DETECTING THE GEOSPATIALNESS OF PREPOSITIONS FROM NATURAL LANGUAGE TEXT

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
degree of

Master Of Technology

in

Computer Science & Engineering

by

PRARTHANA DAS (MT17MCS007)

under the guidance of
Dr. Mansi A. Radke

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)

JUNE 2019

DETECTING THE GEOSPATIALNESS OF PREPOSITIONS FROM NATURAL LANGUAGE TEXT

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
degree of

Master Of Technology

in

Computer Science & Engineering

by

PRARTHANA DAS (MT17MCS007)

under the guidance of
Dr. Mansi A. Radke

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)

JUNE 2019

Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology, Nagpur

Declaration

I, **Prarthana Das**, hereby declare that this dissertation titled " **Detecting the geospatialness of prepositions from natural language text**" is carried out by me in the Department of Computer Science & Engineering of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution / University.

Prarthana Das
Date:

Certificate

This to certify that the dissertation titled " **Detecting the geospatialness of prepositions from natural language text**" is submitted by **Prarthana Das** in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science and Engineering, VNIT, Nagpur**. The work is comprehensive, complete and fit for final evaluation.

Dr. Mansi A. Radke
Assistant Professor
Department of Computer Science & Engineering, VNIT, Nagpur

Dr. U. A. Deshpande
Head, Department of Computer Science & Engineering
VNIT, Nagpur
Date:

Visvesvaraya National Institute of Technology, Nagpur

Non-Plagiarism Certificate

Certified that the MTech Dissertation titled "**Detecting the geospatialness of prepostions from Natural Language Text**" submitted by Prarthana Das, Enrollment No. MT17MCS007, ID No. 21515 has been checked for plagiarism using TURNITIN software and the overlap is found within prescribed limits. A summary of the report is given below:

DETECTING THE GEOSPATIALNESS OF PREPOSITIONS FROM NATURAL LANGUAGE

ORIGINALITY REPORT

12%	10%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Prarthana Das
MT17MCS007
21515

Dr. Mansi A. Radke

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

I would like to gratefully and sincerely thank my supervisor, Dr. Mansi A. Radke, for her encouragement, patience and timely guidance throughout the entire project. I am highly indebted to her for her valuable time. Without her invaluable guidance, this work would never have been successful.

I am also thankful to our HOD, Prof. U. A. Deshpande ,all the staff member & non-teaching staff of Computer Science & Engineering, VNIT, Nagpur.

Finally, and most importantly, I would like to thank my parents Mr. Asim Kumar Das and Mrs. Sibani Das. Their support, encouragement, patience, support, faith and unconditional love have given me the strength and belief. They have been my constant source of inspiration throughout the course of my life.

Prarthana Das
MT17MCS007

ABSTRACT

We often come across terms like vague terms like “near”, “in front”, “some distance away” of some named place in text. There are many instances where we need the precision in location, in cases like describing the site of an event like an accident or crime, or even the location at which a sample of natural environment was found say in applications of geology, soil science or archaeology. There is an increasing interest in detecting the presence of geospatial locative expressions that include spatial relation terms such as near or within some distance. Being able to do so will provide a foundation for interpreting relative descriptions of location. However, there were no clear definitions of geospatialness of prepositions available in literature. Our work is to detect geospatialness of prepositions and thereby calling a sentence that contains it as a geospatial expression. To do so we have broadly divided the prepositions present in the sentence as: geospatial, spatial but not geospatial and non-spatial. A spatial relation specifies how an object is located in space in relation to some reference object and the process of extracting such relations is termed as spatial role labelling. Natural language is unstructured by nature, so, from this sort of labelling we get a lot of information about the locative details a sentence is holding by extracting structured information. Here, we evaluate the use of a spatial role labelling procedure to distinguish geospatial uses of prepositions from other spatial and non-spatial uses and experiment with the use of additional machine learning features to improve the quality of detection of geospatial prepositions. An annotated corpus of nearly 2000 instances of preposition usage was created for training and testing the classiers. We leverage the use of a gazetteer and other geographic databases to customize results as per location.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURES

NER	Named Entity Recognition
NLP	Natural Language Processing
gnn	Geographic Feature Type
loc	Geographic Place Name
POS	Part Of Speech
SRL	Semantic Role Labeling
NBC	Naive Bayes Classifier
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
TPP	The Preposition Project

PUBLICATIONS

1. Mansi Radke, Prarthana Das, Christopher Jones, Kristin Stock, "Detecting the geospatialness of prepositions from natural language text". *14th International Conference on Spatial Information Theory,* Regensburg, Germany, September 9-13 2019. (Accepted; soon to be published).

CHAPTER 1

INTRODUCTION

Identifying spatial relationships between objects in text has been studied in detail over the last decade. A spatial relation explains how an object is located in space with respect to a reference object. The process of extracting such spatial relationships is termed as spatial role labeling. Natural language being unstructured in nature, spatial role labeling helps extracting structured information from the unstructured text. The larger goal in doing this could be to populate spatial ontologies to enrich them. This would facilitate search engines to give direct answers to queries by making use of the underlying structured knowledge bases. For example, consider the query - *"What is the population of Telengana?"* It gives the actual answer (that is the population figure) too along with list of relevant documents containing information about the query. Thus the population of structured knowledge bases (specifically geospatial ontologies) is facilitated by extracting geospatial relations from the natural language.

A spatial role typically consists of a trajector, landmark and a spatial indicator. The spatial indicator is typically is a preposition. When the landmark or both the trajector and the landmark are place names or geographic feature types, we call this triple as a geospatial triple and the sentence as a geospatial sentence. For example, the sentence *"The book is on the table"*, indicates a spatial relationship between the book and the table. Here the preposition "on" is the spatial indicator. However, in the sentence, *"Mumbai is in Maharashtra"*, the trajector and landmark both being geospatial entities, we call the triple <Mumbai, in, Maharashtra> as a geospatial role. We would like to point out here that irrespective of the trajector, if landmark is a geographic place name or geographic feature type, we call the triple as a geospatial triple/role. By geographic feature type we mean words like *"road", "river", "mountain", "church", "canal", "dock", "aqueduct", "street", "park", "monument"*, etc. which are indicative of a geographic feature that can be pinned on the map of the earth. (Note that in future we plan to consider our dictionary of such geographic feature types which would help us identify the geospatialness of a preposition in particular and sentence in general.) Identifying geospatial roles could be used to quantify vague terms in natural language. For example, *Maryland is near New York* and *St. Mary's Church is near Times square*. In these two sentences the first one is talking about granularity of few hundred miles while the later is talking of few miles. So, we can geo-reference the entities in the geospatial triple and quantify the vagueness of these terms like *close to, near to, far from, north of, south of*, etc.

The spatial indicator, in a sentence, is typically considered as a preposition in this work. However, words belonging to other parts of speech like verbs and adverbs could also act as spatial indicators at times. In the literature especially in the methods proposed by Kordjamshidi et al (reference needed), the researchers have confined spatial indicators to prepositions. This paper carried forward this assumption. Our focus is on English language and it uses the language specific features like grammatical properties of the language.

