A Classification Schema for Fast Disambiguation of Spatial Prepositions

André Dittrich
Institute of Photogrammetry and Remote Sensing
Karlsruhe Institute of Technology
Germany
andre.dittrich@kit.edu

Maria Vasardani
Department of Infrastructure Engineering
The University of Melbourne
Australia
mvasardani@unimelb.edu.au

Stephan Winter
Department of Infrastructure Engineering
The University of Melbourne
Australia
winter@unimelb.edu.au

Timothy Baldwin
Department of Computing and Information Systems
The University of Melbourne
Australia
tb@ldwin.net

Fei Liu
Department of Computing and Information Systems
The University of Melbourne
Australia
fliu3@student.unimelb.edu.au

ABSTRACT

In the field of Artificial Intelligence the task of spatial language understanding is a particularly complex one. Textual spatial information is frequently represented by so-called locative expressions, incorporating spatial prepositions. However, apart from the spatial domain, these prepositions can occur in a wide range of senses (e.g., temporal, manner, cause, instrument) as well as in semantically transformed senses (e.g., metaphors and metonymies). Existing practical approaches usually disregard semantic transformations or falsely classify them as spatial, although they represent the majority of cases. For the efficient extraction of locative expressions from data streams (e.g. from social media sources), a fast filter mechanism for this non-spatial information is needed. Hence, we present a classification schema to quickly and robustly disambiguate spatial from non-spatial uses of prepositions. We conduct an inter-annotator agreement test to highlight the feasibility and comprehensibility of our schema based on examples sourced from a large social media corpus. We further identify the most promising existing natural language processing tools in order to combine machine learning features with fixed rules.

IWGS'15, November 03-06 2015, Bellevue, WA, USA

ACM 978-1-4503-3971-1/15/11 ...$15.00
DOI: http://dxdoi.org/10.1145/2833165.2833167.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Algorithms, Experimentation

Keywords

spatial prepositions, preposition sense disambiguation, data streams, semantic transformations, annotator agreement

1. INTRODUCTION

Unrestricted natural language understanding is a complex field drawing on different disciplines. The ability of a computer to recognize and interpret verbal spatial information, such as place descriptions (*I'm waiting in front of the train station*) or route descriptions (*turn right at the post office*), is a particular challenge that has recently attracted attention from a wide range of diverse research communities such as computer science, robotics, computer linguistics, spatial cognition and geographic information science. People use these kind of expressions in their daily life when making decisions; hence, enabling computers for spatial language understanding will be beneficial for human-computer interaction relating to everyday decision support, search, and similar generic information. In particular, the automatic extraction of spatial information from large textual data streams is an interesting and valuable field for possible applications.

These place or route descriptions use a variety of syntactic categories to encode spatial information, such as prepositions, adverbs, nouns and verbs, or any combination of them. In English as well as many other natural languages a very common means for people to express their spatial knowledge is *locative expressions* (LEs), as described in their prototypical form by Herskovits [13]. LEs are spatial expressions incorporating a preposition, its object, and the entity the prepositional phrase modifies, i.e. the subject (e.g., *the spider is on the wall*, cf. [13]). The object of the preposition(*the wall*, in our example) is called the **relatum**, ground, anchor or landmark, and the entity the prepositional phrase modifies (*the spider*, in our example) is called the **locatum**, figure, theme or trajectory in the various literature.

1.1 Research Gap

Applications which involve LEs in verbal interaction are still a major challenge for AI. These include, e.g., extracting spatial information from streaming data, robots following verbal wayfinding instructions, dialog-driven geo-services, emergency response/assistance systems, or querying search engines with spatial language. From a geographic information science point of view, LEs represent a so-far largely untapped resource of spatial information. In particular with the rapid advent of social media platforms in the last decade, the amount of potentially valuable spatial information increases literally every second.

To utilize this information in intelligent systems, it needs to be correctly identified, extracted and modeled in a suitable machine readable format first. This task represents an essential step in the complex processing pipeline from unstructured spatial information to structured spatial knowledge and its potential iconic representations on sketch-maps [37] or geo-referenced maps. However, spatial prepositions can occur in a multitude of different senses apart from their spatial usage. Although several approaches exist to automatically disambiguate prepositional senses [31, 38, 36, 5, 25] the disambiguation of spatial prepositions from their extended uses in metaphors, metonymies, idioms and related figures of speech (e.g., *the thought in the back of my mind*, cf. Section 2.1), generally named semantic transformations [10], has been so far largely disregarded.

