaki Gorge Road" is regarded as the located object (or trajector, locatum or figure) while "Otaki Forks" is the reference location (also referred to as a landmark, relatum or ground).

Automated detection of the presence of locative expressions that describe the spatial relationship of one entity to another is a challenge however in that the preposition terms, such as *near*, *in* and *at*, that are so commonly used to convey spatial relations can also be used in a non-spatial sense. Furthermore when a preposition is used in a spatial sense it is very often the case that the spatial sense is not actually geo-spatial. The distinction that we draw here between generic spatial senses and geo-spatial senses is important, because many locative expressions describe non-geographical situations (for example in so-called table-top space) that cannot normally be geo-referenced.

The task of determining the sense of prepositions (or other parts of speech) is an integral component of processes of automated spatial relation extraction that detect the located object, the spatial relational term and the reference object [Kordjamshidi et al., 2011, Rahgooy et al., 2018, D'Souza and Ng, 2015] and is sometimes referred to as spatial indicator classification [Kordjamshidi et al., 2017]. While there is some previous work on the demanding task of extraction of spatial relations in specifically geo-spatial contexts (e.g. Khan et al. [2013], Zhang et al. [2009, 2011], Zenasni et al. [2018]), little progress has been made on development of methods focused specifically on identifying geo-spatial senses of prepositions and distinguishing them from other spatial senses. A study that was focused specifically on this task by Radke et al. [2019] reported relatively poor classifier performance with the best precision only being 0.63. More effective methods would contribute to progress on this broader task of extracting geo-spatial relational expressions.

A significant motivation for working on aspects of geo-spatial relation extraction, such as preposition disambiguation, is that it contributes to the longer term aims of automated geo-referencing of natural language location descriptions that assert spatial relationships relative to a reference place [Liu et al., 2009, Doherty et al., 2011]. Automated detection of geo-spatial locational descriptions and their subsequent geo-referencing has considerable potential value for many real world applications. For example, in disaster management emergency responders can determine where damage has occurred, or where stranded people need to be evacuated from; health workers can extract and map infectious disease information from social media; and ecologists can extract data on the location of sightings of biological species and events from biological records. Notably the related task of coarser-grained geo-referencing of entire textual documents and social media posts has received significant attention in recent years Melo and Martins [2017], Stock [2018].

This downstrean process of geo-referencing textual expressions involves generating map coordinates which then enable spatial indexing and hence efficient access to documents and to entities and events that are mentioned within documents [Wu et al., 2012, Purves et al., 2018]. Achievement of this goal can require contextually-specific interpretation of essentially vague spatial words such as *near*, *at* and *beside*. Models of the use of vague spatial relations may be learnt from multiple examples of uses of spatial relations which in turn will benefit from automated identification of locative expressions and their subsequent extraction, as well as from direct human subjects experiments [Logan and Sadler, 1996, Worboys, 2001, Robinson, 2000, Hall et al., 2011]. Acquisition of substantial knowledge bases of spatial relations between explicitly defined located and reference objects could also facilitate the direct answering within search engines of questions about spatial relationships.

These potential benefits of effective automation of the process of detecting geo-spatial preposition sense, in combination with the limited progress to date in achieving that goal, have motivated the work that we present in this paper. Our main aim is to address the challenge of obtaining high performance in a machine learning classifier to detect the geo-spatial use of prepositions and distinguish that use from other spatial but not geo-spatial uses of prepositions. Specifically we consider the following research questions:

- Q1 Can machine learning methods achieve high performance for precision and recall in the task of detecting the geo-spatial sense of prepositions in locative expressions and distinguishing them from other spatial and non-spatial senses?
- Q2 For the task geo-spatial preposition sense detection, do transformer-based deep learning models outperform classical machine learning methods?

As noted above, preposition sense detection by itself can be regarded as a sub-task of relation extraction and can be treated as part of a relation extraction pipeline or integrated with the detection of located and reference objects. We regard the presentation of effective methods for preposition sense detection in isolation as being of value in that they have the potential to be integrated with geo-spatial relation extraction methods, but they also serve the purpose of helping to identify the presence of sentences that are communicating geo-spatial information [Stock et al., 2022]. They can therefore serve as a filter prior to application of methods that extract entire spatial relations.

We distinguish geo-spatial from generic spatial senses of prepositions based on the reference object. Both geo-spatial and spatial senses involve a preposition that describes the physical configuration of the located object in space, relative to a reference object. However, the geo-spatial sense describes location that can be geo-referenced to the earth with coordinates that can be

determined directly or through co-reference with other text. Thus the reference object is a feature in geographic space, most often a place name (e.g. "Times Square") or a geographic feature type such as *city*. In contrast, the expression "the book is on the table" contains a spatial preposition "on" that describes a spatial configuration that explains where the book is relative to the table, and the sense of "on" is spatial, but not specifically geo-spatial, because the table is not normally regarded as a geographical object for which geographical coordinates would be derived. In contrast, the expression "the church is beside the Waikato River" contains the spatial preposition "beside" that describes the spatial configuration of the church relative to the Waikato River which is a named geographic feature that can be geo-referenced with coordinates from a gazetteer such as Geonames[1]. Note that all geo-spatial senses of prepositions are also spatial, but not all spatial senses of prepositions are geo-spatial (geo-spatial senses are a sub-set of spatial senses).

In order to distinguish geo-spatial from other spatial and non-spatial senses of prepositions we investigate the use of a variety of types of classifier. These include several deep learning transformer models that employ transfer learning [Ruder et al., 2019] that benefit from prior training on very large corpora, and Naive Bayes, Support Vector Machines (SVMs) and Random Forest classifiers. For the latter non-deep learning classifiers we experiment with several types of features, including the linguistic features of Kordjamshidi et al. [2011] combined with features that record the presence of words that are place names or types of geographic feature. We also use Bag of Words (BoW) represented by a vector with a dimension for each word in the corpus The BoW classifiers learn, in training, associations between particular word usages and the given sense of a preposition.

Our deep learning transformer methods include several BERT-based models [Devlin et al., 2018], the input features of which are the embeddings of the words in the context of the word to be classified, where the word embeddings are multi-dimensional vectors that can be regarded as encapsulating the sense of each word. Our experiments include an adaptation of the textual input to BERT-based models to include tags of place name words, where present, which we show to provide significant advantage.

In the absence of existing labelled data applicable to our main task of geo-spatial sense detection, we annotated two datasets that combine geo-spatial, other-spatial and non-spatial expressions.

The main contributions of the paper can be summarised as follows:

- We demonstrate that BERT-based transformer deep learning classifiers can provide high precision and recall in detecting the geo-spatial sense of prepositions and distinguishing the sense from generic spatial senses.
- We show that a transformer deep learning classifier for detecting the geo-spatial sense of prepositions is superior to those using hand crafted linguistic features, bag of words features and features representing the presence of place names and geo-feature types.
- Our use of the BERT-based transformer classifiers is shown to be enhanced by adding tags to the input text to indicate the presence of place names.
- We publish two new corpora with annotations of prepositions as either geo-spatial, other-spatial (but not geo-spatial) or non-spatial.

The remainder of the paper is organised as follows. Section 2 reviews related work, while Section 3 explains the methods used in this work. Section 4 presents experimental results including a description of the data sets used, their annotation, and the experimental results obtained for the various methods. Section 5 concludes the paper pointing out directions for future work.

2 | RELATED WORK

A number of authors have discussed the nature and properties of spatial language [Coventry and Garrod, 2004, Talmy, 2000, Jackendoff, 1983, Levinson, 2003]. The term *locative expression* has been used to refer to "an expression involving a locative prepositional phrase together with whatever the phrase modifies (noun, clause, etc.)" [Herskovits, 1987, p.7]. The most common syntax of a locative expression consists of a preposition and two noun phrases. In the case of a spatial preposition, the preposition describes the spatial relation (configuration in space) between the objects referenced in the subject and object noun phrases. The combination of the preposition and its reference object (the object noun phrase of the preposition) is known as a prepositional phrase. Examples of prepositional phrases are *across the road*, *along pavements* and *underneath the piers*. This typical association of spatial prepositions with reference objects places them in the class of transitive prepositions that take an object.

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

It distinguishes them from intransitive prepositions that do not have arguments (as for example the use of *in* in the phrase *she is in today*), and selected prepositions where the meaning is determined by a governor word that is usually a verb (as for example the use of *with* in *dispense with cutlery* [Baldwin et al., 2009]).

The distinction between spatial and non-spatial language is not always straightforward, as the same prepositions may be used by both kinds of expressions. For example in the use of the preposition *in* in the phrases *She lived in London* and *She was born in 1772*. Other examples are due to metaphorical uses as in *she's really in a pickle* and *He liked to throw his weight around*. These various distinctions have been addressed by a number of authors such as Tyler and Evans [2003] and Coventry and Garrod [2004].

The common association of prepositions with governor and object words has led to a focus on the use of these head words as distinguishing features that can be used in machine learning. Hovy et al. [2010] studied the effectiveness of different features for machine learning and found that it was the governor and object words and the word to the left that were the most important. This finding was reflected in a later study by Litkowski [2016] that considered many different possible features for machine learning. While specific linguistic features have been a focus for many previous studies in machine learning methods for preposition disambiguation [Cannesson and Saint-Dizier, 2002, Alam, 2004, Litkowski and Hargraves, 2005, 2007, Ye and Baldwin, 2007, O'Hara and Wiebe, 2009, Baldwin et al., 2009, Hovy et al., 2010], more recently the potential of word embeddings of contextual words to serve as features has gained interest (see for example Hassani and Lee [2017] and Premjith et al. [2019]). Notably the word embeddings capture the semantics of words and thus might be of more general value than the use of the individual context words themselves.

The particular task of spatial sense detection of prepositions has been the subject of relatively few studies. Contributions to automated spatial sense detection, treated as an aspect of spatial role labelling, include Kordjamshidi et al. [2011], Hassani and Lee [2017], Kordjamshidi et al. [2017] and Manzoor and Kordjamshidi [2018]. Even less attention has been given to the task of detecting whether a preposition is being used in a specifically geo-spatial (as opposed to generic spatial) sense. As indicated above, distinguishing the geo-spatial sense of a preposition from other spatial senses is a challenge in that it depends on determining that the context of use is geo-spatial. This could be indicated in practice by reference to a named place or to a type of geographical feature or indeed by anaphoric reference (or co-reference) as in the use of 'it' to refer to a geo-spatial feature introduced in a different sentence.

In their study of geo-spatial sense detection Radke et al. [2019] reported classifier performance that was quite low with a best F1 score of 0.64 and best precision of 0.63 (from different classifiers). The machine learning methods did include features to represent the presence of geo-spatial entities in the sentence, but they were derived using a method that employed only a single gazetteer, and a dictionary of place types, and is likely to have missed many actual geo-spatial references. Here we also detect geo-spatial entities, including when emulating their approach. However we use an algorithm that exploits multiple gazetteers and employs various heuristics to avoid false positives (as many place name words also have other meanings), which resulted in superior performance. We also use a dictionary of place types, but our additional use of bag of words and word embeddings as features enables training a classifier to recognise a range of terminology associated with the geo-spatial sense, including terms that might not be present in some dictionaries of place types.