Automated recognition and disambiguation of geographic references in text documents has been receiving considerable amount of attention in recent years, often with the motivation of indexing the documents with regard to geographic space. The methods used to date have been dominated by a focus on identifying geographic names, i.e. toponyms, and using these directly as the basis for geographic footprints for text expressions or entire documents. The assumption, however, is that the references are absolute in the sense that the toponym provides the actual location referred to. While this is a reasonable default assumption, it is very common to refer to locations in an indirect manner using spatial relations, such as near, at, close to, north of etc., relative to a reference location. These expressions often take the form *of triples of a subject (or located object), the spatial relation and an object (the reference location)*, as in *"Kalighat Temple near Rashbehari Crossing"*.

1.1 Problem Statement

Effective methods for modelling vague spatial relations such as "near", "far" and reliable identification of the presence of relative location descriptions in natural language texts are required as part of the process of extracting and interpreting indirect geographic references and to retrieve other geospatial facts that associate an event or some other object with a reference location, as for example in *"Netaji was born in Kolkata"*. Our work is largely based on presenting methods for automatic detection, in sentences, of spatial relational terms, in particular prepositions, that are used specifically in a geospatial sense and we distinguish these from prepositions that have other spatial senses and from prepositions that have no spatial meaning. We are interested in the ability to distinguish between spatial and geospatial senses of prepositions, as this is important for detecting text that can be geo-referenced and thus mapped on a geographical scale (in contrast to text that describes location inside a room, or on a person's body), a goal that is useful in a wide range of application areas.

As mentioned earlier, the geospatial role labelling task finds the spatial relations in a sentence, each of which can be represented as a triple <trajector, spatial indicator, landmark>. When the landmark is a geographic place name or a geographic feature type, we call the triple as a geospatial role, irrespective of whether the the trajector is a geospatial entity or not. We assume that the sentence is aprori tokenized into individual words or phrases as applicable. We use the set of roles defined by Kordjamshidi et al. namely trajector, landmark, spatial indicator and none. Additionally, we retain their definitions of implicit and explicit trajcetor and landmark. The trajector, landmark and spatial indicator of a triple do not overlap in our methods too. A certain token, word or phrase could act as either trajector or landmark or spatial indicator or none in a particular sentence. However, it assumes a single class at a time. There could be multiple spatial/ geospatial roles in a sentence. The trajector of one role could be the landmark of another or vice versa. We plan to extend this notion by defining hierarchical spatial roles wherein there are multiple roles in a sentence and possibly multiple spatial indicators too. This has been elaborated in greater details in chapter 5.

Given a sentence S, the set of spatial indicators of S is denoted as I. The indicator function I defined over all tokens of sentence S is as follows:

- I = 2, if the preposition indicates a geospatial sense in addition to spatial

- I = 1, if the preposition indicates a spatial sense but not geospatial

- I = 0, if the preposition indicates no spatial sense at all

Every geospatial sentence is also spatial, however, the classes geospatial, spatial and non spatial are mutually disjoint and the geospatialness of a preposition takes precedence over its spatialness. Here, we would also like to mention that to declare a preposition as any of the above three we devise the technique described ahead however to declare the sense of sentence/expression as geospatial or spatial but not geospatial or non-spatial we need to have information about all the prepositions present in the sentence and then decide.

1.2 Potential Applications

A potential application of this work could be combining spatial and temporal information together to extract meaningful information which could be populated in ontologies to aid the search engines in giving precise answers. For example *"Rabindranath Tagore was born in Calcutta on 7th May 1861"*. Here, the Named entity recognition software (NER) like Stanford NER could be used to extract date or time information along with detecting the geospatialness of the preposition "in".

Another objective of geospatial relation extraction is to provide information in which one geospatial entity is associated with another geospatial entity indicating some spatial relationship between them such as containment or proximity. Additionally there could be a situation where a geospatial object is associated with some attribute /property such as its type, age, population, owner, or event etc. For example, *John was born in Seattle* or *An earthquake occurred in Japan.*

1.3 Contributions of this Dissertation

The first contribution of our work is to evaluate the methods proposed by Kordjamshidi et. al [7] in identifying geospatial roles. We experiment with applying Kordjamshidi's methods to evaluate whether a preposition in a sentence is geospatial, spatial, or non-spatial. More precisely, we detect geospatial prepositions and distinguish them from those which are "spatial but not geospatial".

The second contribution of our work is to create a corpus for extracting and identifying the geospatial roles from natural language text. This would be a step towards understanding geospatial language and auto generating it too.

1.4 Terminology Used

This section talks about the various terms, tool(s) and approaches that we have frequently addressed in our work.

1.4.1 Spatial Relationships

A spatial relation specifies how some object is located in space with respect to some reference object. One of the essential functions of language is conveying *spatial relationships* between objects and their relative or absolute location in space.

For example, *"The book is on the cupboard"* tells us about the *spatial configuration* between the objects book and cupboard in some space.

1.4.2 Spatial Role Labelling

The process of extracting spatial relations from natural language text is termed as spatial role labelling. As a result of this sort of labelling we get triples from a sentence. The triples are of the form of *{trajector(located object),spatial indicator(spatial relation), landmark(reference location)}*[7].

As an example, for the sentence, *"The book is on the cupboard"*:

- "book" is trajector
- "on" is spatial indicator
- "cupboard" landmark

The triple obtained is *{book,on,cupboard}*.

1.4.3 Geospatial Role Labelling

The geospatial role labeling task finds the spatial relations in a sentence, each of which can be represented as a triple as aforementioned. However, the landmark is a geographic place name or a geographic feature type, we call the triple as a geospatial role, irrespective of whether the trajector is a geospatial entity or not.

By geographic feature type we mean words like "road", "river", "mountain", "church", "canal", "dock", "aqueduct", "street", "park", "monument" etc. which are indicative of a geographic feature that can be pinned on the map of the earth.

1.4.4 Geospatial Expression

Geospatial expressions have the following characteristics:

- They include a spatial relation, which may take a form of a preposition (or a verb or other word or group of words) that describes the spatial location or movement of one object relative to another.
- The reference object of spatial relation is a geographical object which may be outdoor, static (in normal course of events) or in a scale likely to occur in a map. The object in question can be a geographical place name or geographical feature type.

For the sentences,

- *The seat on the verandah* :
  - "on" is a spatial indicator, and verandah is part of a house, the scene is outdoors and the verandah is static.
  - The spatial indicator is geospatial one which in turn makes the sentence a geospatial expression.
- *Taj Mahal is in Agra*
  - "in" is a sptial indicator, and Agra is a city (geographical place name) in India.
  - The spatial indicator is geospatial one which in turn makes the sentence a geospatial expression.

1.4.5 Spatial Expression

Spatial expressions are defined by extension from geospatial expressions and have the following characteristics:

- They include a spatial relation, which may take a form of a preposition (or a verb or other word or group of words) that describes the spatial location or movement of one object relative to another.

- The reference object of spatial relation in this case may be indoor, mobile or in a small scale.

In expressions that contain both geospatial and spatial elements, the geospatial classification takes precedence.

For the sentences,

- *A few bells still exist, hanging on the rood-screens in East Anglian churches.*
  - It contains both spatial (bells hanging on the rood-screens) and geospatial (in East Anglian churches) elements.
  - It's finally classified as geospatial.
- *She is sitting at the back of the room.*
  - it gives the hint of an indoor place in a small scale.
  - The spatial indicator is spatial one which in turn makes the sentence a spatial expression.

1.4.6 Non-spatial Expression

These contain all expressions that are neither geospatial nor spatial. They do not contain any spatial relations or contain spatial relations in a metaphoric sense.