In the spatial extension of the linguistic ontology GUM (Generalized Upper Model) by Bateman et al. [4], for example, the "[C]lauses with idiomatic or metaphorical uses of spatial terms were not considered" for the inter-annotator agreement. Khan et al. [15], however, identified these cases of semantic transformations to be the main reason for high false positive rates in spatial language understanding. Some advanced approaches for spatial information extraction from text, such as the `SpatialML` scheme from Mani et al. [28] have a related goal of identifying and annotating spatial information in text, but focus on geographical and culturally-relevant landmarks, i.e. named places.

In general approaches to sense disambiguation, semantically transformed cases are often classified as spatial, in contrast to temporal, manner and other common senses. In [36], the authors identify 32 distinct classes of prepositions, but still classify, e.g., metaphoric cases as spatial. Moreover, approaches to automatic spatial relation extraction often completely omit these cases [18, 19, 17, 39, 40, 35, 24, 28, 33, 34]. The main reason for omission often lies in the choice of corpus, for example by using corpora that are manually preselected to have a high rate of spatial language utterances (cf. [4]). However, in an unrestricted natural language environment, the number of expressions misleadingly identified as being spatial will be significant. This high rate of false positives has a direct negative impact on the accuracy, on the processing speed, and ultimately, on the feasibility of an automatic system.

Therefore, we are aiming to provide a robust, intelligible and fast classification schema to disambiguate spatial from non-spatial uses of prepositions that are generated by any kind of semantic transformation. To this end, we will identify the key linguistic features as well as external tools that will be beneficial for the automation of the process.

1.2 Fundamental Literature on Spatial Prepositions

Due to the richness of approaches to the topic, only selected views on spatial prepositions that influenced our classification schema are provided.

In semantic theory in general, two types of approaches can be identified: full specification and minimal specification. In full specification, every meaning of a word has a distinct representation in a lexicon. In minimal specification one meaning is seen as central, and from this all others are supposed to be derived by context or semantic transformations such as metaphor and metonomy (cf. [10]). In terms of research on the semantics of prepositions, the minimal specification view is also called the localist view. The localists argue for the spatial sense being the central meaning of prepositions. The first to promote this view was Leibniz, stating that prepositions "are all taken from space, distance and movement" [23], and many others followed (e.g., [29, 14, 21, 1]). Herskovits [14] argued for an ideal meaning, and called any divergence from it a sense-shift, but still as based on the spatial meaning. However, later theories claimed that there is more to prepositional meaning, for example the notion of a control relation as in *John is in a bad mood* (cf. [11]). Neuroscientist O'Keefe described the non-spatial relationships as higher dimensional axes additional to the first dimensions of space and time and called them metaphorical [32]. Coventry [1] supports this view but states that such extended uses are direct extensions of the spatial meaning of the terms rather than novel metaphorical uses.

For a classification schema, we are interested in a rather practical common sense separation in spatial and non-spatial uses of prepositions, for the sole purpose of increasing the accuracy of language parsers. Hence, we argue that for this work an existence of a core meaning is not essential. Dittrich et al. [8] showed that the claimed core meaning, i.e. the spatial meaning, is not necessarily the most frequent one. However, the deep linguistic and cognitive analysis of the cited work (and many more) represents the basis of the current understanding of prepositions, and informs the classification schema for a robust disambiguation.

The remainder of the paper is structured as follows. Section 2 describes the scope and explains consequential constraints for this research. The classification schema to exclude non-spatial prepositions is presented in Section 3, along with the identification of possible features for the implementation. The conducted experimental study is the topic of Section 4, and Section 5 discusses typical problems and errors as well as next steps and some long-term aspects. Section 6 concludes this paper.

2. RESEARCH SCOPE

This section describes the scope and theory of an approach to efficiently extract spatial information from unrestricted natural language. The investigated corpus (cf. Section 4.1) comprises texts from social media sources to represent a variety of modern language usage. The approach identifies prepositions used in a spatial sense in locative expressions. This section explains general constraints on syntax, types of objects as well as the resulting choice of prepositions for this study.

2.1 Spatial vs. Non-Spatial

The notion of spatial in our approach is locating in physical space. Hence, spatial prepositions in LEs can be defined as:

> describing the location (e.g., inclusion, proximity) or movement (e.g., origin, path, endpoint) of physical entities and events relative to other physical entities, actual places or locations.

The definition provides hints as to which locative expressions should be detected and which should be excluded in this practical approach. The involved entities in the scope should support the notion of identifiability as well as being physically anchored. In small-scale environments entities should be accessible for direct or indirect physical interaction, often involving changing their location in space. Following our definition above, some types of expressions should explicitly be excluded as non-spatial. These expressions can incorporate spatial prepositions, of which the sense has changed or shifted from their spatial meaning through semantic transformations. For humans, these non-spatial meanings are relatively easy to identify, due to their understanding of the involved entities and their world knowledge in general.