In Kordjamshidi et al. [2011], the disambiguation of spatial prepositions is considered as an aspect of spatial role labelling and as the first step in a machine learning pipeline to extract triples of a trajector, a spatial preposition (referred to as a spatial indicator) and a landmark. They disambiguate prepositions as either generic spatial (i.e. with no particular attention to geo-spatial sense) or non-spatial. A Naive Bayes classifier was used to disambiguate the preposition once it has been identified with a part of speech (POS) tagger. The features for the classifier were all linguistic being obtained with tools such as a part of speech tagger, a dependency parser and a semantic role labeller. They included the word itself and its part of speech, and words that are dependent on it and on which it depends, along with their parts of speech. They also include the dependency path to such words. This classifier achieved an F1 score of 0.88 on the TPP (The Preposition Project) dataset Litkowski and Hargraves [2005]. As part of their pipeline approach, once a preposition is classified as being spatial, then the trajector and landmark are detected using Conditional Random Fields (CRFs). They also implemented a joint learning approach using CRF to identify all three components of a triple but this did not provide superior performance in the subtask of detecting the sense of a preposition.

Several classical machine learning methods were employed in Stock et al. [2022] to identify expressions that consist of relative geo-spatial descriptions of locations (i.e. locative expressions) and distinguish them from other spatial (but not geo-spatial) and non-spatial expressions. The study differs from ours in that their goal was to document the different forms of speech that can be used in geospatial locative expressions, including the use of verbs, adverbs, adjectives, apostrophes and prepositions, and to develop a classifier to recognise these expressions irrespective of the grammatical form. Our work focuses on the particular

case of detection of geo-spatial prepositions within sentences that could contain multiple prepositions with different senses (of either geo-spatial, other spatial or non-spatial). For this more specialised task, we demonstrate that our best methods, which use deep learning, can achieve an F1 value of .94 (and precision of 0.96) which is a considerable improvement over the F1 value of .90 (precision 0.91) that was achieved in the classifier presented in Stock et al. [2022]. It may also be noted that the expressions classified in the latter work might not contain any geo-spatial preposition as the spatial relations could be communicated with other parts of speech. Also their methods employed only classical machine learning methods.

In Khan et al. [2013], the term degenerative locative expression (DLE) is introduced to refer to a spatial indicator and landmark without the trajector. They distinguish between locative DLEs in which a preposition has an explicit spatial sense and partial DLEs in which it does not though the landmark is a geographic feature. The focus of the work is on the extension of locative DLEs, that use only the prepositions "at", "in" and "on", with a trajector, and the subsequent automated extraction of triples of trajector, spatial indicator and landmark. On extraction of static (non-motion) expressions they report accuracy of 60.5%, that increased to 77.4% when assisted with manual annotation of references to places.

Dittrich et al. [2015] introduce rules to distinguish spatial prepositions from non-spatial prepositions, taking account of factors that constitute non-spatial uses, such as abstract located or reference objects that might be associated with phrases that express emotions or actions, or idiologies (e.g. "in love" or "good at singing") and the presence of collocated phrases such as "to focus on". Preliminary experiments applying their rules provided an F1 of 0.8. The rule based approach of the reported experiments is characterised by hand crafted features that relate to particular uses of prepositions in contrast to the use of more purely linguistic features as in Kordjamshidi et al. [2011], which they also introduce with a view to constructing other classifiers. In our work, as indicated earlier, the bag of words approach is intended to exploit the characteristic use of associated words with the different prepositions' senses and is found to be relatively effective, though is outperformed by the deep learning approach that employs word embeddings that are intended to encapsulate the meaning of words. A few studies have applied deep learning approaches to spatial role labelling, including generic spatial sense disambiguation. An early example is Mazalov et al. [2015] in which a multilayer perceptron (MLP) convolutional neural network was used to extract complete spatial relations. Their approach involves first detecting the spatial indicator term using a simple MLP network that has no convolutional layer. Input is the word embeddings of the words within a 7-word window surrounding the word to be classified, supplemented by embeddings of POS tags of words. No results for this stage of the process are reported though overall F1 for spatial relation extraction when the system detects the spatial indicator was 0.7 on the IAPR TC-12 dataset from SemEval-2013. A deep learning approach dedicated to generic spatial sense disambiguation is presented in Hassani and Lee [2017]. Their most effective method is a hybrid approach that combines word embeddings with a range of linguistic features in a convolutional neural network. The linguistic features include uni-grams and bi-grams and their probabilities, part of speech tags and named entity types. Their evaluation dataset was derived from the Pattern Dictionary of English Prepositions (PDEP) and they report an F1 score of 0.94 in identifying the generic spatial sense. In a joint spatial role labeling task Guo et al. [2020] applied a deep learning technique that uses a novel loss function, Inference Masked Loss, that resulted in an F1 score for spatial indicator detection of 0.95 when applied to the CLEF 2017 mSpRL dataset [Kordjamshidi et al., 2017]. Our work differs from these latter studies in addressing the task of geo-spatial, rather than generic, spatial preposition sense detection. Though as part of our study we also obtain a similar F1 score of 0.95 for generic spatial sense detection. The benefits of BERT-based transformer models for semantic role labeling in general was demonstrated in Shi and Lin [2019]. Application of such methods to spatial role labeling in a medical context of X-ray reports was presented in Datta et al. [2020]. There the task of spatial indicator detection was isolated prior to detection of other roles (of trajector, landmark, diagnosis and hedge) using explicit tagging of the spatial indicator based on the first stage. With their X-ray report dataset they obtained an F1 score of 0.91 for the spatial indicator detection task.

3 | METHODS

Here we describe the methods that we have developed to address our research questions, concerning the development of effective machine learning classifiers for detecting whether prepositions have a geo-spatial or a generic spatial sense, or neither. We investigate the effectiveness of transformer deep learning methods for this task, when compared to classical machine learning approaches. For each preposition in a sentence, as identified with a POS tagger, our objective is to determine whether it conveys a geo-spatial sense or whether it can be classed as having a generic spatial sense that includes a geo-spatial sense. We include consideration of the latter case in order to provide a comparison with previous methods that report only on the detection of the generic spatial sense of prepositions, such as Kordjamshidi et al. [2011], Hassani and Lee [2017], Guo et al. [2020].

In what follows we clarify our distinction between geo-spatial, other spatial and non-spatial preposition sense before describing our various machine learning methods. We believe that the high precision that we obtain with some of these methods is a reflection of the effectiveness of our procedure for extracting place names and geographic feature type that are used as input features for some of the methods. Those feature extraction methods are summarised in subsection 3.3. Essential to implementation of our methods is the use of two datasets that we use to train and test the classifiers. One of these was derived from source texts of the Nottingham Corpus of Geo-spatial Language Stock [2018] while the other, which is intentionally characterised by the sparse occurrence of geo-spatial senses, was derived from the Pattern Dictionary of English Prepositions (PDEP) dataset Litkowski [2014]. More detailed description of these datasets is provided in Section 4 along with details of the implementation of the machine learning experiments. The code for the paper is made available publicly [2].

3.1 | Definition of geo-spatial, other spatial and non-spatial preposition sense

Following Stock et al. [2022]'s definition of geo-spatial expressions, we define geo-spatial prepositions in terms of both the preposition and the reference object. For a preposition to be geo-spatial, it must meet two criteria. First it must describe the physical location in space of an object relative to a reference object, and second it must have a reference object that is geographic. In grammatical terms, the reference object is the object of the preposition and in geo-spatial expressions this is normally the object that is an anchor point for the location description. By geographic, we mean that the reference object can be geo-referenced (geographic coordinates could be determined for it, if sufficient information were available). This may include place names (toponyms), or specific geographic features. Geographic objects are normally found outdoors or in transitional spaces that are large and public [Kray et al., 2013], and are normally of a scale that corresponds to Montello's vista, environmental and geographic spaces [Montello, 1993].

In contrast, the reference object of a preposition with our other-spatial sense is not geographic. This situation can be regarded as equivalent to what is sometimes referred to as table-top space, especially in the context of applications in robotics [Kelleher and Costello, 2009, Tellex and Roy, 2009]. The reference objects are often objects that are movable, can be picked up or manipulated, such as cups, pens and computers, or they could be a person or a part of a person such as an arm or a hand.

Our third category of non-spatial includes all prepositions that are used non-spatially, meaning that they do not describe a physical location in space. This may include temporal, metaphorical, metonymic or figurative uses of prepositions that are otherwise used spatially (e.g. "as my friend, you should be on my side").

An example of a sentence that uses geo-spatial and non-spatial prepositions is "You can paddle and portage carry the canoe overland to the next lake for days weeks even months camping on the shores of a different lake every night pulling fresh walleye or northern pike from its crystal clear waters for dinner". It includes the following prepositions (in the order in which they appear), along with their classification according to the scheme.

- for: non-spatial - "**for** days" - sense describes time, not location
- on: geo-spatial - "camping **on** the shores..."
- of: geo-spatial - "shores **of** ...lake..."
- from: geo-spatial - "pike **from** ...waters..."
- for: non-spatial - "pike....**for** dinner" - sense describes purpose/function, not location

Here, the second and third prepositions ("on" and "of" respectively) are classed as geo-spatial because their reference objects ("shores" and "lake" respectively) are geographic in nature, being geo-referenceable and typically in vista or environmental space. The question of what counts as a geo-spatial/geo-refenceable reference object can be difficult to resolve in some cases, including for example those that refer to generic parts of the environment such as "waters" in the example above. In this case, we consider the preposition "from" geo-spatial because it clearly refers to the waters of a specific lake. However, there are cases in which such terms are used more generically, and in those cases might not be considered geo-spatial. Stock et al. [2022] provide a detailed discussion of the challenges of classifying spatial language, providing an extensive explanation of borderline examples and specific types of challenges such as descriptions of weather and hypothetical or metonymic references. We also discuss some other borderline cases when we describe the annotation process in Section 4.1.2.

[2] https://figshare.com/s/9899575b0617a6a9eaa5

The following example provides all three classes of prepositions: “Yvonne and I at South Queensferry a couple of summers ago across the road from the Hawes Inn at the slipway underneath the tall stone piers of the rail bridge the mile wide river bright before us people promenading along pavements...”:

- at: geo-spatial - “Yvonne and I **at** Queensferry”
- of: non-spatial - “a couple **of** summers”
- across: geo-spatial - “**across** the road”
- from: geo-spatial - “**from** the Hawes Inn”
- at: geo-spatial - “**at** the slipway”
- underneath: geo-spatial - “**underneath** the ...piers”
- of: geo-spatial - “piers **of** the rail bridge”
- before: other-spatial - “river bright **before** us”. Note that the preposition describes a spatial relationship, but the reference object is the observer, not regarded as a geographic object, so the preposition sense is spatial, but not-geo-spatial.
- along: geo-spatial - “people promenading **along** pavements”

Another set of examples of each type, with the preposition highlighted in italics, is mentioned below:

- geo-spatial
  1. “The Kadets led other radical deputies ***across*** the border to Vyborg in Finland where they issued a manifesto calling for protest in the form of passive resistance.”
  2. “The orthodox theory of the eighteenth-century constitution was provided by the influential jurist and lawyer Sir William Blackstone in his lectures ***at*** Oxford in 1765.”
- other-spatial
  1. “On April 18 , 1943 at 0700hrs Yamamoto climbed ***aboard*** a Mitsubishi G4M Betty bomber and set off for Bougainville , his formation was accompanied by another Betty and six Zeros.”
  2. “These sticks would be twisted round until the bag was tightly pressed and the essential oil oozed ***out of*** the petals.”
- non-spatial
  1. “Their conclusions were kept ***under*** wraps.”
  2. “This internecine strife ***within*** the Christian community was a sad diversion of effort at a time when faith was rapidly decaying.”