For the sentences,

- *I thought perhaps you were still mad at me*
  - The above expression contains no spatial relation.
- *You are accusing me on false grounds.*
  - The above expression contains spatial relation in a metaphoric sense.

1.4.7 Part of Speech Tagger

A Part-Of-Speech Tagger (POS Tagger)[17] is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

1.4.8 Dependency Parser

Dependency parsing is the task of recognizing a sentence and assigning a syntactic structure to it. The most widely used syntactic structure is the parse tree which can be generated using some parsing algorithms. These parse trees are useful in various applications like grammar checking or more importantly it plays a critical role in the semantic analysis stage.[16]

1.4.9 Semantic Role Labeler

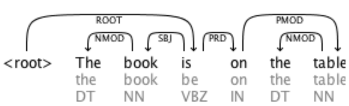
In natural language processing, semantic role labeling (also called shallow semantic parsing) is the process that assigns labels to words or phrases in a sentence that indicate their semantic role in the sentence, such as that of an agent, goal, or result. It consists of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. [5]

1.4.10 Parsing a sentence by POS Tagger, Dependency Parser and Semantic Role Labeler

Example of the above three from the LTH software app (http://barbar.cs.lth.se:8081/parse [4]), for the sentence *The book is on the table*, is given in the image below.

	The	book	is	on	the	table

Parsing sentence required 5ms.



ID	Form	Lemma	PLemma	POS	PPOS	Feats	PFeats	Head	PHead	Deprel	PDeprel	IsPred	Pred
1	The	_	the	_	DT	_	_	_	2	_	NMOD	_	_
2	book	_	book	_	NN	_	_	_	3	_	SBJ	_	_
3	is	_	be	_	VBZ	_	_	_	0	_	ROOT	_	_
4	on	_	on	_	IN	_	_	_	3	_	PRD	_	_
5	the	_	the	_	DT	_	_	_	6	_	NMOD	_	_
6	table	_	table	_	NN	_	_	_	4	_	PMOD	_	_

Figure 1.1: **Example of a Sentence Parsed by the LTH Software Application**

The abbreviations used in the software are stated below:

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition or subordinating conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun

18. PRP Personal pronoun
19. PRP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non3rd person singular present
32. VBZ Verb, 3rd person singular present
33. WDT Whdeterminer
34. WP Whpronoun
35. WP$ Possessive whpronoun
36. WRB Whadverb

Figure 1.2: **Expansion of Abbreviations used in Output Obtained by the LTH Software Application**

1.4.11 Expat

It is a command line natural language processing application uses word patterns to identify important elements in sentences.

Expat generates twenty seven attributes from a sentence, of which two are of our interest viz. geographic location(location) and geographic feature type(gnn). (both are numeric fields).

For the sentence, *"Accolades include ABIA Australian Bridal Industry Awards Winner of 2010 Club Reception Venue for Victoria Winner of 2011 Club Reception Venue for Victoria"*, the output generated by Expat[14] is:

NOUN SVB ORG NOUN of 2010 GNN NOUN for LOCATION NOUN of 2011 GNN NOUN for LOCATION",0,1,0,1,0,0,0,0,0,0,0,0,0,0,1,5,0,0,0,2,0,1,2

Of which the underlined values are of our interest. They give the values of geographic location(location) and geographic feature type(gnn) respectively.

1.4.12 Evaluation metrics

The evaluation metrics of this work are precision, recall, F1 measure and accuracy. These values are obtained with the help of a confusion matrix.

1.4.12.1 Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which a classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Table 1.1: **Example of Confusion Matrix**

	Actual Value : C1	Actual Value : C2
Predicted Value : C1	True Positive	False Positive
Predicted Value : C2	False Negative	True Negative

The terms in the cells of the table can be defined as follows. Here, we assume that we have two classes C1 and C2:

- True Positive : When an element is predicted as belonging to C1 and it is actually C1 so.
- False Positive : When we predict that an instance belongs to C1 but in reality it belongs to C2.
- False Negative : When an instance is actually belonging to C1 but is predicted that it belongs to C2.
- True Negative : When an instance does not belong to C1 and is predicted the same too.

Now, let us have a look at the definition of the evaluation metrics with respect to class C1.

1.4.12.2 Precision

Tells us about that percentage of items that are selected is correct. It is given by the formula:

$$\text{Precision}=\frac{TP}{TP+FP}$$

1.4.12.3 Recall

Tells us about the percentage of correct items that are selected. It is given by the formula:

$$\text{Recall}=\frac{TP}{TP+FN}$$

1.4.12.4 Accuracy

Tells us about the percentage of correct predictions among all the predictions made. It is given by the formula:

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

1.4.12.5 F1 Measure

It is the harmonic mean of the values of precision and recall.It is given by the formula:

$$\text{F1}=\frac{2*Precision*Recall}{Precision+Recall}$$

The remainder of this thesis is organized as follows. Chapter 2 details out the various works in literature that are related to this work or have influenced it in one way or other. Chapter 3 explains the methodology or the approach of identifying the geospatialness of the prepositions in English sentences in detail. Chapter 4 gives the details of the data set used, the anootation of the data set, the experiments performed and also about the results obtained and their analyses. Chapter 5 concludes the paper pointing out some directions for future work.

CHAPTER 2

LITERATURE REVIEW

A method designed specifically to detect whether a preposition has a spatial sense was presented by Kordjamshidi et al. [7] in the paper on spatial role labelling in the context of spatial relation extraction. For the sake of completeness, we reiterate over the definition of spatial relation and spatial role labelling. A spatial relation specifies how an object is located in space with respect to some reference object. The process of extracting such relations is termed as spatial role labelling. By spatial role labelling, we get triples of the form : *<trajector(located object), spatial indicator(spatial relation), landmark(reference location)>*.

There are two approaches mentioned, in Kordjamshidi et al. [7], for spatial role labelling. The first one is the pipeline approach. Here, an input sentence is made to undergo a tokenizer stage of the pipeline from where each token of the sentence is passed through a Part of Speech Tagger (PoS) tagger. The sentence under consideration is also made to undergo a dependency parser and a semantic role labeller (the LTH software http://barbar.cs.lth.se:8081 from [4]). If a word is marked to be belonging to the category of prepositions by the POS tagger,then a Naive Bayes classifier is used to predict as to whether or not it is used in a spatial sense in that particular sentence context. The features used by the classifier are based on output from the POS tagger, the dependency parser and the semantic role labeller. If the preposition is determined to have a spatial sense, then it is passed to a second stage of the pipeline which identifies the trajector and the landmark with respect to the spatial indicator. This second stage uses probabilistic graphical models, in particular a Conditional Random Fields classifier, which again takes a variety of features generated by the initial parsing of the sentence. A triple, as mentioned previously, is returned as output by the pipeline. The second approach offered by Kordjamshidi et al. [7] uses joint learning in which all three of trajector, spatial indicator and landmark are detected simultaneously. In this work, we modify the first step of the pipeline method i.e their method for detecting the spatial sense of prepositions by adding additional features. We leverage the use of a gazetteer for this purpose.