(1) The thought **in the back of** my mind.

In (1), it is understood that the relation *in the back of* is not a spatial one, i.e. the *thought* is not physically in the back of the *mind*. The locatum is an abstract entity and as such it cannot be assigned to an explicit location in physical space. It is also known that thoughts *being in one's mind* is just the way people conceptualize this abstract relation because of its resemblance to the general spatial relation of an object in a container. Although locational uses of prepositions are often implicitly spatio-temporal, we only consider the spatial aspect in our current approach.

2.2 Syntactic Constraints

Spatial relations can be encoded by other word categories such as adverbs (e.g. *here*, *downstairs*, *nearby*) and verbs (usually indicating a directed path, e.g., *to enter*, *to descend*, or implicitly describing a spatial arrangement, e.g., *to follow* or *to surround*), but we focus on the closed group of prepositions as indicators of spatial relations in this research. The reasons for this are as follows: (i) path-indicating verbs can in general be expressed by a simpler verb denoting movement (usually the manner) and a preposition providing the direction, such as to go *in(to)* instead of *to enter*; and (ii) adverbial terms lack an explicit relatum, and cannot usually be decoded without discourse or context knowledge. For example:

(2) I'm working **nearby**.

(2) is only fully understandable if a reference object is mentioned in the preceding discourse or is obvious in a specific situation (e.g., the listener knows the current location of the speaker and therefore can assume the reference object equals the current location). However, discourse analysis (co-reference resolution) spanning over the boundaries of a sentence is not yet considered in this work. In terms of syntax, an (optional) modifier (MOD), a preposition (P) and a complement (C) establish a prepositional phrase (PP). A PP denotes a single sentence constituent, which in general cannot be separated. For an extensive analysis of syntactic and semantic cases of LEs that go beyond the scope of this paper, see [20].

$$PP \Rightarrow (MOD) + P + C$$

It is important to distinguish between (transitive) prepositions and particles (= intransitive prepositions) in verb particle constructions, which do not take a complement (e.g., *He blacked out*) or can be moved to the right of the following noun phrase (NP) as in *turn off the light* and *turn the light off*. Hence, they are not constituents, as the NP is a direct object of the verb and not of the preposition. Verb particle constructions (VPCs) often form a semantic unit with the verb, where the particle does not carry its own semantic meaning and thus is not the head of a PP [3], as in (3).

(3) We looked up the answer.

However, in (4) the word *up* is in fact the head of the following NP and therefore a (transitive) preposition.

(4) We looked up the street.

Prepositions can also take different types of complements such as participial VPs (*John left before eating dessert.*), sentences (*He was nervous before the President called.*), NPs (*The book was placed on the table.*) or other PPs (*She jumped out from behind the tree.*). Only the latter two are of interest for this research because the former two do not usually describe spatial relations. It follows that the present approach exclusively studies transitive prepositions (i.e., taking an NP as complement) and complex PPs (i.e., taking one or more PPs as complement) where the last preposition has a NP complement. In these complex PPs, not every preposition will be recursively disambiguated, but rather assessed as one compound preposition that will get one class label. Thus, in *the function was called from inside of itself*, the compound preposition would be *from inside of*.

2.3 Choice of Prepositions

In this research, we study an extensive list of English prepositions that are typically considered to be potentially spatial. The source of this list is Dittrich et al. [8], who investigated more than sixty prepositions in the domain of short message services (Twitter). They compiled a comprehensive summary including a statistical probability for each preposition to encode a spatial relation based on their corpus. In this study their list is altered only slightly, i.e., *outwith* is excluded as being only used in Scottish English and *about*, *past* and *throughout* are added to test if they occur in spatial relations in the corpus used in the present study. Prepositions that, to the authors' best knowledge, do not (or not anymore) occur with a spatial sense in natural language, such as *after*, *as*, *because of*, *despite*, *during*, *for*, *in line with*, *in the face of*, *like*, *since*, *until*, *with*, and *without* are not investigated, i.e. they are directly excluded as non-spatial. We admit that *after* and *until* could in some cases be interpreted as temporal and spatial, however, the temporal aspect is the more prominent one. Additionally, there are a few English prepositions which can denote spatial relations but are archaic (e.g. *betwixt* and *nigh*) or domain specific (e.g. nautical terms such as *athwart* and *abeam*). They are also excluded because they are extremely rare in everyday language. The final list of prepositions considered in this study is presented in Table 1.

3. CLASSIFICATION SCHEMA

In this section we present a classification schema covering the scope of Section 2. Each rule is explained with examples, drawn from the corpus of this study wherever possible. Additionally, possible features and tools that can be exploited for the implementation of the schema are given for each presented rule, as well as general linguistic features (Section 3.4). In contrast to the manual classification, the identified features or external tools will rarely serve as a definite exclusion factor, but only as indicators with possible different weighting in the automatic classification algorithm. As input the approach assumes sentences that meet the requirements of the syntactic constraints explained in Section 2.2 and that contain at least one of the prepositions in Table 1. The pre-processing is described in Section 4.2.