3.2 | Machine Learning Classifiers Overview

We present four types of classifier, being the three classical methods of Naive Bayes, SVM and Random Forests that serve as forms of baseline for our main method that employs transformer deep learning models. For the first three types of classifier we compare the use of two main types of feature. The first type of feature consists of those used in a previous study of geo-spatial sense detection Radke et al. [2019], which uses a combination of linguistic features and features that indicate whether place names or geo-feature type terms are present in the context of the preposition to be classified. The latter place names and geo-feature types are a significant aspect of input to our classifiers and our methods for extracting them are summarised in the following section. The second type of feature uses a bag of words approach in which the feature is a vector recording data on the presence of which words from the entire vocabulary are present in a window of text surrounding the preposition to be classified.

The transformer model classifiers, for which we use several BERT-based models, take as input the text in a window surrounding the preposition to be classified. That text is initially converted to word embeddings which are then progressively updated in the BERT model so that they adapt to the contexts of the training data. We also experiment with a variation on the standard input to BERT in which we tag place name words present in the input text.

Prerequisite: None
Input: An instance/sentence S from the dataset
Output:Count of the Geo-Feature types and Placenames in the instance
begin

```
1  S ⟵ findDates(S)
2  < GeoFeatureCount, S >⟵ findGeoFeatures(S)
3  < OrgNameList, PlacenameList, S >⟵ StanfordNER(S)
4  < S, tempList >⟵ StanfordPOSTagger(S)
5  foreach Entry L in temp_list do
6      if GeotextLookup(L) then
7          insert into PlacenameList
       end
8      else
           if GeonamesLookup(L) or OsmLookup(L) or
            OrdnanceLookup(L) then
9              insert into PlacenameList
           end
       end
   end
10 < GeofeatureCount > ⟵
    ExtractGeoFeaturplace/Org(GeoFeatureCount, PlacenameList, OrgnameList)
11 < GeofeatureCount, S >⟵
    Extract_GeoFeature_String_Stemmer(S)
12 return ((length of Placename-count) + GeoFeatureCount)
end
```

Algorithm 1: Count the Geographic feature Types and Placenames

3.3 | Extraction of geographical place names and geographic feature types from input sentence to use as features in classifiers

The definition of a geo-spatial sense of a preposition depends on the reference object of the preposition being required to be a geographic feature (an object in geographic space). Some of the classifiers presented in the following sections reflect this requirement by relying on the detection of the presence of features that include the number of place names and the number of geographic feature type words occurring in the context of the preposition to be disambiguated.

The geo-parsing algorithm proposed is explained in detail in this subsection. The gazetteer lookup module is capable of extracting the count of place names and geographical feature types from a sentence. In this work the OpenStreetMap, Ordnance Survey OpenNames, Geonames, and Geotext gazetteers are used. The use of multiple gazetteers is intended to optimize the performance of the geo-parsing module.

The numbers of place names and geographic feature types are extracted with the help of various natural language processing tools and gazetteers. Unlike the study in Radke et al. [2019] that used a single gazetteer, Geonames, to detect place names, we use the multiple gazetteers listed above. Our gazetteer lookup module works as a wrapper around all the gazetteers and additionally uses some natural language processing principles to extract the number of place names and geographic feature types from the input sentences. The algorithm for the gazetteer lookup module which counts the number of place names and geographic feature types is as shown in Algorithm 1.

The input to the algorithm is a sentence i.e. one instance from the dataset. The algorithm finds date entries in any format in the sentence and removes those in line 1 of the algorithm through a function call $findDates()$. In line 2, the function to count the geo-features in the sentence is called and returns the count as well as the modified sentence S with those geo-features removed from it. In line 3 it calls the Stanford NER Finkel et al. [2005] on the instance and it returns two lists, namely the list of place

Prerequisite: Placename list returned from Ordnance or OSM or Geonames or Geotext lookup and Orgname list returned from Stanford NER parser

Input: An entry of a Placename and an Orgname

Output: Extracted GeoFeatures from Placename and Orgname list

begin

1 $M \longleftarrow length\ of\ placename\ list$
2 $N \longleftarrow length\ of\ orgname\ list$
3 $i \longleftarrow 0$
4 $j \longleftarrow 0$
5 $c \longleftarrow 0$
6 **while** $i < M\ \&\&\ j < N$ **do**
7 $X \longleftarrow split\ (placename[i])\ delimiter(' ')$
8 $Y \longleftarrow split\ (orgname[j])\ delimiter(' ')$
9 $(c, GeofeatureCount, S) \longleftarrow Extract\_placename(X, S)$
10 $(GeofeatureCount, S) \longleftarrow extract\_orgname(Y, S)$
11 $i \longleftarrow i + 1$
12 $j \longleftarrow j + 1$
end
13 **while** $i < M$ **do**
14 $X\ (placename)\ delimiter(' ')$
15 $(c, GeofeatureCount, S) \longleftarrow Extract\_placename(X, S)$
16 $i \longleftarrow i + 1$
end
17 **while** $j < N$ **do**
18 $Y \longleftarrow split\ (orgname)\ delimiter(' ')$
19 $(GeofeatureCount, S) \longleftarrow Extract\_orgname(Y, S)$
20 $j \longleftarrow j + 1$
end
21 return (c,GeoFeatureCount,S)

end

Algorithm 2: Extract GeoFeature_Placename/Orgname

names and the list of organization names. It also returns the modified sentence with those entries removed from it. In line 4, a call to the Stanford part of speech Tagger Toutanova et al. [2003] is made which returns a temporary list *tempList* in which all extracted proper nouns are stored. For each entry in this list, it checks if the entry is present in any of the four gazetteers, namely Geonames [3], Ordnance Survey [4], Geotext [5], or OpenStreetMap [6] and if so inserts the entry in the placename list $PlacenameList$. This is done in lines 5-9 of the algorithm. In line 10, Algorithm 2 (ExtractGeoFeatureplace/Org) is invoked to check if any of the place names or organization names in our lists $PlacenameList$ and $OrgnameList$ are actually geo-feature types. If they are, they are removed from the placename list. The geo-feature count which was initially calculated in line 2 is updated accordingly. Some of the geo-feature types which occur in the sentence S in some other form than their lemma are now extracted in line 11, wherein it stems the remaining words in S and checks if any root form is there in the geo-feature dictionary (ADL dictionary of geo-feature types) [7]. For this it uses the SnowBall stemmer [8] to bring the words to their root forms. Finally the length of the placename list, which is the count of place names, and the count of geo-feature types is returned by the algorithm. Algorithm 2

[3] https://www.geonames.org/
[4] https://www.ordnancesurvey.co.uk
[5] https://pypi.org/project/geotext
[6] https://www.openstreetmap.org
[7] https://www.legacy.alexandria.ucsb.edu
[8] https://snowballstem.org/

Prerequisite: Instance S with the dates removed
Input: Instance from Dataset as S
The ADL dictionary of geo-feature types
Output: String S with the geo feature types extracted
begin

```
split_S ⟵ split(S)(delimiter =' ')
foreach Entry E in geofeaturetypes dictionary do
    if E in S then
        split_E ⟵ split(E)(delimiter =' ')
        if set(split_E) is subset of set(split_S) then
            if Count of E in S >0 then
                GeoFeatureCount ⟵
                 GeoFeatureCount + CountofE
                remove all instances of split_E from S
            end
        end
    end
end
return (GeoFeatureCount,S)
```

end

Algorithm 3: Find the geographic feature types

checks if any of the words or phrases in the sentence S are present in the ADL dictionary of geo-feature types and returns their count. [9] Note that this function assumes that the geo-feature type is present in the sentence in singular lowercase, i.e. root, form.

3.4 | Bayesian Learning with Linguistic features

The first baseline approach in this work makes use of the Naive Bayes classifier with the linguistic features used in Kordjamshidi et al. [2011] to which we add geo-spatial features, namely the number of geographical place names and the number of geographic feature types in the input sentence. The linguistic features are specific to each preposition to be classified and include the word to be classified, its part of speech and words (and their parts of speech) that are dependent on it and on which it depends, plus the dependency path to these words. We combine the geo-spatial features in three different ways, emulating the methods of Radke et al. [2019] for detecting geo-spatial sense, as summarised below. A full list of the linguistic features can be found in Kordjamshidi et al. [2011] and Radke et al. [2019].

The different combinations of additional features are summarised in the following three bullet points. They replicate those reported in Radke et al. [2019] and we use the same terminology of "Kord-Geo" etc (see below):

- “Kord-Geo” uses the number of place names and number of geographic feature types occurring in the input sentence (two additional features).
- In “Kord-Geo-Sum”, the sum of the counts of number of place names and number of geographic feature types in the input sentence has been used as a feature.
- In “Kord-Geo-Or” the binary value of 1 or 0 indicates whether place names or geographic feature types are present or not in the input sentence.

[9] https://www.legacy.alexandria.ucsb.edu

3.5 | Bag of words with SVM and an SVM Metaclassifier

In this approach, rather than employing hand crafted features, we use as features a Bag of Words (BoW) that creates a vector with an element for every word in the corpus (i.e. all our source sentences). For each preposition to be classified, the feature vector has a non-negative weight value for those words present in a window surrounding the preposition to be classified. This vector becomes the input feature to various machine learning models of Naive Bayes, Support Vector Machine (SVM), and Random Forest. Prior to extraction of the BoW vector, we pre-processed the data, including removing stopwords and applying a stemming algorithm. Each word in a window on either side of the preposition in each expression was represented by a *tf-idf* (term frequency - inverse document frequency) weight to reflect its relative significance in the document collection.

In addition to the basic BoW classifiers, we used a meta-classifier approach comprising two SVM classifiers where the predicted classes and class probabilities of the first classifier act as input features to the second one, which may include other features in addition to the predictions. Our additional features for the second classifier were counts of the geographic feature types and place names for each sentence containing a preposition to be classified.

3.6 | Transfer Learning with various BERT-based models

Deep learning based models involving transformers [Vaswani et al., 2017] are state-of-the-art in many Natural Language Processing (NLP) tasks. These models aim to understand the context of a token and output a vector called a Contextual Word Representation or Contextual Word Embedding for each token in the input sentence, based on adapting initial pre-trained word embeddings to the contexts of the training data. This means that these models are not only capable of differentiating homonyms but are also able to understand the contextual meaning of out of vocabulary words, i.e. words which are not in the vocabulary of the tokenizer used for training the model. This makes these models useful for a wide variety of tasks even on domains different from the original domain over which the model was trained. A widely used transformer model is Bidirectional Encoder Representation from Transformers (BERT) Devlin et al. [2019], which is pre-trained over English Wikipedia and BookCorpus using Masked Language Modeling and Next Sentence Prediction methods. BERT has more than 340 million parameters making it a computationally intensive model. The architecture of BERT consists of a set of encoding and decoding layers in which the encoder successively transforms embeddings (i.e. multidimensional vector representations) of the textual input data to a form that is then decoded to generate what becomes the predicted output of the network. All layers of the model employ a self-attention process in which the initial input embeddings of individual words are adapted by learning from the surrounding words of the actual text (or sentence) that is input. The attention process modifies the embeddings of individual words as weighted sums of the surrounding word embeddings. The result is that the initial individual word embeddings are adapted to the context of their use. This is unlike more conventional embedding methods such as GloVe Pennington et al. [2014] and Word2Vec Mikolov et al. [2013] in which in their basic form the embedding is fixed for each particular word. Retraining of these latter types of embedding is computationally a very expensive process. Each layer of the BERT encoder itself has two layers, one of which is a (multi-head) self-attention mechanism followed by a simple feed-forward neural network. The decoder layers each have these two components plus a further multi-head attention sub-layer that "attends" to the output of the stack of encoder layers, as well as that of the other self attention sub-layer, prior to input to the feed-forward network. The decoder layers are also characterised by the use of masks to prevent predictions at one position being dependent on subsequent positions. The input representation of each token is an embedding that is the sum of the embedding of the token, the embedding of the segment, which simply distinguishes whether the token belongs to the first or the second sentence (or just one sentence if only one), and the position embedding that records the position of the token in its sentence. The token embeddings are WordPiece embeddings [Wu et al., 2016] which have a relatively small (30,000) size vocabulary due to breaking up and separately representing components of compound words.