A method for detecting just the spatial relation and the reference object of spatial relations was described by Liu [9] where these partial relations were described as degenerate locative expressions (DLE). The approach is analogous to methods of Kordjamshidi et al, though they employed a smaller set of features for machine learning, that did not include dependency relations or semantic roles. An evaluation of the method in [10] obtained an F1 score of 0.76 when applied fully automatically to their TellUsWhere corpus on which it was trained. Note that no distinction was made in that work between geospatial and other spatial senses of prepositions. The method of [9] to extract DLEs was also exploited in Khan et al. [6] in which locative DLEs which explicitly encode spatial relations, with prepositions such as "near" and "in", were distinguished from partial DLEs where a preposition such as "to" was not regarded as conveying explicit spatial information. A rule based approach was employed to extend the latter to an explicit spatial DLE when it was used as part of a spatial relation such as "next to". This technique was part of a procedure to extract spatial triples by matching structures from the Stanford parser, of the form <governor, preposition, dependent>, with locative DLEs that used the same preposition. The governor would then serve as the located object of a spatial triple.

As part of a process of creating a corpus of geospatial sentences, Stock et al.[15] employed a set of language patterns to detect various ways in which geospatial information is described.This included a pattern to recognize when a place name or place type is preceded by a spatial relation which could be a preposition (though other parts of speech were also considered to represent spatial relations). They obtained a precision of 0.66 when applying these methods to detect geospatial expressions. A specialized collection of spatial relational expressions was created by Wallgrun, Klippel and Baldwin[18]. They used search patterns to query the web to find expressions that contained any of the three relations of near, close and next to. Their approach therefore constrained the results to include the specified spatial relation. They also confined the expressions to include specified types of located and reference objects. Our work differs from that in allowing any spatial relation that is classed as a preposition and in using a machine leaning approach to determine the geospatial or other spatial sense of the preposition.

However, a clear, unequivocal and specific definition of what counts as spatial language has not been given yet in literature. While the definition of locative expressions may seem simple and clear, in practice there are many borderline cases that rise some questions about the definition of spatial language. The challenges become even more complex when we attempt to define geospatial language, and distinguish it from spatial language, and the definition and syntax of geospatial language as distinct as spatial language have been given.

In this work we aim to provide a clear definition of what we mean by spatial and geospatial language, the first of which differs from the locative expression described by Herskovits(1987)[3] in that it is broader, and the second of which is narrower than the notion of geospatial language that relies only on place names. We also beyond the place name focus of Wolf, Henrich, Blank, (2014)[19] by describing a number of challenges and borderline cases in the definition of spatial and geospatial language, including the question of what counts as a spatial relation.

CHAPTER 3

PROPOSED APPROACH

In this chapter, we discuss about the methods followed to detect the geospatialness of a preposition in a sentence. First, we start by giving the work flow diagram that we used followed by the explanation of each of the modules in it. We finally conclude this section by giving the details of the classifier used and the data handling methods.

3.1 Work Flow Diagram

The basis of the work for this project is based on the methods devised in Kordjamshidi et al. [7]. The following diagram walks us through the work flow in pipeline model of [7]. In brief pipeline module works in two steps:

- The first stage identifies the spatial indicator i.e. preposition of the sentence and classifies the preposition as having a spatial sense or not depending on the context and usage.
- The second stage identifies the trajector and landmark of the preposition if the first stage marks it as spatial otherwise ignores it.

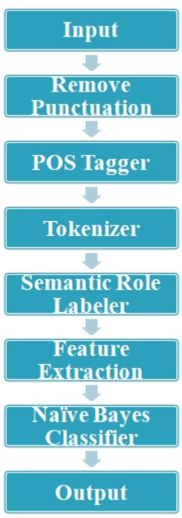


Figure 3.1: **Work Flow Diagram for Pipeline Approach of Kordjamshidi et al. [7]**

3.2 Details of the Modules:

The mode of working of each of the modules of the work flow in 3.1 is described as follows:

- Input consists of sentences in English language from various types of data sets (which is discussed later on in detail).
- Sentences from the data set are taken and passed through RemovePunctuation module which necessarily removes the punctuations from the sentences.(Sentences are labeled apriori as belonging to the class 0(non spatial), 1(spatial but not geospatial), or 2 (spatial and geospatial.)
- The resultant sentence is passed through Stanford POS(Part of Speech) Tagger which identifies prepositions in the sentence.
- Each of the sentence is passed through Tokenizer module which considers each word of the sentence as a token.
- Output of the previous module is passed through SRL (Semantic Role Labeller) Pipeline which does semantic role labeling of each of the tokens.
- Feature extraction of semantically labeled token is done here.
- Naive Bayes Classifier is used to predict the class of the given sentences of the data set and the confusion matrix is used to calculate the performance metrics.

3.3 Preliminary Work Done

Our major contribution is in the enhancement of the Feature Extraction module of 3.1 of pipeline method of [7]. Originally, the sense of a preposition was used to tell whether a preposition is spatial or non-spatial. Now, instead of providing that we are providing the label the sentence already had to declare if a preposition is spatial or non-spatial. We performed various rounds of experiments with this. The following table lists the features used by Kordjamshidi et al. in [7]. As indicated previously, these are obtained from a combination of a POS (Part of Speech) tagger, a dependency parser and a semantic role labeller.

Table 3.1: **Features from [7] used in Detecting the Sense of a Preposition**

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

3.4 Detecting Geosptaial Natural Language

3.4.1 With Kordjamshidi et al. [7] Features Only

This is an extension to the work done in the preliminary experimentation. In the preliminary work, only two tags were used, spatial (1) and non-spatial (0). We propose the notion of geospatialness and now consider the data with three tags which are

- non-spatial (0)
- spatial but not geospatial (1)
- geospatial (which is spatial also) (2)

As we did not have such tagged data, we first annotated it as mentioned in section 4.2. With this tagged data, we experimented with the methods proposed by Kordjamshidi et. al. in [7] and evaluated their effectiveness on this data. The results are given in detail in section 4.5.2.1. We further added some new features by leveraging the use of gazetteer and Expat. The geospatialness of an sentence/expression depends on the spatial sense of all the prepositions in the sentence/expression.

3.4.2 Addition of New Features

In this work, we modify the first step of the spatial role labelling pipeline method of [7], i.e. their method for detecting the spatial sense of prepositions, by adding additional features for machine learning. The features used in the original classifier are listed in 3.1. We extend the method to add additional features that indicate whether a place name or a geographic place type is present in the expression that includes the target preposition. The presence of a place name is detected with the GeoNames gazetteer [2], while the presence of a place type is detected with a dictionary of geographic place types. The Expat[14] application was used to generate these features (location and geographic feature type patterns).

3.5 Detecting Geospatialness of Preposition

To detect the spatial sense of a preposition we need details of just the preposition in question and nothing more. For the sake of our work we employ the methods discussed in the above section, the difference being that we only concentrated on a single preposition rather than all in the sentence, we also add the following.

3.5.1 Preposition Level Analysis and Expat

A typical sentence may contain more than one preposition and each of them may not emanate the same spatial sense. We were concentrating on the entire sentence whilst detecting geographic name place (location) and geographic feature types (gnn).

For example, the sentence : *"19th century restaurant on Tettye Hill with large interior and beer garden"* has two prepositions : *on* and *with*. But, since, we are doing sentence level analysis, we pass the same sentence twice through Expat and essentially get the same outputs. The output being:

- *19th century GNN on LOCATION with NOUN GNN", 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 2*

To combat with the situation we extracted the head words (head1 and head2) from each preposition-specific sentence from the set of Kordjamshidi et. al. [7] features that has been previously generated.