3.1 Abstract Locatum or Abstract Relatum

3.1.1 Rule Explanation

If the locative expression has an abstract locatum **or** an abstract relatum it should be excluded as non-spatial, i.e. the locatum and relatum have to denote physical entities. Typical examples for abstract entities include emotions (5), ideologies (6), actions (7) and cognitive content (8).

(5) [I']m already in love with someone else...

(6) ...individuals can be drawn into world capitalism...

(7) I wish I was good at singing...

(8) ...we've settled into a pattern...

Terms that usually describe physical entities but are used in an abstract sense, here count as abstract entities and should be excluded as non-spatial as well.

(9) The author lures her reader into dark and dangerous territory.

The term *territory* in its most common sense can be described as a confined geographic area. In (9), however, it is used as an imaginary or abstract instance of its physical equivalent.

3.1.2 Identified Features

To determine whether an entity is abstract, external knowledge sources such as WordNet [9] and DBpedia are called upon. WordNet's sense frequencies will also be included as a-priori probabilities for each sense. Different existing algorithms for word sense disambiguation will be tested, in particular noun sense disambiguation, as additional support for a specific sense. Many of the examples that contain imaginary instances also often refer to the human body or body parts as relatum (*butterflies* ***in*** *my stomach*, *you* ***in*** *my heart*, etc.). This could also serve as an auxiliary hint of non-spatial usage, i.e. as an input feature to a machine learning approach.

3.2 Idioms

3.2.1 Rule Explanation

If the potential locative expression itself is in fact a (frozen) idiom or if it is used idiomatically, the example should be excluded as non-spatial. Idioms often contain potentially spatial prepositions, but utilize them in a non-spatial meaning.

(10) Peter **is over the hill**.

(11) She **felt under the weather.**

In (10), the preposition *over* does not imply that the subject (i.e., *Peter*) is located or is living over *the hill*, but rather that his career is over. (11) also does not locate an entity under *the weather*, but describes a status of being ill, sick or intoxicated.

3.2.2 Identified Features

The identification of specific indicators for an idiom is difficult, if not impossible to generalize into patterns. Therefore, different online dictionaries, thesauri and other digital linguistic resources of the English language will be employed to compile an extensive list of idioms. This list can be used for fuzzy matching of examples in question against all idioms and thus identify potential idioms and provide a score of how likely the identification is.

3.3 Others Missing Locative Purpose

The last group of non-spatial expressions incorporating a preposition is a rather heterogeneous one. Here, all non-spatial examples are subsumed that are not "caught" by the previous two classification steps.

3.3.1 Rule Explanation

If the preposition used in the example does not locate the locatum relative to the relatum, it should be excluded as non-spatial. Examples for this rule include, but are not limited to, prepositions that denote a temporal relation, the material of an object (12), the agent of an action (13), or the topic of some means of communication (14).

(12) The paint is **made from** resin.

(13) He was misunderstood **by** the customers.

(14) I read the paper **on** construction sites.

3.3.2 Identified Features

For the automatic exclusion of other examples missing locative purpose, the approach will mainly rely on two features. First, the dependency trees provided by an NLP parser such as the Stanford CoreNLP parser [16] and the Malt parser [30] will be used. These parsers can indicate, for example, when the very frequent preposition *by* is used to refer to an agent of an action. Secondly, as an additional indicator for the non-spatial usage of prepositions, typical collocations involving prepositions will be analyzed. In (12) the collocation of the verb *to make* plus the preposition *from* usually refers to the material of an object rather than a spatial relation. There are several other common verb + preposition combinations that indicate non-spatial usage such as *suffer from*, *focus on*, or *participate in*, quite often in the form of passive constructions. In contrast, in (14) the combination of the action *read* with the preposition

Table 1: Potentially spatial English prepositions

about	aside	beside	forth from	into	off	past	toward
above	at	between	from	left of	on	right of	under
across	atop	beyond	in	near	on top of	south of	underneath
against	back to	by	in (the) back of	next to	onto	southeast of	up
ahead of	before	close to	in (the) front of	north of	opposite (of)	southwest of	upon
along(side)	behind	down	in the middle of	northeast of	out (of)	through	via
amid(st)	below	east of	in the midst of	northwest of	outside (of)	throughout	west of
among(st)	beneath	far from	inside (of)	of	over	to	within

on and the generic relatum *construction sites*, indicates that the preposition describes the topic and not the location of the *paper*. This could be generalized to a non-spatial pattern *means of communication + on + generic noun phrase*.