Following BERT, several alternative versions have been developed. Here we briefly summarise the ones that we have used in our experiments in addition to BERT. DistilBERT was proposed by Sanh et al. [2019] and as the name suggests is a reduced version of BERT. This model preserves about 97% of BERT's performance while reducing the number of parameters by 40% which makes it faster by 60% compared to BERT. It was found by Liu et al. [2019] that BERT was quite under-trained. Hence, they trained it for longer, increased the batch size and tested on huge corpora. They focused on changing the masking pattern on the training data dynamically as required and trained on longer sequences as opposed to predicting the next sentence, which led to the model RoBERTa.

In Lan et al. [2020], ALBERT was presented in an effort to avoid computing intensive operations and to reduce TPU/GPU dependencies. It has fewer parameters compared to BERT, is faster in training time and has lower memory consumption leading to a more scalable version of BERT without compromising on the efficiency and accuracy of the model.

Transformer-XL [Dai et al., 2019] is a modified form of transformer model which has the capability to learn long term dependencies as opposed to the base transformers which learn fixed length dependencies. When regression of a variable against itself is done, it is termed as autoregression. Here the predicted future values of the variable are based on the past values of the input and the current input. The BERT model is not an autoregressive model but it is an autoencoder model. An autoregressive model can see the context in backward or forward direction while the autoencoder model of BERT can see the context in both directions. However, BERT, though an autoencoder model, is known to have pre-train/fine-tune discrepancy due to the use of masking. This is due to the fact that masking should not be done in fine-tuning. BERT achieves better performance than approaches based on autoregressive methods when it comes to pre-training.

XLNet [Yang et al., 2019] is a BERT-like model that incorporates the ideas from Transformer-XL into the pretraining procedure. It has been shown to outperform BERT significantly in many NLP tasks including natural language inference (NLI), question/answering (QA) and sentiment analysis. The main idea of XLNet is that it provides a new method by which an autoregressive model learns from the bi-directional context to avoid the disadvantages brought by the masking method in the autoencoder language model.

Lim and Madabushi [2020] introduce new features by calculating the TF-IDF vector for each sentence and concatenating it to the corresponding BERT output. This vector is then fed to a fully connected classification layer. Usually, to handle the concatenated features, an additional dense layer is added to the network. A drawback of this method is that the feature vector becomes too long and the parameters of the model increase which in turn leads to a rise in the training time of the pre-trained model.

In our use of BERT-based models we experiment here with incorporating an additional feature to indicate that a word in the input text is a named geographic feature. As indicated above, BERT-based models take, as input, text the tokens of which are converted initially to pre-trained embeddings. We incorporate the knowledge that a word is a named geo-feature by prepending and appending a location tag before and after any place name that has been detected in the input sentence. Consider the input sentence - "John lives in New York". The sentence is first parsed using an NER (Named Entity Recognition) Tagger. The tagger returns occurrences of place names in the sentence, here "New York". A location tag is prepended and appended at the start and end of the place name, resulting in this case in "John lives in <LOCATION> New York </LOCATION>." Note that tags handle multi-word place names as in this example. These new tags are added to the vocabulary of the tokenizer so that the information about whether a place name exists in the sentence or not is conveyed to the pre-trained model. The motivation behind this process is that prepositions with a geo-spatial sense are frequently, but certainly not invariably, associated with occurrences of place names. This architecture is illustrated in Figure 1.

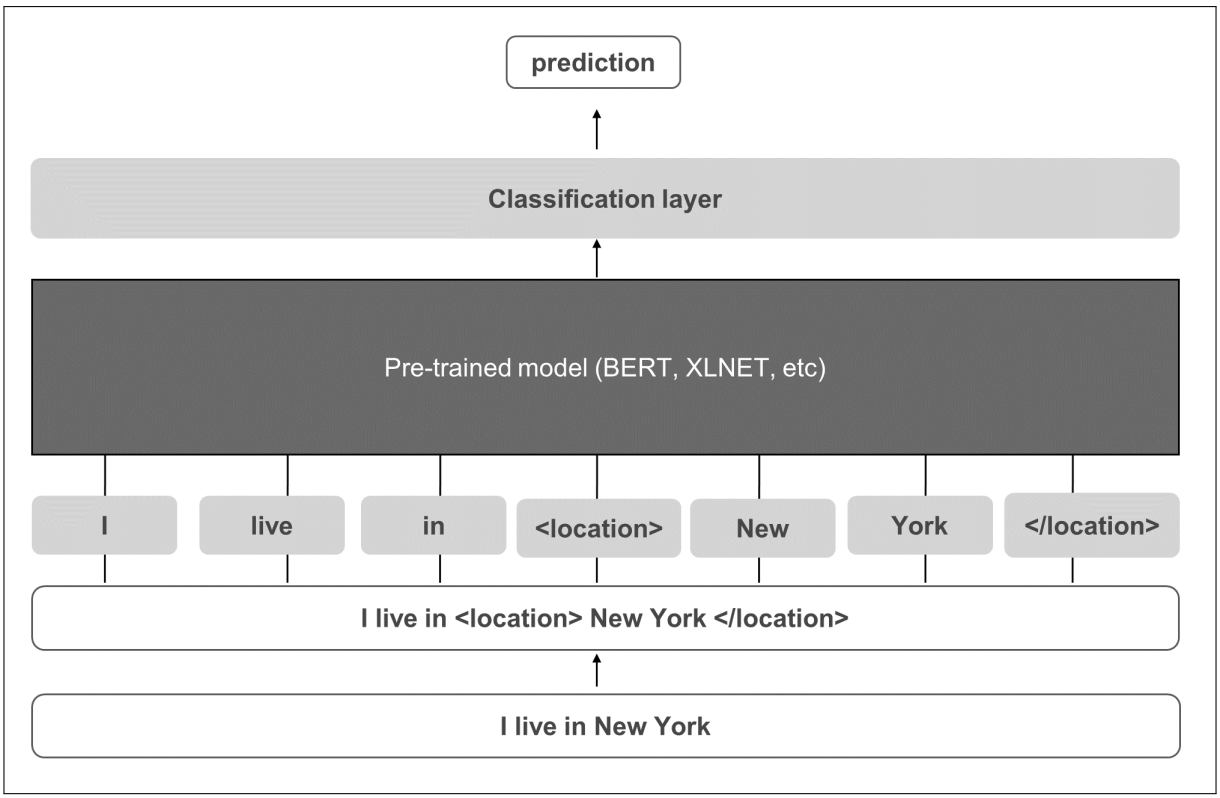
The specific use of the above BERT-based models for our classification task is described in detail in section 4.2

4 | EVALUATION

In this section we describe details of the experiments that we conducted including how we applied the various methods described in Section 3. For all experiments the aim is to classify the sense of a selected preposition as either geo-spatial, generic spatial or non-spatial. This was done in practice by conducting individual binary experiments in which the target class was either geo-spatial, other spatial or non-spatial, before then merging results to generate precision, recall and F1, in the first instance for geo-spatial *vs* other spatial or non-spatial and in the second (generic spatial) case for geo-spatial or other-spatial *vs* non-spatial.

For the method "Bayesian Learning with Linguistic features" we generated the same features used in Radke et al. [2019] (which themselves include the features from Kordjamshidi et al. [2011]) for each preposition to be classified. Thus these features are all generated relative to that target preposition including for example the head words of the preposition, and their parts of speech and lemmas, and the dependency paths to the head words. For the Bag of Words methods, the input data for each preposition to be classified was the words within a window surrounding the preposition. It is these words that are used to generate the bag of words vectors. In the case of the BERT-based methods the input text for training and testing the methods was again the text within a window surrounding the preposition to be classified. Note that an individual sentence can have more than one preposition and hence the context words in these windows is different in each case.

FIGURE 1 Architecture of BERT based models used



4.1 | Data Set Preparation and Annotation

We use two datasets in this work, both of which have been made available publicly on the Figshare repository[10,11]. The process of annotation that we adopted for these datsets followed the guidelines set out in Artstein [2017] and Carletta [1996], in which the use of more than one annotator with verified agreement ensures consistent and reproducible annotation.

4.1.1 | Dataset 1 (Derived from the source of the Nottingham Corpus of Geo-spatial Language)

Dataset 1 was created in the following manner. 6221 unique sentences were randomly selected from the original source data of the Nottingham Corpus of Geo-spatial Language (NCGL). The data was first harvested using an automated method described in Stock et al. [2013], and then manually filtered to remove expressions that were not geo-spatial. We used the data harvested after the first step, but not yet filtered, in order to create a corpus with examples of all three classes, but with a higher proportion of geo-spatial and other-spatial sentences than would appear in randomly selected text.

Each sentence contains one or more prepositions. For example, if an input sentence under consideration s_1 has 2 prepositions, say p_1 and p_2, we created 2 instances of the same sentence $<s_1, p_1>$ and $<s_1, p_2>$. In this manner we derived 18828 instances from the total 6221 unique sentences in the dataset. Each instance $<s_n, p_n>$ was then manually annotated by the paper authors with one of the three values described in Section 3.1 according to the use of the preposition p_n: non-spatial, other-spatial or geo-spatial.

The annotation process began with a training phase, in which two authors of the paper annotated a sample of expressions. The annotators were provided with written guidelines/definitions of the classes before starting, and worked independently of each other. The guidelines were based on the descriptions contained in Section 3.1 and were an adapted version of those included in Appendix A and B of Stock et al. [2022] to address only prepositions (since that paper provides examples of spatial relations with a range of parts of speech), including both the definitions of each class and multiple examples. Then another author,

[10] https://figshare.com/s/5ff1f127948145681af5
[11] https://figshare.com/s/7407f37544f910fdbb9e

TABLE 1 Example of PDEP Sense Metadata

Preposition	Sense	Definition
near	1(1)	at or to a short distance away from (a place)
near	2(2)	a short period of time from
near	3(3)	close to (a state); verging on
near	4(3a)	a small amount below (another amount)
near	5(4)	similar to

who was more experienced in geographical data annotation, reviewed a sample of 1000 annotations (500 from each annotator) and provided feedback, identifying incorrectly annotated instances and pointing out any patterns in the errors made by annotators. Following this training phase, the first two annotators annotated approximately half of the prepositions each independently. A sample of 500 was then randomly selected from each group to be annotated by the other annotator for cross-checking purposes, resulting in 1000 expressions that had been annotated by both annotators. The inter-annotator agreement (Cohen's kappa) was calculated as 0.818 for these 1000 prepositions (indicating strong agreement [McHugh, 2012]). This method follows the guidelines set out in [Artstein, 2017] to ensure a correct and reliable process.

The final annotated data set consisted of 5557 geo-spatial prepositions; 3027 other-spatial prepositions (and hence 8584 generic spatial prepositions), and 10,238 non-spatial prepositions, thus having a reasonable representation of each class.