- Head1: words that directly depend on the preposition
- Head2 : words on which the preposition is directly dependent.

Continuing with the example stated above, the head words are:

- Head1 and Head2 for “on” : restaurant and hill respectively
- Head1 and Head2 for “with” : restaurant and interior respectively

The output given by Expat for these:

- *"restaurant on Hill",,"GNN on GNN", 0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,2*
- *"restaurant with interior",,"GNN with NOUN", 0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1*

We performed our set of experiments with this approach.

3.6 Classification Task

Detection of spatial sense ultimately boils down to a classification problem. We have used a Naive Bayes multi-class classifier with three output classes of Geospatial,Spatial but not geospatial, and Neither geospatial nor spatial. We also used Naive Bayes binary classifiers for each one of these three classes vs the other two classes. Also, different types of cross validation methods were adopted by us. These methods are used for forming the training and testing data sets used for classification.

3.6.1 Naive Bayes Classifier

The Naive Bayes Classifier[12] applies to learning tasks. A training set is created with a group of attribute values all of them with a certain target value. When a completely new a group is presented to this it is the task of the NBC to assign a target value to it.

Bayesian conditional probability is used to mathematically represent the aforementioned information:

$$v_{MAP} = argmax_{(v_j \varepsilon V)} P(v_j | a_1, a_2, ..., a_n)$$

where,

- v_{MAP} is the target value to be assigned
- v_j is a target value
- V is a finite set of target values
- $(a_1, a_2, ..., a_n)$ is the set of attributes

Rewriting the above equation using Bayes theorem we get,

$$v_{MAP} = argmax_{(v_j \varepsilon V)} P(a_1, a_2, ..., a_n | v_j) P(v_j)$$

It is extremely easy to estimate $P(v_j)$ simply taking the frequency of each v_j. However, computation of $P(a_1, a_2, ..., a_n | v_j)$ is not feasible unless we a huge data set. The problem is that we we need to see every instance in instance space many times in order to obtain reliable estimates.

Herein, NBC uses a simplifying assumption : attribute values are conditionally independent for a given target value. v_{NB} being the target output value by NBC given by ,

$$v_{NB} = argmax_{(v_j \varepsilon V)} P(v_j) \prod_i P(a_i | v_j)$$

For our work, we grouped the outputs generated from the NBC. *(This is applicable only for geospatial language detection.)* This is due to the fact that a sentence may have more than one preposition so an individual prediction by the classifier does not really make much sense and at times is anomalous. For example, a sentence a 3 prepositions then in the feature file it has 3 entries. An entry individually does not make any sense also if the 3 entries are giving separate predictions then it's anomalous.

After obtaining the predictions from the classifier we grouped the output based on preposition count and the final prediction was based on the priority of the element which has highest priority among them. The priority is (highest to lowest): geospatial (2), spatial (1) and non-spatial (0). Same grouping is done to the test data set as well. More about the classifiers is explained in the later section.

3.6.2 Cross Validation

For a prediction problem, a model is generally fed with a data set of known data: training data set, and a set of unknown data against which the model is tested: test data set. The target is to have a data set for testing the model in the training phase and then provide insight on how the specific model adapts to an independent data set.This test data set may or may not belong the training data set. However, it is better to have a completely different test data set.[13]

This forms the heart of a prediction problem. Since, the performance of the same largely depends how well it is trained. Let us know look at the cross validation methods used by us.

- **Hold Out Validation** :Here, the splitting (90/10, 80/20 or 70/30) was done in such a way that the training part consisted of the initial 90% part (say) and the rest 10% part was for testing.
- **Randomized Cross Validation** : Here, data splitting is done as well but keeping in mind the continuity. By continuity we mean that a sentence may contain more than one preposition and we are considering each such case. So the continuity ascertains that a chunk (preposition and sentence) remains in the train/test set and not undergo any partition.

  Let us imagine that our corpus consists of 6 sentences and the PrepCount File (contains the number of prepositions per sentence) looks something like this:

Table 3.2: **Example of File Required for New Cross Validation Model Implementation**

Sentence	Preposition Count
S1	2
S2	2
S3	1
S4	2
S5	6
S6	5

The feature file generated contains 18 features in the same order. So we cannot just randomly pull out an entry in feature file and put it in train/test set. We must adhere to continuity; that is governed by the preposition count.

Suppose, we want to do a 70/30 split then for 3.2, we need to do the following:

- In that case we need ceiling(6*0.3)=2 entries in test set.
- Randomly we select to 2 such entries from PrepCount file (this file contains the number of prepositions per sentence). Let they be for S2 and S4.
- Now, whilst iterating the feature file, whenever we encounter the index for S2 we need to the next 2 values into the testing realm. S3 goes to train set as for S4 it goes to test set for 2 counts.

This gives a very randomized training data set not compromising at the continuity.

CHAPTER 4

EXPERIMENTS AND RESULTS

This chapter can be broadly divided into two parts. The first part gives the details of the data set used and its annotation. The second part summarizes the experimental results obtained and their analyses.

4.1 Dataset Used

The data sets used by us for experimentation purposes are listed below:

4.1.1 TPP Dataset

It contains 34 separate XML files, one for each preposition, totaling over 25000 instances ; each sentence contains one example of the respective preposition. This standard test and training data was provided by the SemEval-2007 challenge [Litkowski and Hargraves 2007][8].

4.1.2 The Leftovers of The Nottingham corpus

Approximately 25000 labeled (spatial or non-spatial) sentences of which 5000 were used.

4.1.3 The Data Labelled by The Amazon Mechanical Turk

About 7000 sentences (6735 to be specific) obtained during the creation of the Nottingham corpus and The Preposition Project (TPP) in 1:2 ratio were used for the purpose of experimentation. This was done by the subject matter experts of Amazon Mechanical Turk[1].

4.1.4 Corpus Creation for Inclusion of New Features

A dataset of 696 sentences was derived from two sources. 196 of the sentences came from the source of about 26,000 sentences that were used in the process of creating the Nottingham Corpus of Geospatial Language (NCGL) [15]. These sentences were harvested from the web using the algorithm described in [15], and was thus biased towards retrieving geospatial content, but also included very few spatial (but non-geospatial) expressions as well as some uses of prepositions that are non-spatial in their sense.

To make up for the balance between the three classes, we used the data set labeled by the experts of Amazon Mechanical Turk. We selected the first 500 sentences which were marked spatial by them.

Each sentence could have multiple prepositions in it. We considered a tuple <Sentence, Preposition> as a unique instance. So, if a sentence S had two prepositions p1 and p2, we created two instances out of it, namely <S, p1> and <S,p2>. Thus 1877 instances were generated from a total of 696 instances (based on the number of prepositions per sentence).

This dataset is used to implement the method proposed in 3.5.1

4.1.5 GeoNames

It is a gazetteer which covers all countries in the globe and contains over eleven million geographical place names. They are available for download; devoid of any charges [2]. The presence of a place name is detected by making use of this gazetteer.

4.2 Annotation of the Data Set

The following annotations were done manually by us:

- It was discovered that the tags of Nottingham Corpus were not very reliable. So, we manually tagged the first 1000 sentences of the corpus into spatial or non-spatial.
- Manually annotated 100 sentences of Nottingham Corpus for the purpose of geospatial role labeling.