3.4 General Features

Apart from the specific tools and features explained in the previous sections, a large set of general linguistic and syntactic features can be used in learning algorithms for the automatic disambiguation of spatial and non-spatial cases. These features can be grouped into three categories according to the word group they are related to.

3.4.1 Verb-related Features

- the main verb itself
- the verb stem
- the verb tense
- voice (active or passive)
- if the verb is a motion verb
- if the verb is a typical non-spatial verb
- the noun subject and direct object of the verb, e.g. *author* and *reader* in (9)

3.4.2 Noun-related Features (for locatum and relatum)

- if the noun is singular or plural
- if the noun can be identified as a toponym
- if the noun can be identified as a name of a person
- the accompanying adjective(s) of the noun, if existent
- the determiner of the noun, if existent, e.g. articles (definite or indefinite), demonstrative and possessive pronouns (*this*, *those*, etc. and *my*, *his*, etc.), quantifiers (*many*, *some*, *a lot*, *most*, etc.)

3.4.3 Preposition-related Features

- the a priori probability of the preposition being spatial, with a confidence interval (cf. [8])
- the modifier of the preposition, if existent

4. EXPERIMENTAL STUDY

This section introduces the corpus and the pre-processing steps of the actual annotation. Then, we detail the setup of our annotator agreement study, followed by the presentation of the results. Finally, experiences and results from a first proof-of-concept implementation conclude this section.

4.1 Corpus

In this study, a processed version of the mixed social media corpus is used that was originally compiled in [2] for the purpose of investigating and testing the common assumption of strong noisiness in textual data from social media sources. The original corpus comprised the following sources:

- **Twitter1 and Twitter2** — posts (tweets) from Twitter; 1M documents respectively
- **Comments** — user comments on Youtube; 874772 documents
- **Forums** — posts from the top-1000 valid vBulletin-based forums in the Big Boards forum ranking; 1M documents
- **Blogs** — blog posts from tier one of the ICWSM-2011 Spinn3r dataset [7]; 1M documents
- **Wikipedia** — text from an English Wikipedia dump; 200K documents
- **BNC** — all documents from the written part of the British National Corpus (BNC: [6]), a balanced corpus of British English used mainly as a point of comparison to the social media corpora; 3141 documents

The authors further restricted the corpus to English documents by applying automatic language identification (LANGID.PY: [27]). For a deeper description of the original corpus and the processing tools used see [2]. Liu et al. [26] further sampled the corpus down to a selection of 100K random sentences from each source. Additionally, they extracted 500 sentences from each source for their hand-annotation. These 3500 sentences in total depict the base corpus investigated in this paper for classifying spatial prepositions. The examples are in CoNLL format and include part-of-speech (POS) tags and chunk tags.

4.2 Pre-Processing

For the purpose of annotator agreement testing, the examples were pre-processed to comply with the syntactic rules described in Section 2.2.

1. Regular expressions to exclude sentences which do not contain at least one of the prepositions in Table 1.
2. Exploiting the POS tags to exclude sentences where the identified preposition is not tagged as `IN` or `TO`, i.e. not identified as a (transitive) prepositional use of the term (Penn Treebank Style)
3. Exploiting the chunk tags to exclude intransitive prepositions (i.e., not followed by a noun phrase)

The preprocessing resulted in 1265 examples. From these, a random subset of 500 examples was generated for the annotator agreement test.

Table 2: Example of sentence with compound preposition

ID	EXAMPLE	P1
1	The rabbit came FROM INSIDE OF the hole.	1

Table 3: Outcome of Annotations A and B

	A	B
true positives (TP)	159	149
false positives (FP)	25	17
false negatives (FN)	24	34
true negatives (TN)	596	604

4.3 Study Setup

The annotators were provided with the three classification steps (Section 3) plus compact explanations and examples for each rule. For the classification, we provided a spreadsheet that contained one example per row. The prepositions that needed to be classified were completely in upper case letters. The successive columns were reserved for the annotation. The class labels were **1** for spatial and **0** for non-spatial. An example could contain several prepositions. In these cases, the annotators were advised to use one column for each preposition for the classification. Additionally, the annotators were advised to always classify prepositions that are in upper case and directly following each other, as one compound preposition, as in Table 2. Due to the possibility of multiple prepositions in one example, these 500 examples contained 804 prepositions for classification.