4.1.2 | Dataset 2 (Derived from the PDEP corpus)

The second dataset is derived from the PDEP corpus[12] and is intended to test the ability of our method to successfully classify sparse data, in which there are few geo-spatial sentences. Given that in non-specialised text of the kind harvested from many web sites or other text documents, geo-spatial sentences are relatively infrequent, this can be regarded as a realistic scenario. Dataset 2 contains 45279 instances ($<s_n, p_n>$ pairs), 2122 of which are geo-spatial, making the ratio of geo-spatial instances to the other instances intentionally skewed. We created this dataset using a two stage process: in Stage 1, we identified instances with prepositions that had senses that were likely to be spatial (whether geo-spatial or other-spatial), and in Stage 2 we manually annotated these with our three target classes, as described below.

Stage One

Dataset 2 was created by first filtering the PDEP (Pattern Dictionary of English Prepositions) corpus, by removing rows containing incomplete information or empty or null columns. Then, one of the authors annotated the set of senses provided with the PDEP corpus in order to identify prepositions used in a spatial sense (which could encompass both our geo-spatial and other-spatial classes). The PDEP corpus metadata lists prepositions, their corresponding senses and definitions for each. An example is shown in Table 1 for the preposition *near*. As can be seen, only the sense 1(1) listed here is clearly spatial in nature. The others refer to time, similarity, quantity and state.

Again following recommended practice for sound creation of annotated data sets [Artstein, 2017], a second author annotated a sample of 100 senses to verify the process, achieving a Cohen's Kappa of 0.76 (see Figure 3), considered to indicate moderate agreement [McHugh, 2012]. Many of the discrepancies between the two annotators were to do with the more liberal interpretation of senses, with the second annotator marking as spatial any sense that might possibly include a spatial use, while the first only marked as spatial those that were exclusively for spatial uses. This stricter interpretation was thought to be more appropriate given the use of this process to identify potential geo-spatial and other-spatial expressions ahead of manual filtering in the following step (see below).

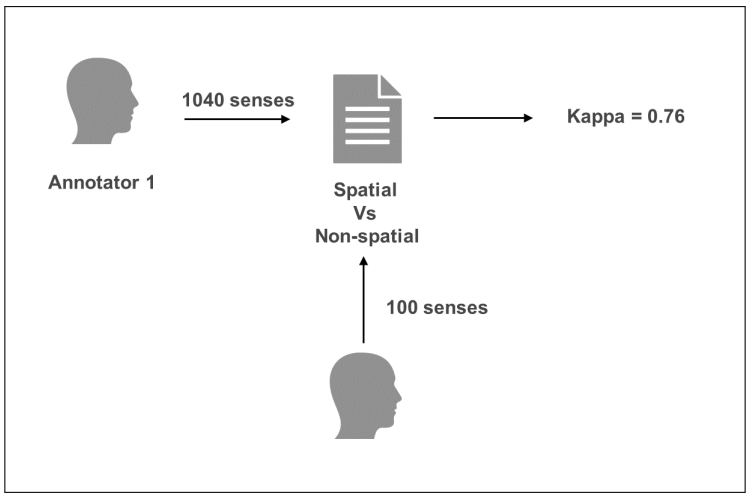
The PDEP corpus marks each preposition with a sense, and we next extracted $<s_n, p_n>$ tuples from the PDEP corpus that were annotated with one of the senses that the first (stricter) annotator marked as spatial (encompassing both geo-spatial and other-spatial classes in our classification scheme). From the resulting set of 34834 instances, we randomly selected 6000.

Stage Two

While the preposition senses in the PDEP dataset suggest that the 6000 randomly selected instances with spatial senses from Stage 1 of the process are likely to be spatial, for our sense-detection method, we need to distinguish geo-spatial and other-spatial

[12]https://www.clres.com/pdep.html

FIGURE 2 Stage 1 of the Annotation Procedure for Dataset 2



senses for training and testing purposes, as well as verifying that the expressions that are assumed to be spatial based on their selected PDEP senses, are indeed spatial. Thus in the second stage of annotation, the sample of 6000 expressions extracted in the previous stage was manually classified using our ternary classification scheme (geo-spatial, other-spatial and neither). The 6000 expressions were divided among three of the paper authors. A sub-sample of 200 was randomly selected and annotated by all three authors, achieving an average Cohen's kappa value of 0.73 (average calculated between three binary comparisons), and the remainder were annotated by one of the three authors (see Figure 3). While the Cohen's kappa value of 0.73 is lower than is ideal, the average accuracy of the binary pairwise comparisons is 0.85. We note greater discrepancies among annotators for geo-spatial/other-spatial annotations, while disagreement about which expressions are non-spatial is rare. This is due to the challenges involved in annotation of geo-spatial language more generally [Aflaki et al., 2018, Wallgrün et al., 2014], with some examples provided below.

The final data set was then constructed as follows:

- 2122 geo-spatial instances resulting from the second stage of manual annotation.
- 2994 other-spatial instances resulting from the second stage of manual annotation.
- 40163 non-spatial instances, being 904 instances that were annotated as non-spatial in Stage 2 of the annotation process plus 39259 instances from the PDEP data set that were annotated with senses that we had classified as non-spatial in Stage 1 of the annotation.

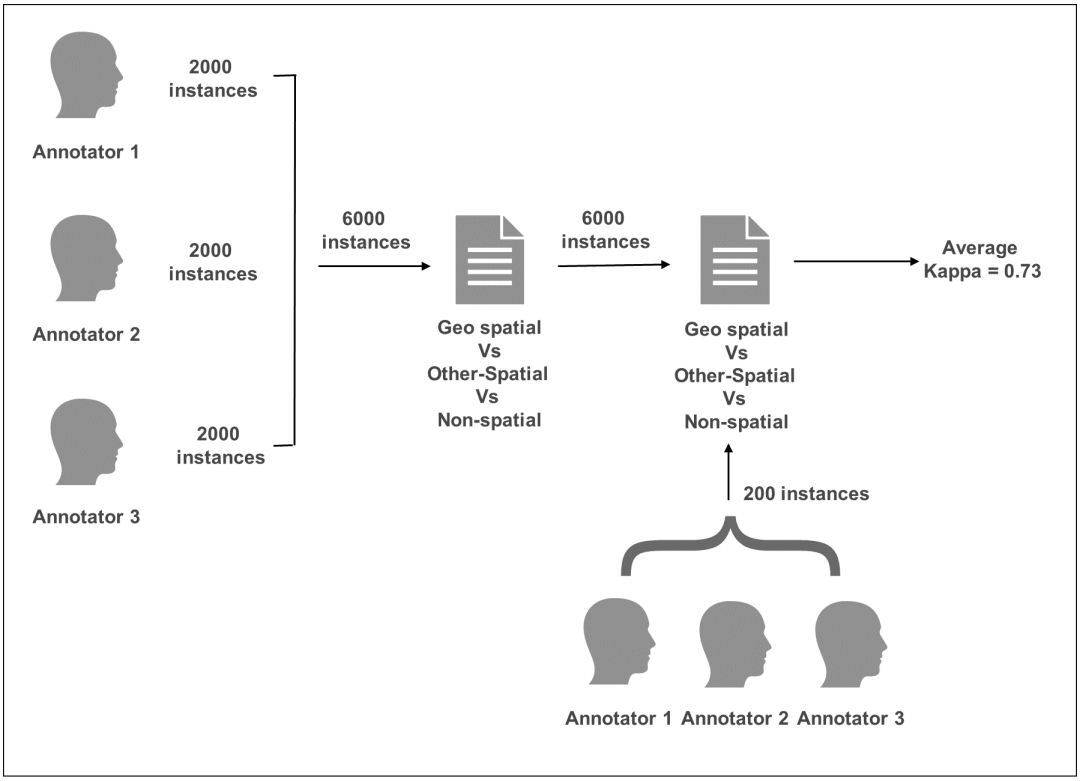
As explained above, Dataset 2 is already sparse, but to test the model further, we considered an even more skewed ratio between the classes, and hence an even sparser dataset, by taking 1000 geo-spatial, 1500 other-spatial and 40163 non-spatial instances for (geo spatial vs others) and (generic spatial vs others). We call this dataset the "Sparser version of Dataset 2".

The task of annotating geo-spatial, other-spatial and non-spatial sentences is challenging for humans, due to the multiple nuances; and thus of course for the classifier too. Some examples of sentences which were difficult to annotate from Dataset 1 are given below:

1. *"According to this article the first direct passenger plane between the Republic **of** China Taiwan and the Peoples Republic of China happened early this morning."*
   Disagreement about whether China here is meant as a country (and thus geospatial), or a political entity (metonymic use of a place to represent a government).

2. *"**After** 50m you will reach a road with wide verges where you turn left toward Lambley."*
   Disagreement about whether after is purely temporal, or also spatial since it describes the process of travelling along

FIGURE 3 Stage 2 of the Annotation Procedure for Dataset 2



for 50m and then reaching a road, and thus indicates the spatial configuration of different locations (as well as having a temporal aspect).

3. *"'It has seemed like that', he said sombrely, looking **off** into the distance."*
   The reference object (distance) may be considered abstract, rather than physical, and thus not georeferenceable.

4. *"FIRE broke out yesterday **on** a cross-Channel ferry sailing from Dieppe to Newhaven."*
   The reference object is mobile, and thus may be considered not geo-referenceable. However, descriptions of transport routes are numerous and could be mapped.

5. *"Nobody is seriously being invited to stand **on top of** the mountain of verbiage to get an overview."*
   The preposition ("on top of") is immediately followed by a geographic feature type ("mountain"), but refers to a mountain of verbiage, and is thus metaphoric.

6. *"Later, these functions will move **outboard of** the servers and will be implemented as either an in-band or out-of-band function in the storage network itself."*
   The preposition "outboard of" describes the location of software functions on a server/network, and there is a question about whether this refers to physical, or purely digital space.

4.2 | Experiments and Results

In this work, experimental results for two main preposition sense classification problems of geo-spatial and generic spatial are presented based on combining individual results for each of the three classes used in annotation of the data. These three classes are geo-spatial (gs), other-spatial (sp), and non-spatial (nsp). In the first experiment the target class is geo-spatial (gs) and the other two classes (sp and nsp) are merged as one alternative class. Note that every geo-spatial sentence is also spatial, but the classes geo-spatial, other-spatial and non-spatial are mutually disjoint. The second experiment detects the generic spatial class in which the results for the target class are computed by merging geo-spatial (gs) and other-spatial (sp) results. The target class

TABLE 2 Naive Bayes using features in Kordjamshidi et al. [2011]+ additional features, with Dataset 1

	Geo-spatial i.e. (gs) vs (nsp+sp)			Generic Spatial i.e. (gsp+sp) vs (nsp)		
Features used	**P**	**R**	**F1**	**P**	**R**	**F1**
Kord-Geo	0.746	**0.902**	0.816	**0.913**	0.611	0.732
Kord-Geo-Sum	0.792	0.882	**0.834**	0.852	0.730	0.786
Kord-Geo-Or	**0.927**	0.744	0.826	0.800	**0.797**	**0.798**

generic spatial of this second experiment corresponds to the spatial class of for example Kordjamshidi et al. [2011] and Hassani and Lee [2017].

In Table 2 the results of experiments performed using a Naive Bayes classifier with Dataset 1 are presented. In all three experiments, the linguistic features from Kordjamshidi et al. [2011] in combination with some additional features allow us to compare our results directly with those of Radke et al. [2019] who present results for the task of geo-spatial sense detection with prepositions. The linguistic features were summarised earlier in the part of the Related Work section that describes Kordjamshidi et al. [2011]. Extraction of these features included the use of the Stanford POS tagger and dependency parser and the LTH semantic role labelling tool [Johansson and Nugues, 2007] at barbar.cs.lth.se:8081.