- For the corpus creation discussed in 4.1.4, initially, 1000 sentences(with 2804 instances) from Nottingham Corpus and 500 sentences (with 1376 instances) from Amazon MTurk(marked as spatial) were annotated manually.

  Annotation was conducted through an iterative process. One person annotated all sentences, a subset of which were then checked by the other followed by a discussion of disagreements. Finally, an inter-annotator agreement[11] of 0.75 was reached for the 696 sentences.

  Some disagreements that occurred while annotating:
  - *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.*
    * Annotator -1: Marked "of" of Republic of China as geospatial since China is a country.
    * Annotator-2: Marked the same as non-spatial. Since, China is of course geospatial, but this of preposition refers to its status as a republic (the kind of political system it uses), which is not geospatial.
    * Annotator-2 had the last word.
  - *After 50m you will reach a road with wide verges where you turn left toward Lambley.*
    * Annotator-1 : "After" does not have any spatial sense.
    * Annotator-2 : You are travelling along for 50m and then you reach a road – after is used to explain the geospatial arrangement of different locations (as well as temporality).
    * Annotator-2 had the last word.

4.3 Classifiers in Detail

4.3.1 Spatial Role Labelling

The Stanford POS Tagger is used to identify preposition in a sentence. This is used in implementation of the method discussed in 3.3. The preposition detected can be any of the following:

- Tag 0 : Non-spatial
- Tag 1 : Spatial

Here, tag is the value assigned to differentiate between classes in a classifier.

4.3.2 Geospatial Role Labelling

Stanford POS Tagger is used to identify preposition in a sentence. The preposition detected can be any of the following, where tag is the value assigned to differentiate between classes in a classifier:

- Non-Spatial vs Spatial but not geospatial vs Geospatial aka Experiment-1
  - Tag 0 : Non-spatial
  - Tag 1 : Spatial but not geospatial
  - Tag 2 : Geospatial

A multi-class Naive Bayes Classifier is used to implement the method discussed in 3.4. Along with this we implemented the following binary classifiers as well:

- Non-spatial vs (Geospatial or Spatial but not geospatial) aka Experiment-2
  - Tag 0 : Non-spatial
  - Tag 1 : Geospatial or Spatial but not geospatial
- Geospatial vs (Spatial but not geospatial or Non-spatial) aka Experiment-3
  - Tag-0: Non-spatial or Spatial but not geospatial
  - Tag-1: Geo-spatial
- Spatial but not geospatial vs (Geospatial or Non-spatial) aka Experiment-4
  - Tag 0 : Spatial but not geospatial
  - Tag 1 : Geospatial or Non-spatial

This enables a finer grained analysis of language content.

4.4 Experiments Performed

The hardware used for the experiments:

- Processor : Intel(R) Core(TM) i5-7200 CPU@2.50GHz 2.70 GHz

- Random Access Memory : 4 GB

The softwares used for the experiments:

- Ubuntu 14.01 LTS
- Eclipse 4.4.2 IDE
- Python Scikitlearn Library
- Stanford Tokenizer
- Stanford Part of Speech (POS) Tagger
- Semantic Role Labeler
- Expat 1.0.0
- Stanford CoreNLP

The programming languages used:

- Java programming language
- Python programming language

4.4.1 Preliminary Work Done Implementation

The implementation of the method discussed in 3.3 is pretty straightforward using Kordjamshidi et. al. [7] features only. All we did was:

- Implemented binary Naive Bayes Classifier for the classifications.
- The classifier was trained on TPP Dataset and tested with Nottingham Corpus (with existing labels). Incidentally, we used about 5000 sentences (4979 to be specific) for this purpose. 10-fold cross validation on Nottingham corpus with existing labels was also performed.
- The classifier was trained on TPP Dataset and tested with Nottingham Corpus however, we used 100 sentences manually labelled by us. 10-fold cross validation on Nottingham corpus (manually labelled) was also performed.
- Training and testing was done in splits of 90/10,80/20 and 70/30 respectively.
- The performance measures of these methods were also computed.

4.4.2 Implementation of Detection of Geosptaial Natural Language

Here, we discuss the various experiments performed to implemented the methods discussed in 3.4 and 3.5. They are divided in the following ways:

4.4.2.1 With Kordjamshidi et al. [7] Features Only

- Both training and testing was done using Amazon Mechanical Turk data using the cross validation methods mentioned in 3.6.2.
- Training and testing was done in splits of 90/10,80/20 and 70/30 respectively.
- The data obtained from classifier is made to undergo grouping so that a sentence/expression has a spatial sense based on all of the prepositions present in it.
- We used Kordjamshidi et. al. [7] features only.
- Experiments : 1-4 as mentioned in 4.3.2 were implemented for both the aforementioned cases.The tweaking of the data obtained from the classifier is done in the same way as discussed in 3.4.
- The performance measures of each of these methods were also computed.

From the already extracted features we could already tell if a certain expression is geographical place name or geographical place type.Using the GeoNames geographical database and Expat with the help of the said features *we could tell if an expression is a feature type or geographical place name by a simple OR operation.* .We conducted all the work stated above along with this new addition.

4.4.3 Implementation of Detection of Geospatialness of a Preposition

4.4.3.1 With New Features

New features are the geographical place(loc) names and geographical feature types(gnn) present in a sentence; it is extracted using Expat. Here, each of the experiments mentioned in 4.3.2 were conducted (along with new ones) using the following feature sets:

- Only on Kordjamshidi et. al. [7] features using Naive Bayes Classifier.

- The two new features (location and gnn) were included as independent ones to Kordjamshidi et. al. [7] features using Naive Bayes Classifier.
- Binary OR of the two new features (location and gnn) were included along with Kordjamshidi et. al. [7] features using Naive Bayes Classifier.
- Sum of the two new features (location and gnn) were included along with Kordjamshidi et. al. [7] features using Naive Bayes Classifier.

The following points should be stated:

- This chunk of experiments is applicable to sections 3.5 at both sentence level and preposition level (with respect to head words).
- The data set used here is the corpus created in 4.1.4.
- The training and testing was done using the randomized cross validation method method in 3.6.2.
- Training and testing was done in splits of 90/10,80/20 and 70/30 respectively.
- The data obtained from classifier is used as it is without any further grouping.
- The performance measures of each of these methods were also computed.

4.5 Results Obtained and Analysis

This section is one of the crucial part of our work. It gives us an empirical view of our work. It gives us the numbers that tells us about the performance of our approach. The analysis of these results are also here in this section.

4.5.1 Results of Preliminary Work Done

Table 4.1: **A Comparative Analysis of Training with TPP and Testing on the Nottingham corpus(leftovers) based on tag**

Type of Tag	Precision	Recall	Accuracy	F1
Existing	0.409	0.584	0.495	0.48
Manual	0.395	0.847	0.433	0.539

Table 4.2: **A Comparative Analysis of Training with TPP and Testing on the Nottingham corpus(leftovers) with 10-fold Cross Validation based on tag**

Type of Tag	Data Split	Precision	Recall	Accuracy	F1
Existing	90/10	0.679	0.9217	0.7026	0.7815
Existing	80/20	0.6778	0.92598	0.7056	0.78268
Existing	70/30	0.682	0.917	0.7052	0782
Manual	90/10	0.902826	0930996.	0.8875	0.915659
Manual	80/20	0.849352	0.949408	0.83617	0.895606
Manual	70/30	0.869084	0.951116	0.847143	0.907206

Analysis: The set of experiments were performed both with existing tags and manually labelled by us. We can observe a sharp rise in precision, recall, accuracy and F1 values. The rise is remarkable for cross-validation methods this might be due to the fact that the training data is very enriched due to 10-fold cross-validation. In any case, manually labelled tags show much better results. So, we can infer that the existing tags are not very reliable.