4.4 Results

The manual classification was conducted independently by three annotators. One of the classifications was done by the first author, and will be referred to as the reference annotation (RA). The RA yielded a percentage of 22.8 % of spatial examples which illustrates the need for efficient filtering steps when processing high velocity data such as streaming data. The remaining classifications will subsequently be called Annotation A and B. In addition to the typical measure of precision, recall and F1-Score, we considered the following complementary measures: the negative predictive value (NPV), the specificity and the negative agreement. Thus, we account for the correct handling of negative examples. Finally, the value of Fleiss's Kappa is calculated to measure the agreement between all three classifications without considering the special status of the RA.

4.4.1 Agreement between Annotation A/B and RA

The results of Annotation A and B are displayed in comparison to the RA in Table 3.

Based on these values, the statistical measures can be calculated. These measures are summarized in Table 4.

4.4.2 Inter-Annotator Agreement

For the inter-annotator agreement (IAA), all three annotations were taken into account without any weighting or preferential treatment. Fleiss' Kappa (κ) was used, as it allows the calculations of the agreement between more than two annotators in case of nominal data, taking into account

Table 4: Statistical measures for the evaluation of the agreement of Annotation A and B vs. RA

Measure	A	B
Accuracy	0.939	0.938
Precision	0.864	0.898
Recall	0.869	0.814
F_1-Score	0.866	0.854
NPV	0.961	0.947
Specificity	0.960	0.973
Negative Agreement	0.961	0.960

Table 5: Outcome of the proof-of-concept test

	Test
true positives (TP)	36
false positives (FP)	13
false negatives (FN)	5
true negatives (TN)	193

the probability of agreement occurring by chance. This leads to a conservative estimation of the IAA. The value range is dependent on the number of annotators ($n = 3$) and the number of classes ($q = 2$) but not on the number of examples to annotate ($m = 804$), i.e. the possible value range in this setup is -0.5 to 1. The value of κ was computed to be 0.81 according to the formulae in [12]. According to the guidelines provided in [22], the presented result is at the lower limit of an "almost perfect agreement" which ranges from 0.81 to 1.0.

4.5 First Proof-of-Concept Implementation

To show the feasibility of our identified features and tools for the task of automatically disambiguating spatial from non-spatial uses of prepositions, a first proof-of-concept test was implemented. For the test, a random subset of 247 examples was chosen and the respective locata and relata were manually marked.

Based on the first rule described in Section 3.1, triplets were classified and analyzed according to whether the locatum and relatum of the triplet both denote physical entities. However, a pre-processing step of classifying typical non-spatial collocations of words as non-spatial proved highly valuable. The corpus analysis in this study and preceding studies facilitated the compilation of a comprehensive list of these typical non-spatial collocations, which contain elements such as *to expect from*, *to upgrade to*, *to focus on*, *to be good/bad at*, and *to be interested in*.

After automatically classifying examples as non-spatial based on this list, the remaining triplets are further investigated. For an estimate of how likely a certain term (i.e. the locatum and relatum of the triplet) refers to a physical entity, the test relies on WordNet.

For every sense of a term in question, the script determines the frequency count of the specific lemma that is identical to the term. The ratio of the accumulated frequencies for the lemmas denoting physical entities and the accumulated total frequencies is then used as the score for the term to refer to a physical entity. For those locata and relata which are not in the WordNet database the system relies on the

Table 6: Statistical measures for the evaluation of the agreement of the proof-of-concept test vs. RA

Measure	Test
Accuracy	0.927
Precision	0.735
Recall	0.878
F_1-Score	0.800
NPV	0.975
Specificity	0.937
Negative Agreement	0.955

tags of the STANFORD NAMED ENTITY TAGGER, if existent. Thus, entities that refer to persons or locations can be identified as physical entities. Moreover, the system heuristically assumes that most personal pronouns (subject and object form) denote persons as well, i.e., *I*, *me*, *you*, *he*, *him*, *she*, *her we*, *us*, *they*, *them* are regarded as physical entities.

For the actual classification, we currently use threshold-based rules. However, the preliminary results from automatic classification experiments (Table 5 and Table 6) are promising.

5. DISCUSSION

Despite the considerable agreement among all annotations (IAA) and the agreement of Annotations A and B with the RA, there were still some cases where the annotators analyzed the utterances differently. These cases were in fact quite often ambiguous concerning the actual triplet, especially concerning the context of the utterance. This section first identifies systematic or recurring cases within the false positives and false negatives, followed by an outlook on further research.

5.1 Problems

The misclassified examples can be divided into False Positives and False Negatives, which will be discussed in turn below.