Regarding extraction of the features for the three variants of the Kordjamshidi et al. [2011] method, i.e. "Kord-Geo", "Kord-Geo-Sum" and "Kord-Geo-Or", presented in Radke et al. [2019], and summarised here in Section 3.4, we employed the methods described in Section 3.3.

For these three variants, all of which used 10-fold cross validation for training and test, the Kord-Geo-Or experiment (with a feature indicating the presence of either place names or geographic feature types) provides the best results, with a geo-spatial F1 value of 0.826 and generic spatial F1 of 0.798. The precision for the geo-spatial class is relatively strong at 0.927, but with poorer recall, which is highest for geo-spatial with the Kord-Geo features. These results are significantly better than those reported by Radke et al. [2019] despite using a similar test corpus. We believe that our improved results are a consequence of providing more effective detection of place names and place type words, as summarised in Section 3.3 in which we employ multiple gazetteers in combination with several heuristics.

To compute the bag of words (BoW) features, the words in a window extending 5 words to the left and to the right of the target preposition were used. For the machine learning techniques of SVM and Random Forests, the hyperparameters were tuned using grid search Pedregosa et al. [2011]. In grid search, we specify a list of values for each hyperparameter which needs to be tuned. Grid search tries out every combination of the hyperparameter values and determines the combination which best fits the data. The resulting optimal values were used to perform the training and testing process. For SVM, we used the RBF (Radial Basis Function) Kernel. The value of C was set to 10. Gamma took values from the set 0.1, 1, 10 and varied according to the experiment at hand. The test results for the basic BoW experiments were obtained with 10-fold cross validation. For the meta-classifiers, following the 10-fold cross validation on the first SVM classifier, the probabilities for each fold were calculated. The obtained average probabilities along with the counts of the geographic feature types and place names for each instance become the input features to the second classifier. Ten-fold cross validation was then performed on the second classifier.

In Table 3 we report the results for the BoW geo-spatial experiment with various classifiers, again with Dataset 1: SVM, Naive Bayes and Random Forest (RF) using the Bag of Words features as described in Section 3.5. Of these experiments the Random Forest classifier provides better geo-spatial precision, at 0.945, than the previous linguistic features approaches but this is accompanied by lower recall. The difference in precision for BoW with the three types of classifier is quite marked and indicates that the decision tree approach of the Random Forest classifier may be particularly well suited to detecting the contextual geo-spatial words that are characteristic of the geo-spatial prepositions. The F1 value for geo-spatial sense detection with SVM bag of words is higher than for any of the methods using linguistic and additional geographically-based features, but for precision it was not able to match the Random Forest classifier, or the Naive Bayes Kord-Geo-Or method.

Table 4 reports the results of two meta-classifiers with Dataset 1. The first row in the table refers to an SVM classifier in which the features are the output probabilities from the SVM Bag of Words classifier (the probabilities of a preposition belonging to the relevant class) combined with features representing the counts of place names and geographical feature types in the context of the preposition to be classified, while the second row refers to a similar classifier that uses the output predictions rather than probabilities from the BoW classifier. While the F1 value for the geo-spatial experiment for the second meta-classifier is slightly higher than the best Bag of Words result (SVM), the improvement is very modest, and the precision does not show any

TABLE 3 Results of geo-spatial classification i.e. gs vs (sp+nsp) with Bag of Words, with Dataset 1

Method used	Precision	Recall	F1
Naive Bayes with BOW	0.773	**0.882**	0.824
SVM with BOW	0.874	0.845	**0.864**
RF with BOW	**0.945**	0.649	0.770

TABLE 4 Metaclassifier Results using Bag Of Words (SVM+SVM), with Dataset 1

	Geo-spatial i.e. gs vs (nsp+sp)			Generic Spatial i.e. (gsp+sp) vs nsp		
Features Used	**P**	**R**	**F1**	**P**	**R**	**F1**
o/p probabilities + Geo	**0.946**	0.770	0.849	0.782	**0.831**	0.805
o/p predictions + Geo	0.901	**0.851**	**0.875**	**0.850**	0.768	**0.807**

significant improvement. However the overall performance, when considering both precision and recall, is better for the first meta-classifier than for the RF BoW classifier (which had the highest precision for the geo-spatial sense). The F1 values for the generic spatial experiment are only very slightly improved over the Kord-Geo-Or (the best of the linguistic feature methods), and the precision is lower than that of Kord-Geo.

A notable outcome specifically for the geo-spatial preposition classification task, when comparing the BoW methods with those using linguistic features methods, is that despite the relative simplicity of the BoW approach (being based just on *tf-idf* values of individual words) it is able to match and sometimes outperform the methods based on quite complex hand-crafted linguistic features. The experiments that follow, using the BERT-based methods, in turn are able to significantly outperform the BoW methods, illustrating the power of representing the semantics of words as multi-dimensional word embeddings in a deep learning environment.

The final method which is presented involves the use of various BERT-based models to determine the geo-spatial or generic spatial sense of prepositions (see Section 3.6). Here the BertForSequenceClassification model is employed with uncased text in the Hugging Face[13] implementation. This version of BERT adds a final classifier layer that is used as part of the process of fine tuning to output the class labels. For the BERT experiments, the pre-trained 'BERT tokenizer' (which is based on the WordPiece tokenizer) is used to tokenize the sentences and is provided by the transformer library. The tokenizer maps each word of the input text 'sentence' to a numeric value and creates a numeric vector for each sentence. It uses a variety of techniques including byte pair encoding (BPE) to handle words such as "can't". To mark the start and end of each instance, the [CLS] (classifier) and the [SEP] (separator) tokens are added to the start and the end of each sentence. In our experiments the input text consisted of all words in a window that extended 5 words to the left and to the right of the preposition to be classified. The sentences are padded using a padding (PAD) token to ensure that all of them are of the same length. The value of length is a maximum value equal to the number of tokens in the longest instance. The entire dataset is divided into train (80%), test (10%) and validation (10%) datasets. The validation data is used to set the hyperparameters. The recommended batch size of 32 is used to train the model and sample batches randomly from the dataset for the training process. The batches are randomly selected for training purpose so that the model trains in an unbiased manner. The number of output labels is set to 2 as both experiments have binary class labels. The Adam Optimizer was employed with a learning scheduler for the training process, specified by a warm up period followed by a gradual increase in the learning rate. The learning rates for each BERT model were set as follows: XLNet: 4.00E-04; Transformers XL: 4.00E-04; Roberta: 2.00E-05; BERT: 2.00E-05; ALBERT: 1.76E-03; DistilBERT: 2.00E-05. The models were trained for 4 epochs, and for each epoch the training *vs* validation loss was calculated to keep a check on overfitting. The F1 scores were recorded for the test set. The whole process was repeated 10 times and each time the training, testing and validation datasets were randomly sampled from the entire dataset. Averaged values were computed for Precision, Recall and F1 score.

The analysis of the GPU training times for the various BERT-Based models is shown in Figure 4. The blue line indicates the GPU training times for the geo-spatial versus others (gsp *vs* (sp+nsp)) experiment. The red line indicates the GPU train times for the generic spatial versus non-spatial experiment ((gsp+sp) *vs* nsp).

[13] https://huggingface.co

FIGURE 4 Analysis of the GPU training time for the experiments

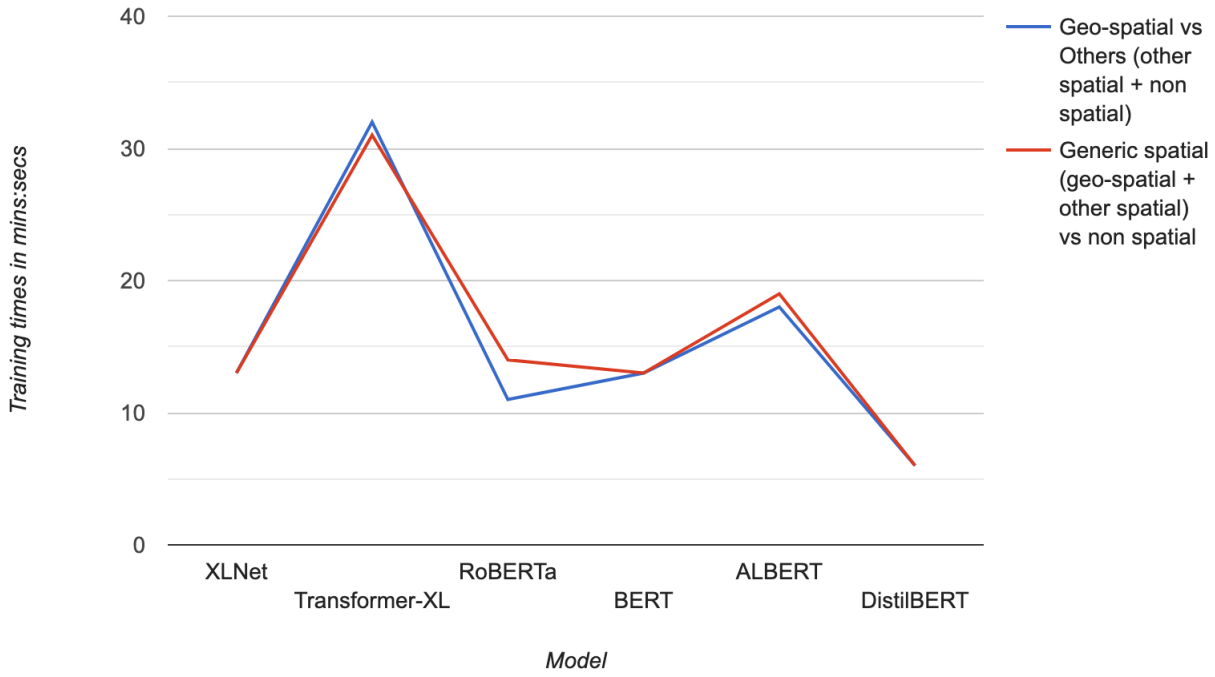
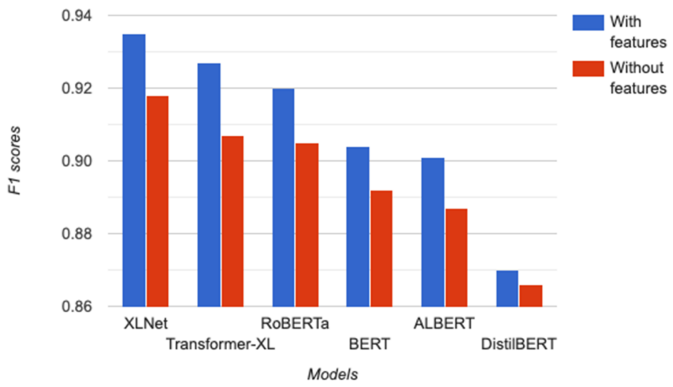


FIGURE 5 Impact of the proposed modification to the BERT architecture on F1 scores for the Geo-spatial versus (other spatial + non-spatial) experiment for Dataset 1



The results of modifying the input to the transformer models to include tags indicating the presence of place names, as explained in Section 3, are reported in Figures 5 and 6 which relate to Dataset 1. The approach was found to provide significant improvement in performance of all six transformer models, as reflected in the precision and F1 scores. Therefore all subsequent BERT-based model experiments on Dataset 2 were conducted by incorporating this feature. All results in Table 5, for both datasets relate to the use of the place name (location) tags.