4.5.2 Results of Detection of Geospatial Natural Language

4.5.2.1 With Kordjamshidi et al. [7] Features Only

Table 4.3: **Results for Experiment-1 by Hold-Out Validation**

Data Split	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
90/10	0.66	0.94	0.78	1	0.056	0.11	0.81	0.55	0.66
80/20	0.601	0.953	0.74	0.8	0.024	0.046	0.791	0.475	0.594
70/30	0.614	0.9	0.73	0.8	0.038	0.073	0.72	0.48	0.577

Table 4.4: **Results for Experiment-2 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.741154562	0.980295567	0.844114528
80/20	0.711610487	0.983182406	0.82563824
70/30	0.716666667	0.958833619	0.82024945

Table 4.5: **Results for Experiment-3 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.773134328	0.773134328	0.773134328
80/20	0.768987342	0.800658979	0.784503632
70/30	0.794994041	0.717204301	0.754098361

Table 4.6: **Results for Experiment-4 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.886855241	0.998127341	0.939207048
80/20	0.863521483	0.999025341	0.926344329
70/30	0.868824532	0.999346832	0.929526124

Analysis : For both the hold-out validation and randomized cross-validation methods implemented by us; the results were not very satisfactory. But then again, we used only Kordjamshidi et. al. [7] features. Since, F1 measure tells us about the trade off between precision and recall values, so the results are not very satisfactory. This triggered us to include a new feature to the existing ones for our job.

Table 4.7: **Results for Experiment-1 by Randomized Cross Validation**

Data Split	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
90/10	0.587	0.92	0.72	0.983	0.031	0.059	0.802	0.54	0.64
80/20	0.58	0.92	0.71	0.96	0.024	0.046	0.79	0.51	0.616
70/30	0.535	0.955	0.686	0.796	0.0226	0.044	0.853	0.518	0.644

Table 4.8: **Results for Experiment-2 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.686431524	0.960726789	0.800532179
80/20	0.696110924	0.961715522	0.807513175
70/30	0.698304659	0.962792965	0.809378316

Table 4.9: **Results for Experiment-3 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.736386926	0.769007043	0.75192523
80/20	0.716512572	0.769070515	0.741536266
70/30	0.718042156	0.768878472	0.742461538

Table 4.10: **Results for Experiment-4 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.857240536	0.999014177	0.92269761
80/20	0.853048038	0.999355688	0.920381066
70/30	0.856941473	0.999547697	0.92275187

4.5.2.2 With Inclusion of Binary OR of New Features

Here, the new features are geographical place name or feature type obtained from GeoNames geographical database. We kept track of the frequency of these two features and ten used the binary OR of these to get our final feature.

Table 4.11: **Results for Experiment-1 by Hold-Out Validation**

Data Split	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
90/10	0.72	0.97	0.83	1	0.098	0.18	0.83	0.67	0.74
80/20	0.69	0.96	0.81	1	0.072	0.135	0.79	0.65	0.713
70/30	0.7	0.92	0.79	0.95	0.087	0.158	0.745	0.642	0.69

Table 4.12: **Results for Experiment-2 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.754253308	0.982758621	0.853475936
80/20	0.72519084	0.983182406	0.834706205
70/30	0.734774067	0.962264151	0.833271444

Table 4.13: **Results for Experiment-3 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.855384615	0.829850746	0.842424242
80/20	0.863560732	0.855024712	0.859271523
70/30	0.879708384	0.778494624	0.82601255

Table 4.14: **Results for Experiment-4 by Hold-Out Validation**

Data Split	Precision	Recall	F1
90/10	0.88538206	0.998127341	0.938380282
80/20	0.863521483	0.999025341	0.926344329
70/30	0.867838911	0.999346832	0.928961749

Analysis : For both the hold-out validation and randomized cross-validation methods implemented by us; the results show improvement than the previous condition. The F1 measures are comparatively on the better side. This triggered the need for the next step of actions.

Table 4.15: **Results for Experiment-1 by Randomized Cross Validation**

Data Split	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
90/10	0.66	0.937	0.77	0.95	0.048	0.092	0.814	0.664	0.73
80/20	0.647	0.94	0.765	0.9	0.034	0.066	0.81	0.64	0.71
70/30	0.62	0.95	0.75	0.92	0.048	0.091	0.81	0.68	0.74

Table 4.16: **Results for Experiment-2 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.717008155	0.967756233	0.823448956
80/20	0.722648379	0.964708275	0.82624523
70/30	0.712769465	0.965981867	0.820211839

Table 4.17: **Results for Experiment-3 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.801884706	0.816249667	0.808529383
80/20	0.820107142	0.814211915	0.817148867
70/30	0.82718983	0.804452382	0.814997658

Table 4.18: **Results for Experiment-4 by Randomized Cross Validation**

Data Split	Precision	Recall	F1
90/10	0.837994494	0.999024487	0.911397077
80/20	0.86917649	0.999675853	0.929856362
70/30	0.85177348	0.999633594	0.919796514

4.5.3 Detection of Geospatialness of Preposition

4.5.3.1 Expat at Sentence Level and Preposition Level

We tested our approach with a number of feature sets to get the most optimum results.The features used here are coded in the following table for the sake of our understanding. Experiments to employ features consisting of a binary value to record whether a place name or geo-feature were present and, separately, of a value that is the sum of the numbers of place names and geo-feature types, did not improve on sentence level performance and are not listed here.

Table 4.19: **Features used in the Experiments**

Kord	All features used for preposition sense detection in [7]
Kord-Geo	The features from Kord plus the number of placenames and the number of geographic feature types found in the head words of the preposition
Kord-Geo-S	The features from Kord plus the number of place names and the number of geographic feature types found within the entire sentence in which the preposition occurs
Kord-Geo-All	The features from Kord-Geo-S plus the sum of the numbers of place names and a binary value of true if either a place name or a geographic feature type is present
Geo-Baseline-S	The number of place names and the number of geographic feature types found within the entire sentence in which the preposition occurs

The results obtained from the classifiers are not grouped further on based on their precedence.