5.1.1 False Positives

Two types of FP classifications occurred in the annotation experiment. The most common ones were the misinterpretations of actual abstract entities as physical ones (e.g. *project*, *demand*, *capital allocations*, *voice*, *cost*, *word*, etc.). Annotator A produced 21 errors of this type and Annotator B 12. However, often these examples included two entities that could arguably be taken as locatum. In (15) the physical entity (person) *you* and the abstract entity *difference* can be seen as being the subject of the preposition *on*. In a further study, this error source could be reduced by highlighting the complete potential locative expression (i.e. locatum + preposition + relatum) to the annotators instead of only the preposition. This means of course, that there might be cases where one prepositional phrase will have two locata that are possibly different types of entities (i.e. abstract or physical). Moreover, this issue still needs to be taken care of concerning the automatic extraction of the locatum as one of the most important disambiguation features.

(15) [...], I don't know if you'd notice a huge difference on the street.

5.1.2 False Negatives

The FN classifications showed two smaller groups of common misinterpretation sources but the majority of cases were very heterogeneous. The first group consists of cases where the annotators rejected a spatial example with a place as relatum. The rejected places often were either very large (16) or they were just not very common toponyms (or unfamiliar to the annotators (17)), and as such hard to classify as actual places. In general, Annotator A produced 4 errors of this type (total FN=24) and Annotator B 16 (total FN=34).

(16) it moves inside Mercury's orbit and [...]

(17) Chornovil , [...] in Lvov oblast [...]

The second group is degenerate locative expressions (DLEs, cf. [14]), i.e. locative expressions without explicit locatum. Especially in informal communication, people tend to disregard the subject of the preposition. The omitted locatum can often be assumed to be an implicit *I* (18) or implicit *you* (19) or at least a person. Still, DLEs are harder to classify with confidence. Errors caused by DLEs occurred 4 times for Annotator A and 3 times for Annotator B.

(18) arrived at home

(19) live at home anymore?

5.2 Outlook

The immediate next steps in this research will focus on implementing a parser that automatically identifies locative expressions, distinguishing them from non-literal uses of literal prepositions. The identified features will be further investigated and practically tested for their actual contribution to classification and speed. By means of the conducted annotator agreement experiment, we not only showed the appropriateness and comprehensibility of the classification schema, but also revealed complexities and possible error sources that will help us to implement the schema more efficiently. Moreover, we will test and compare different learning algorithms to determine what approach works best over our data.

We are aiming for a real practical solution to the problem of automatically disambiguating spatial from non-spatial uses of prepositions. Hence, we are not claiming that we conducted a holistic linguistic analysis of prepositional senses. As such, we recognize that in this approach, different aspects are disregarded such as certain meta-operators. These meta-operators include, for example, negations and aspects of existentiality, thus the examples (20) and (21) would be classified as spatial. One could of course argue that in (20) the actual relation that should be extracted is *outside_of(pencil,box)* rather than *in(pencil,box)*. However, taking the "inverse" preposition to handle negations cannot account for the many different spatial configurations a single preposition can describe (cf. [5, 10, 13, 14], inter alia) and would go beyond the goal of this research — i.e. parsing natural language and extracting spatial relations in a fast and robust way.

(20) The pencil is not in the box.

(21) The pencil was/might be/will be in the box.

Apart from these meta-operators, there are structural ambiguities that cannot be solved without explicit context knowledge. In (22), possible spatial relations are *in(dog,park)* and *in(man,park)*, but there is also the possible scenario that the dog is actually outside of the park looking in, seeing a man, thus making *in(dog, park)* an invalid spatial relation.

(22) The dog saw a man in the park.

Despite these complex cognitive/linguistic examples, our classification schema is a feasible approach for fast practical applications incorporating natural language locative expressions.

6. CONCLUSION

In this paper we presented a classification schema that allows for fast identification of locative expressions. The main contribution of our schema is its specific capability to recognize the most common metaphoric uses of prepositions, which we exclude as non-spatial. Together with other non-spatial cases, they represent the majority of uses and thus need to be filtered out, in particular for the efficient processing of textual streaming data. Nonetheless, they are usually disregarded in current approaches to natural language understanding, or they are explicitly assigned to the class of spatial prepositions. The manual classification conducted according to this schema shows a strong agreement between annotators and the reference annotation. This illustrates the fitness of our schema and its general comprehensibility. We also discussed typical cases of disagreement and pointed out advanced cognitive/linguistic aspects. Moreover, we identified the most promising ways to exploit existing tools for the implementation of the classification schema and a large set of linguistic indicators. Both the output of the tools as well as the linguistic indicators will be used as features for a machine learning approach, together with fixed rules.

7. ACKNOWLEDGMENTS

This is joint research between the Karlsruhe Institute of Technology and The University of Melbourne. The research stay abroad was funded by the Karlsruhe House of Young Scientists (KHYS).

8. REFERENCES

[1] A. Baddeley, V. B. H. L. Roediger, and J. R. Pomerantz, editors. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions.* Taylor & Francis, Hove, UK, 2004.

[2] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, 2013.

[3] 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.