As illustrated in Table 5, the F1 results for all BERT-based models when applied to Dataset 1 exceed those for the other (non-BERT-based) experiments for both geo-spatial and generic spatial sense detection. This was the case for the best BERT-based methods even when the location tags were omitted. While the XLNet model provides the best overall performance, considering both senses, it may be noted that precision of the geo-spatial sense prediction obtained with the Bag of Words metaclassifier matched the XLNet performance (being 0.946 vs 0.943), though its recall was poorer than all of the BERT-based models. We observed that BERT is slightly better at predicting the generic spatial class than the geo-spatial class, in contrast to the methods that combine linguistic and additional geographic features (Kord-Geo, Kord-Geo-Sum and Kord-Geo-Or) and the

FIGURE 6 Impact of the proposed modification to the BERT architecture on F1 scores for the Generic spatial (geo-spatial + other spatial) versus non-spatial experiment for Dataset 1

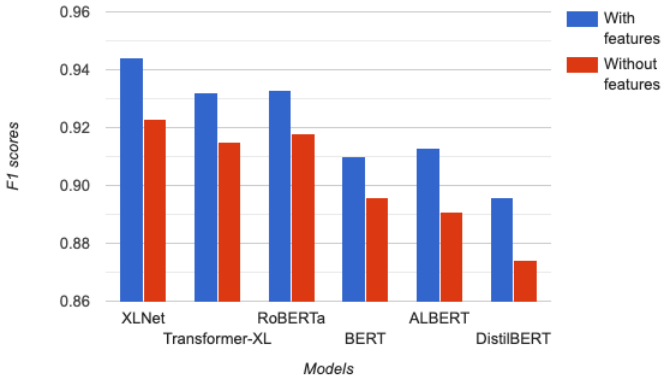


TABLE 5 Experiment with Transfer Learning on Dataset 1 and Dataset 2

		Geo-spatial i.e. (nsp+sp) vs (gs)			Generic Spatial i.e. (nsp) vs (gsp+sp)		
Dataset	**Algorithm**	**P**	**R**	**F1**	**P**	**R**	**F1**
Dataset 1	XLNet	**0.943**	0.928	**0.935**	**0.953**	0.937	**0.944**
	Transformer-XL	0.917	0.937	0.927	0.915	**0.948**	0.932
	RoBERTa	0.902	**0.939**	0.92	0.954	0.909	0.933
	BERT	0.915	0.895	0.904	0.902	0.917	0.91
	ALBERT	0.936	0.867	0.901	0.931	0.896	0.913
	DistilBERT	0.871	0.869	0.87	0.893	0.899	0.896
Dataset 2	XLNet	**0.958**	0.917	**0.938**	0.941	0.952	**0.947**
	Transformer-XL	0.931	0.932	0.931	0.925	**0.957**	0.941
	RoBERTa	0.912	**0.942**	0.926	**0.948**	0.929	0.939
	BERT	0.925	0.898	0.911	0.924	0.918	0.92
	ALBERT	0.927	0.899	0.913	0.922	0.913	0.918
	DistilBERT	0.864	0.901	0.882	0.898	0.897	0.898

meta-classifier approaches. However the absolute performance on generic spatial sense prediction of the BERT-based models significantly outperforms the other approaches. The improvement of geo-spatial over generic spatial prediction in the linguistic and meta-classifier experiments may be due to the incorporation of the features that indicate the presence or number of place names and geographic feature types, which are key to the definition of the geo-spatial class (see Section 3.1). It may be noted that the superiority in geo-spatial *vs* generic sense detection was not found in the study of Radke et al. [2019] and we attribute the difference here to our more sophisticated methods for detecting place names that employ multiple gazetteers.

Table 5 illustrates the results of applying our methods to Dataset 2, which in addition to coming from a different data source that was designed for preposition sense detection, is characterised by a much lower proportion of geo-spatial and other spatial prepositions when compared with Dataset 1. The precision of XLNet for this sparse dataset was actually superior at 0.96 to that obtained for Dataset 1, while F1 was very similar, both being 0.94 when rounded to 2 significant figures. Generic sense detection performance with the best BERT-based models was very similar to that for Dataset 1.

When the best model XLNet was tested on an even sparser version of dataset 2, illustrated in Table 6, the classifier's performance degraded somewhat, but still obtained a precision of .93 for geo-spatial sense detection, with an F1 of 0.88. Precision and F1 for generic spatial sense detection were both 0.90. This lower performance is attributed to the reduction in both cases in the quantity of training data for the target class.

Experiment	Model	P	R	F1
Geo-spatial : gsp vs (sp+nsp)	XLNet	**0.926**	0.844	0.883
Generic spatial: (sp + gsp) vs nsp	XLNet	0.898	**0.907**	**0.902**

TABLE 6 Experimental results on sparser version of Dataset 2

TABLE 7 Examples of correct prediction of geo-spatial prepositions by the XLNet classifier.

ID	**Sentence** (**preposition in bold**; *context window in italics*)
1	*Above us, waterfalls tumbled* **down** *the mountainside from glaciers that* hung over the lip of high cols.
2	Critics complained of the 'drab uniformity' of such seemingly *endless streets, which frequently concealed* **behind** *them a network of older* and certainly less well ordered thoroughfares.
3	John *Elsley spoke for many booksellers* **up and down** *the country: 'Unemployment and the* fear of unemployment have been a major adverse factor, during the year and particularly in the period up to Christmas.
4	The annual report by the UK government's Radioactive Waste Management Advisory Committee suggested that water could *seep from deep volcanic rocks* **beneath** *the site to sandstone nearer* the surface , introducing a continuous rivulet of radioactive waste into the environment.
5	Although hierarchy *was not expressed by ritual* **along** *the coastal strip, inequality remained* fundamental to perceptions of caste.
6	*A 1500ft volcano* **in** *the Timanfaya National Park.*

TABLE 8 Examples of incorrect prediction of geo-spatial prepositions by the XLNet classifier. The predictions (Pred.) were either other-spatial (sp) or non-spatial (nsp)

ID	**Sentence** (**preposition in bold**; *context window in italics*)	Pred.
7	*They edged* **round** *the bend.*	nsp
8	Peter's dream of unclean animals is recorded in Acts, Chapter 10 : 9 On the morrow, as they went on *their journey, and drew nigh* **unto** *the city, Peter went up* upon the housetop to pray for about the sixth hour.	sp
9	Liberal opinion began to waver as the extent of the violence was revealed on newsreels, *on television, in Paris Match,* **across** *the globe*.	sp
10	**Alongside** *the rugose corals low cushions* or branching masses of a different kind of calcite coral are often found.	nsp

In summary the best F1 performance for both geo-spatial and generic sense detection was obtained with the XLNet transformer model. This outcome reflects results in other domains in which BERT-based transfer learning approaches often outperform methods that use hand-crafted features as well as some alternative deep learning approaches. In our case we can interpret this as indicating that the contextual embeddings that the model learns in training can be more powerful in communicating significant linguistic features than the explicit linguistic features such as part of speech tags and dependency paths, or the presence of individual words as in the bag of words methods. This benefit was however only clearly expressed with regard to improving recall values. Thus the bag of words classifier, in which features are *tf-idf* values of individual words, provided the best precision, though this was almost the same (differing only in the third decimal place) as that obtained with XLNet. An important outcome of the present study is however the fact that explicit tagging of place names in the BERT-based input resulted in a significant boost in performance of up to 2%.

Study of results of XLNet classification on individual instances casts light on the effectiveness and limitations of the BERT-based approach. In Table 7 we illustrate examples of successful prediction of the geo-spatial sense. Note that the words in italics are those in the context window that extends 5 words to the left and right of the preposition to be classified, and which is input to the classifier. With one exception, we intentionally focus on the more challenging cases in which there is no place name in the sentence. For example in sentence 1, in which the preposition *down* is correctly classified, the classifier has successfully

identified the geo-spatial semantics of terms such as *mountainside* and *glaciers*. Other examples of terms that we might expect to be encoded with geo-spatial semantics within the classifier's adaptive word embeddings include *streets* in sentence 2, *country* in sentence 3, *rocks*, *site* and *sandstone* in the gelogical context of sentence 4, and *coastal strip* in sentence 5. Sentence 2 is of particular note in that the preposition *behind* has as its object (landmark) the word *them* which is an indirect (anaphoric) reference to *streets* that occurs elsewhere in the context window. The sixth example is a more straightforward case in which the sentence contains a place name (*Timanfaya National Park*). Sentence 3 is of note for illustrating the fact that our approach includes processing multi-word prepositions as in *up and down*. Table 8 illustrates examples of some failed cases. The examples are characterised by the use of relatively rare prepositions such as *unto* in sentence 8, and infrequently used terms that might not have have been learnt as having geo-spatial semantics due to inadequate examples in training, such as *bend* in sentence 1, *globe* in sentence 9 and *rugose coral* in sentence 10. It may also be remarked that in sentence 7 the short text has resulted in a smaller than normal context window and hence less evidence for the classifier.

While all of the BERT-based methods perform well, the results for DistilBERT, and to a lesser extent ALBERT and BERT are lower than for XLNet, Transformer-XL and RoBERTa. This is unsurprising given that DistilBERT and ALBERT intentionally sacrifice performance in order to reduce processing load (see Section 3.6), confirmed in particular for DistilBERT by our training time comparisons (Figure 4); and XLNet, Transformer-XL and RoBERTa were all specifically designed as improvements over BERT. To further examine the differences in classification between the BERT-based methods, we examined cases of prepositions that were classified differently by different methods. Given the high precision for all BERT-based methods, we specifically identified challenging cases for comparison, including prepositions that were multi-word, slang (e.g. 'thru'), obselete (e.g. 'abaft'), unusual (e.g. 'modulo'), vague or frequently used for senses other than the one used in our data. One of the most common types of instances that were misclassified by ALBERT and DistilBERT, but not by the other methods, included those that used prepositions metaphorically, but often in the company of place names or geographic feature types (e.g. "the realities of France **in** the throes of revolution", "was achieved **through** citizenship of the United Kingdom"). Metonymic uses of place names to represent governments or organisations also proved challenging for the reduced forms of BERT (e.g. "Macdonalds all **over** the world contribute"), especially if other geographic feature types or place names were present. There were also cases in which geospatial prepositions were misclassified as spatial and vice versa by DistilBERT and ALBERT, with reference objects such as bodies "bodies are dropping all **over** the place" and indoor rather than outdoor items (bedroom, furniture). We can see by these examples that while all of the BERT-based methods perform well, the full, and particularly the more recent/extended versions for the original BERT are better able to identify the nuances in the particularly challenging expressions that include misleading or ambiguous elements.

The improvement that resulted from the use of LOCATION tags around place names (see Section 3.6) is evident in cases such as the following, that were correctly classified with XLNET when the location tags were included, but not when they were excluded (using only the text):

1. "they descended from the hills **at** West Quantoxhead."
2. "right **on** the sea front opposite St Johns Church."
3. "The trumpeter defected **to** the West in the mid-Seventies."

These examples demonstrate the importance of multi-word place names (as well as single-word) (Example 1), names that are not immediately adjacent to or the direct object or subject of the preposition (Example 2), and vague names (Example 3) in the detection of geo-spatial language. It may also be remarked that our best methods proved effective in classifying expressions as geo-spatial in those cases where the reference object is geographic but is not a place name, as in "Note Clock Tower and track to right leading to the Cricket ground".