Table 4.20: **Results for 3-class classifier Predicting Geospatial, Spatial (but not Geospatial) or Non-spatial**

	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Kord	0.442	0.578	0.501	0.747	0.744	0.745	0.763	0.664	0.710
Kord-Geo	0.514	0.614	0.559	0.751	0.762	0.757	0.772	0.696	0.732
Kord-Geo-S	0.566	0.638	0.6	0.732	0.802	0.765	0.783	0.665	0.719
Kord-Geo-All	0.6	0.692	0.643	0.749	0.797	0.772	0.796	0.692	0.740

Analysis: In table 4.20, a multi-class Naive Bayes classifier was used to predict each of the three classes of geospatial, spatial (but not geospatial) and neither. There were several versions of the classifier that use different combinations of features (summarised in table 4.19). One of these (Kord) just uses the features from [7] described above. It resulted in an F1 value of 0.50 for Geospatial and better values of 0.745 for Spatial and 0.710 for Neither. This was extended by adding the two features of the number of place names and number of geographical features detected in the head words of the preposition that is being tested (Kord-Geo). Note that the head words are features generated by the parsing procedure used to generate the features used in [7]. They correspond to the subject and object of the preposition. A further variation (Kord-GeoS) records these latter numbers at the sentence level, which was found to improve upon the performance when only observing head words (though note that the quality of performance will depend upon the performance of the script to detect place names and geo-feature types). Experiments to employ features consisting of a binary value to record whether a place name or geo-feature were present and, separately, of a value that is the sum of the numbers of place names and geo-feature types, did not improve on sentence level performance and are not listed here. However, combining these latter data items with those in Kord-Geo-S did provide an improvement (referred to as feature set Kord-Geo-All) with an F1 for Geospatial of 0.643.

Table 4.21: **Results for Three 2-class Classifiers Predicting Geospatial, Spatial (but not Geospatial) and Non-spatial**

	Geospatial			Spatial			Non-spatial		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Kord	0.37	0.647	0.471	0.696	0.79	0.74	0.762	0.751	0.756
Kord-Geo	0.423	0.680	0.521	0.704	0.798	0.748	0.760	0.755	0.757
Kord-Geo-S	0.48	0.704	0.57	0.688	0.846	0.759	0.755	0.753	0.754
Kord-Geo-All	0.542	0.728	0.621	0.672	0.837	0.745	0.750	0.771	0.761
Geo-Baseline-S	0.625	0.419	0.502	0.494	0.889	0.635	0.422	0.326	0.368

Analysis: In addition to the three class classifiers we implemented several 2-class classifiers (see table 4.21 with target classes of geospatial (vs spatial or neither), spatial vs (geospatial or neither) and neither (vs geospatial or spatial). Just as with the 3-class classifiers we used either just Kordjamshidi features (Kord), and place name and geographic features from the preposition's head words (Kord-Geo) and from the whole sentence in which the preposition occurred (Kord-GeoS). We also tested the method using Kord-Geo-All features, which gave the best 2-class performance for geospatial sense with an F1 of 0.621 but this did not improve on the result from the 3-class classifier.

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

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Our work is mainly concentrated on developing a method for detecting the geospatial nature of prepositions in sentences. This was developed by using a machine learning approach that was developed in [7] for generic spatial role labelling.However, we followed a number of approaches and then finally came up with the latest one. Initially, we tweaked the mentioned machine learning approach for spatial role labelling then extended that approach to geospatial role labelling. Along the way, we kept on testing our approaches with different experiments. The classifiers were trained and tested by various types of datasets. Next, we thought of including new features along with the Kordjamshidi et. al. [7] features and that is where our work is wrapped up.

Using a corpus of sentences annotated as either geospatial,spatial (but not geospatial) or neither geospatial nor spatial, we found that, when trained on this corpus, the original method was not able to detect geospatial prepositions with an F1 value greater than 0.50. However, it detected the spatial (but not geospatial) class with F1 of 0.745 and it detected prepositions that are used with either a geospatial or a spatial sense with an F1 of 0.832. We have adapted the method in an effort to improve its performance for detecting geospatial sense by adding features (for machine learning) that record whether a place name or a geospatial feature type is present in the head words that serve as subject and object of the preposition or, alternatively, whether they are present in the entire sentence. Using the sentence level features provided better performance with an F1 of 0.643 for geospatial sense. It also resulted in an improvement in detection of the spatial (but not geospatial) class with an F1 of 0.772. It may be noted that a classifier using only the presence of a place name or geographic feature type in the sentence provided better performance than the basic spatial role labelling method.

5.2 Future Work

In future, we wish to investigate methods to make further improvements to the performance of the methods presented here. In particular, we will address a limitation of the current method with regard to detection of place names and feature types by using a richer gazetteer and extending the dictionary of geographical feature types. We will also consider the use of alternative approaches to classifying prepositions, for example through the use of word embeddings that take more account of the textual context of the prepositions.

The current work can also be extended to the task of hierarchical spatial role labelling. It is the case when one (or more) triple exits within another. Successful extraction of these triples and their respective trajector and landmark can be done by string matching algorithms by checking for containment.

Also, our work is based on the assumption that spatial indicators can be prepositions only. But it can be other parts of speech or even phrases. Simply put the spatial indicator may not be a single word always. We can get spatial information from a group of words too and it might be even more informative.

Another assumption in this case was that the trajector or the landmark were of single words mostly but in real life we often come across trajectors/landmarks with more than a word. We wish to deal with those also in our upcoming work.

For example, for the sentence, *The book is on the left corner of the table*. Here, the triple for the preposition *on* is <book, on , left corner of the table> and for the preposition *of* is <corner, of, the table>. So, this sentence consists of a hierarchy spatial roles also the landmark consists of multiple words. We wish to cover more ground on this in future.

REFERENCES

[1] Amazon mechanical turk. `https://www.mturk.com/`.

[2] Geonames gazetteer. `https://www.geonames.org/`.

[3] A. Herskovits. *Language and Spatial Cognition*. Cambridge University Press New York, NY,USA, 1987.

[4] R. Johansson and P. Nugues. Lth: Semantic structure extraction using nonprojective dependency trees. In *The Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230. Association for Computational Linguistics, 2007.

[5] J. H. Jurafsky, Daniel Martin. *Speech and Language Processing*, chapter 18. Pearson Education UK.

[6] A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *COMP@ SIGSPATIAL*, page 62, 2013.

[7] P. Kordjamshidi, M. Van Otterlo, and M. Marie-Francine. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36, December,2011.

[8] K. C. Litkowski and O. Hargraves. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[9] F. Liu. Automatic identification of locative expressions from informal text. `http://minerva-access.unimelb.edu.au/handle/11343/38520`. Masters by Coursework Shorter thesis, University of Melbourne, Melbourne, Australia, 2013.

[10] F. Liu, M. Vasardani, and T. Baldwin. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web, LocWeb '14*, pages 9–16, New York, NY,USA, 2014. ACM.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*, chapter 8, pages 164–166. Cambridge University Press, 2009.

[12] T. M. Mitchell. *Machine Learning*, chapter 6, pages 177–179. McGraw-Hill Science/Engineering/Math, 2013.

[13] T. M. Mitchell. *Machine Learning*, chapter 4, pages 111–112. McGraw-Hill Science/Engineering/Math, 2013.

[14] S. Russell. Expat. `https://github.com/shaun-russell/expat-nlp/tree/master/expat`.

[15] K. Stock, R. C. Pasley, Z. Gardner, P. Brindley, J. Morley, and C. Cialone. *Creating a Corpus of Geospatial Natural Language*, pages 279–298. Springer-Verlag New York, Inc., September,2013.

[16] S. University. Neural network dependency parser. `https://nlp.stanford.edu/software/nndep.html`).

[17] S. University. Stanford log-linear part-of-speech tagger. `https://nlp.stanford.edu/software/tagger.html`.

[18] J. O. Wallgrun, K. 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.

[19] S. J. Wolf, A. Henrich, and D. Blank. Characterization of toponym usages in texts. In *Proceeding'14 : Proceedings of the 8th Workshop on Geographic Information Retrieval*, Article No.7, Dallas, Texas, November 2014. ACM New York, NY, USA.