[4] J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071, 2010.

[5] C. Boonthum, S. Toida, and I. Levinstein. Preposition senses: Generalized disambiguation model. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3878 of *Lecture Notes in Computer Science*, pages 196–207. Springer Berlin Heidelberg, 2006.

[6] L. Burnard. User guide for the British National Corpus, 1995.

[7] K. Burton, N. Kasch, and I. Soboroff. The ICWSM 2011 Spinn3r dataset. In *Proceedings of the Fifth Annual Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, Spain, 2011.

[8] A. Dittrich, D. Richter, and C. Lucas. Analysing the usage of spatial prepositions in short messages. In G. Gartner and H. Huang, editors, *Progress in Location-Based Services 2014*, Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Berlin/Heidelberg, Germany, 2015.

[9] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, USA, 1998.

[10] P. Gärdenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces.* The MIT Press, 2014.

[11] S. C. Garrod and A. J. Sanford. Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, 6:147–160, 1989.

[12] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[13] A. Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9:341–378, 1985.

[14] A. Herskovits. *Language and Spatial Cognition: An interdisciplinary study of the prepositions in English.* Cambridge University Press, Cambridge UK, 1986.

[15] A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, COMP '13, pages 62:62–62:69, New York, NY, USA, 2013. ACM.

[16] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Sapporo, Japan, 2003.

[17] P. Kordjamshidi, P. Frasconi, M. Van Otterlo, M.-F. Moens, and L. De Raedt. Relational learning for spatial relation extraction from natural language. In S. Muggleton, A. Tamaddoni-Nezhad, and F. Lisi, editors, *Inductive Logic Programming*, volume 7207 of *Lecture Notes in Computer Science*, pages 204–220. Berlin/Heidelberg, Germany, 2012.

[18] P. Kordjamshidi, M. Van Otterlo, and M.-F. Moens. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

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

[20] M. Kracht. On the semantics of locatives. *Linguistics and Philosophy*, 25(2):157–232, 2002.

[21] B. Landau and R. Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265, 1993.

[22] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

[23] G. W. Leibniz. *Nouveaux essais sur l'entendement humain*. v. R. E. Raspe, Amsterdam, Netherlands, 1765.

[24] H. Li, T. Zhao, S. Li, and Y. Han. The extraction of spatial relationships from text based on hybrid method. In *Information Acquisition, 2006 IEEE International Conference on*, pages 284–289, 2006.

[25] 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.

[26] 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, Shanghai, China, 2014. ACM.

[27] M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 25–30, Jeju, Republic of Korea, 2012.

[28] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. SpatialML: Annotation scheme, corpora, and tools. Technical report, The MITRE Corporation, 2008.

[29] G. A. Miller and P. Johnson-Laird. *Language and Perception*. Harvard University Press, Harvard, USA, 1976.

[30] J. Nivre, J. Hall, and J. Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2216–2219, Genoa, Italy, 2006.

[31] T. O'Hara and J. Wiebe. Preposition semantic classification via Penn Treebank and FrameNet. In *Proceedings of the Seventh Conference on Natural Language Learning*, CONLL '03, pages 79–86, Edmonton, Canada, 2003. Association for Computational Linguistics.

[32] J. O'Keefe. The spatial prepositions in English, vector grammar, and the cognitive map theory. In P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett, editors, *Language and Space*, chapter 7, pages 277–316. MIT Press, Cambridge, MA, USA, 1996.

[33] J. Pustejovsky, J. Moszkowicz, and M. Verhagen. ISO-Space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 1–9, Oxford, UK, 2011.

[34] J. Pustejovsky, J. Moszkowicz, and M. Verhagen. A linguistically grounded annotation language for spatial information. *Revue TAL–Traitement Automatique des Langues*, 53(2):87–113, 2013.

[35] Q. Shen, X. Zhang, and W. Jiang. Annotation of spatial relations in natural language. In *Proceedings of the International Conference on Environmental Science and Information Application Technology, 2009*, volume 3, pages 418–421, Wuhan, China, July 2009.

[36] V. Srikumar and D. Roth. An inventory of preposition relations. *CoRR*, abs/1305.5785, 2013.

[37] M. Vasardani, S. Timpf, S. Winter, and M. Tomko. From descriptions to depictions: A conceptual framework. In T. Tenbrink, J. Stell, A. Galton, and Z. Wood, editors, *Spatial Information Theory*, volume 8116 of *Lecture Notes in Computer Science*, pages 299–319. Springer, 2013.

[38] D. Villanueva, A.-D. Cuevas-Rasgado, O. Juárez, and A. Guzmán-Arenas. Using frames to disambiguate prepositions. *Expert Systems with Applications*, 40(2):598–610, 2013.

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

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