5 | CONCLUSION AND FUTURE WORK

In this paper we have addressed the problem of disambiguation of the sense of prepositions with a particular focus on geo-spatial senses that refers to a geographical context. The motivation is that detection of such geo-spatial uses of prepositions supports the increasingly important field of geographical information retrieval which is aimed at finding documents and extracting textual information that relate to geographical locations. The use of BERT-based transformer architectures was compared to a variety of alternative methods. These included an extension of the linguistic features of Kordjamshidi et al. [2011] with features that detect

the presence of place names and geographic features; a bag of words method based on a window surrounding the word to be classified; and meta-classifiers that combined class predictions and probabilities from the bag of words method with features that indicate the counts of place names and geographic feature types in the local context. The transformer models, that are notable for incorporating transfer learning in which a pre-trained deep learning model is fine-tuned to the training data, provided clear superiority in performance for both geo-spatial (F1 0.94) and generic spatial (F1 0.95) sense detection. This performance for geo-spatial sense detection improves very significantly on previous work focused on that task. The success of the transformer model deep learning approach for geo-spatial sense detection demonstrates the benefits of the context-dependent word embedding approach relative to the use of hand-crafted linguistic features and to the use of a bag of words approach. With regard to precision however, the BERT-based methods did not improve on the SVM meta-classifier that combined predictions from a bag of words classifier with counts of place names and of geofeatures. The BERT-based methods did provide markedly superior results for recall. While the BERT-based methods provided the best F1 performance with standard text input, a significant outcome of our study of spatial preposition sense detection is to demonstrate the benefit with the BERT-based models of explicit tagging of words that represent place names. Although the generic spatial sense detection was a subsidiary aim of this work, it may be noted that F1 performance on our datasets matched results of previous published work on the application of deep learning to that task with different datasets.

We regard a significant contribution of this work to be its role in enabling much finer grained geo-referencing than is currently the norm. Most work on geo-referencing to date has been focused on determining representative coordinates for entire documents or for social media posts. Effective automation of detection of the geo-spatial use of relational terms (as presented here) will support extraction and geo-referencing of locative expressions that refer to geo-spatial entities with relative spatial relationships. Such locative expressions commonly occur in natural language descriptions of for example impacts of natural disasters; the location of road accidents; and the vast numbers of historic records of the locations at which biological or other natural environmental samples were obtained. Future work in this field can be envisaged on tasks such as improved extraction of all components of locative expressions, and more effective modelling of the applicability of individual spatial relational terms to enable determining the exact locations referred to by such expressions.

In other future work we will investigate augmenting the transformer language model approach with tagging of geographic features types (in addition to place names), as well as supplementing the text input with linguistic features. We also plan to extend these methods to address the problems of detecting other parts of speech that communicate geo-spatial relations, in particular focusing on the role of verbs when used in a spatial sense. Further investigations could also address detection of the distinction between static and dynamic uses of spatial relations in geo-spatial contexts.

References

N. Aflaki, S. Russell, and K. Stock. Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions (Short Paper). In S. Winter, A. Griffin, and M. Sester, editors, *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:6, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-083-5. . URL http://drops.dagstuhl.de/opus/volltexte/2018/9348.

Y. S. Alam. Decision trees for sense disambiguation of prepositions: Case of over. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 52–59, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-2608.

R. Artstein. Inter-annotator agreement. In N. Ide and J. Pustejovsky, editors, *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.

T. Baldwin, V. Kordoni, and A. Villavicencio. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149, 2009. . URL https://www.aclweb.org/anthology/J09-2001.

E. Cannesson and P. Saint-Dizier. Defining and representing preposition senses: a preliminary analysis. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 25–31. Association for Computational Linguistics, July 2002. . URL https://www.aclweb.org/anthology/W02-0804.

J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996. URL https://aclanthology.org/J96-2004.

K. R. Coventry and S. C. Garrod. *Saying, Seeing and Acting : The Psychological Semantics of Spatial Prepositions*. Psychology Press, July 2004. ISBN 978-1-135-43199-0. . URL http://www.taylorfrancis.com/books/9781135431990.

Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

S. Datta, Y. Si, L. Rodriguez, S. E. Shooshan, D. Demner-Fushman, and K. Roberts. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of Biomedical Informatics*, 108:103473, 2020. ISSN 1532-0464. . URL https://www.sciencedirect.com/science/article/pii/S1532046420301027.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

A. Dittrich, M. Vasardani, S. Winter, T. Baldwin, and F. Liu. A classification schema for fast disambiguation of spatial prepositions. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 78–86. ACM, 2015.

P. Doherty, Q. Guo, Y. Liu, J. Wieczorek, and J. Doke. Georeferencing Incidents from Locality Descriptions and its Applications: a Case Study from Yosemite National Park Search and Rescue. *Transactions in GIS*, 15(6):775–793, 2011. ISSN 1467-9671. . URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2011.01290.x.

J. D'Souza and V. Ng. Utd: Ensemble-based spatial relation extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 862–869, 2015.

J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. . URL https://aclanthology.org/P05-1045.

Q. Guo, H. R. Faghihi, Y. Zhang, A. Uszok, and P. Kordjamshidi. Inference-masked loss for deep structured output learning. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2754–2761. ijcai.org, 2020. . URL https://doi.org/10.24963/ijcai.2020/382.

M. Hall, P. Smart, and C. Jones. Interpreting spatial language in image captions. *Cognitive Processing*, 12(1):67–94, 2011.

K. Hassani and W.-S. Lee. Disambiguating spatial prepositions using deep convolutional networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

A. Herskovits. *Language and Spatial Cognition*. Cambridge University Press, New York, NY, USA, 1987. ISBN 978-0-521-26690-1.

D. Hovy, S. Tratz, and E. Hovy. What's in a preposition? dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*, pages 454–462, Beijing, China, Aug. 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-2052.

R. Jackendoff. *Semantics and cognition*. The MIT Press, Cambridge, MA, US, 1983. ISBN 978-0-262-10027-4.

R. Johansson and P. 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.

J. Kelleher and F. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.

A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In S. Scheider, B. Adams, K. Janowicz, M. Vasardani, and S. Winter, editors, *ACM SIGSPATIAL International Workshop on Computational Models of Place, COMP 2013, November 5, 2013, Orlando, Florida, USA*, page 62. ACM, 2013. . URL https://doi.org/10.1145/2534848.2534857.

P. Kordjamshidi, M. Van Otterlo, and M.-F. Moens. Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36, Dec. 2011. ISSN 1550-4875.

P. Kordjamshidi, T. Rahgooy, M.-F. Moens, J. Pustejovsky, U. Manzoor, and K. Roberts. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *CLEF*, 2017.

C. Kray, H. Fritze, T. Fechner, A. Schwering, R. Li, and V. J. Anacta. Transitional Spaces: Between Indoor and Outdoor Spaces. In T. Tenbrink, J. Stell, A. Galton, and Z. Wood, editors, *Spatial Information Theory*, Lecture Notes in Computer Science, pages 14–32. Springer International Publishing, 2013. ISBN 978-3-319-01790-7.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

S. C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press, 2003. .

W. M. Lim and H. T. Madabushi. Uob at semeval-2020 task 12: Boosting BERT with corpus level information. *CoRR*, abs/2008.08547, 2020. URL https://arxiv.org/abs/2008.08547.

K. Litkowski. Pattern dictionary of English prepositions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1274–1283, Baltimore, Maryland, June 2014. Association for Computational Linguistics. . URL https://aclanthology.org/P14-1120.

K. Litkowski and O. Hargraves. The preposition project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, 2005.

K. Litkowski and O. 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.

K. C. Litkowski. Feature ablation for preposition disambiguation. *Technical Report, Damascus, MD: CL Research*, 16-02, 2016.

Y. Liu, Q. H. Guo, J. Wieczorek, and M. F. Goodchild. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, 23(11):1471–1501, Nov. 2009. ISSN 1365-8816. . URL https://doi.org/10.1080/13658810802247114. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658810802247114.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL http://arxiv.org/abs/1907.11692. cite arxiv:1907.11692.

G. Logan and D. Sadler. A computational analysis of the apprehension of spatial relations. *Language and space*, pages 493–529, 1996.

U. Manzoor and P. Kordjamshidi. Anaphora resolution for improving spatial relation extraction from text. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 53–62, New Orleans, June 2018. Association for Computational Linguistics. . URL https://aclanthology.org/W18-1407.

A. Mazalov, B. Martins, and D. Matos. Spatial role labeling with convolutional neural networks. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, GIR '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450339377. . URL https://doi.org/10.1145/2837689.2837706.

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

F. Melo and B. Martins. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38, 2017. ISSN 1467-9671. . URL http://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12212.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

D. R. Montello. Scale and multiple psychologies of space. In A. U. Frank and I. Campari, editors, *Spatial Information Theory A Theoretical Basis for GIS*, Lecture Notes in Computer Science, pages 312–321. Springer Berlin Heidelberg, 1993. ISBN 978-3-540-47966-6.

T. O'Hara and J. Wiebe. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2): 151–184, 2009. . URL https://www.aclweb.org/anthology/J09-2002.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014.

B. Premjith, K. Soman, M. Anand Kumar, and D. Jyothi Ratnam. Embedding linguistic features in word embedding for preposition sense disambiguation in english—malayalam machine translation context. In *Recent Advances in Computational Intelligence*, pages 341–370. Springer, 2019.

R. S. Purves, P. Clough, C. B. Jones, M. H. Hall, and V. Murdock. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, 12(2-3):164–318, 2018. .

M. Radke, P. Das, K. Stock, and C. B. Jones. Detecting the geospatialness of prepositions from natural language text (short paper). In *14th International Conference on Spatial Information Theory (COSIT 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

T. Rahgooy, U. Manzoor, and P. Kordjamshidi. Visually guided spatial relation extraction from text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 788–794, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

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

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

P. Shi and J. Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

K. Stock. Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, 71:209–240, Sept. 2018. ISSN 0198-9715. . URL http://www.sciencedirect.com/science/article/pii/S0198971518301133.

K. Stock, R. C. Pasley, Z. Gardner, P. Brindley, J. Morley, and C. 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., Sept. 2013. ISBN 978-3-319-01789-1.

K. Stock, C. B. Jones, S. Russell, M. Radke, P. Das, and N. Aflaki. Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 36(3):547–584, 2022.

L. Talmy. *Toward a cognitive semantics*, volume 2. MIT press, 2000.

S. Tellex and D. Roy. Grounding spatial prepositions for video search. In *Proceedings of the 11th International Conference on Multimodal Interfaces, ICMI 2009, Cambridge, Massachusetts, USA, November 2-4, 2009*, pages 253–260, 2009. .

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 173–180, USA, 2003. Association for Computational Linguistics. . URL https://doi.org/10.3115/1073445.1073478.

A. Tyler and V. Evans. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press, 2003.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

J. O. Wallgrün, A. Klippel, and T. 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.

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

D. Wu, G. Cong, and C. S. Jensen. A framework for efficient spatial web object retrieval. *The VLDB Journal*, 21(6):797–822, 2012. . URL https://doi.org/10.1007/s00778-012-0271-0.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

P. Ye and T. Baldwin. MELB-YB: Preposition sense disambiguation using rich semantic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 241–244, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S07-1051.

S. Zenasni, E. Kergosien, M. Roche, and M. Teisseire. Spatial information extraction from short messages. *Expert Systems with Applications*, 95:351–367, 2018.

C. Zhang, X. Zhang, W. Jiang, Q. Shen, and S. Zhang. Rule-based extraction of spatial relations in natural language text. In *2009 International Conference on Computational Intelligence and Software Engineering*, pages 1–4, 2009.

X. Zhang, C. Zhang, C. Du, and S. Zhu. Svm based extraction of spatial relations in text. In *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, pages 529–533, 2011.