Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language

PARISA KORDJAMSHIDI,
MARTIJN VAN OTTERLO
and
MARIE-FRANCINE MOENS
Katholieke Universiteit Leuven

This article reports on the novel task of *spatial role labeling* in natural language text. It proposes machine learning methods to extract spatial roles and their relations. This work experiments with both a step-wise approach, where spatial prepositions are found and the related trajectors and landmarks are then extracted, and a joint learning approach, where a spatial relation and its composing indicator, trajector and landmark are classified collectively. Context-dependent learning techniques, such as a skip-chain conditional random field, yield good results on the GUM evaluation data (Maptask) data and CLEF-IAPR TC-12 Image Benchmark. An extensive error analysis, including feature assessment, and a cross-domain evaluation pinpoint the main bottlenecks and avenues for future research.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Parsing and understanding*; *Text analysis*; H.3.1 [**Information Storage and Retrieval**]: content analysis and indexing—*Linguistic processing*

General Terms: Experimentations, Languages

Additional Key Words and Phrases: Semantic labeling, Spatial relations, Spatial information extraction

1. INTRODUCTION

An essential function of language is to convey *spatial relationships* between objects and their relative/absolute location in a space. The sentence *"Give me the gray book on the large table."* expresses information about the *spatial configuration* of two objects (book, table) in some *space*. Understanding such spatial utterances is a problem in many areas, including robotics, navigation, traffic management, and query answering systems [Tappan 2004]. Although the current work focuses on natural language processing, our long-term research considers spatial information extraction in a multimodal environment and aims to obtain and represent spatial relations using formal representations, allowing further spatial reasoning. For example, an interesting multimodal environment is the navigation domain, where we expect a robot to follow navigation instructions [Kollar et al. 2010]. When a camera is placed on the robot, it should be able to both recognize objects and their location and search for particular items based on verbal instruction. Another example is answering queries about objects' locations using both textual descriptions and visual data; combining the evidence provided by recognizing objects in the texts and images could generate answers that are more reliable. Spatial information extraction from language could also play an important role in semantic search, i.e., extracting information based on meaningful categories.

We recently introduced *spatial role labeling* problem as the extraction of generic spatial semantics from natural language [Kordjamshidi et al. 2010b]. We defined a semantic labeling scheme to annotate spatial information. It tags natural language with the spatial roles carried by words according to the *holistic spatial semantic theory* (HSS) [Zlatevl 2007]. The core problem of spatial role labeling is assigning specific tags to words or phrases in natural language sentences to express their roles in terms of spatial semantics. For example, in *"John is sitting on the ground"*, the preposition *"on"* is an indicator of a *spatial relation* between *"John"* and *"the ground"*. Many prepositions never carry a spatial meaning, whereas some have spatial *sense* depending on the *context*. The preposition *"on"* in this sentence has spatial sense, though it has no such sense in the sentence *"I can count on him."* *"John"* is the first argument of the *on*-relation and is a *trajector*. The phrase *"the ground"* is the second argument of the *on*-relation and is a *landmark*. In the related research in this domain, restricted languages extract very specific and application-dependent relations from text [Kelleher 2003; Tappan 2004; Li et al. 2007]. Previous research has not systematically covered spatial relation and role extraction from *unrestricted* natural language with machine learning methods, but this paper aims to do so. Statistical machine learning models are promising approaches to address the intrinsically ambiguous nature of spatial information in natural language.

A major obstacle when dealing with unrestricted language is the scarcity of annotated data available for training machine learning models. We therefore start with the available resources. In our leading experiments, we learn prepositions' spatial senses by exploiting annotated data from the preposition project (TPP) employed in SemEval-2007 [Litkowski and Hargraves 2007] and then use the results of preposition disambiguation in a *spatial role labeler* that identifies *trajector* and *landmark* roles. We use linguistically motivated features and evaluate several context-dependent classification algorithms. We successfully evaluate spatial role labeling on texts from the GUM (General Upper Model spatial ontology) evaluation data [Bateman et al. 2007] and CLEF IAPR TC-12 Image Benchmark data [Grubinger et al. 2006].[1]

One advantage of our pipelining approach is that knowledge from another linguistic resource is *injected* into the learning system. The TPP data are exploited here to solve the first part of our relation extraction algorithm, i.e., finding prepositions that have a spatial sense. We use annotated data from a larger source outside our training and test data in the extraction task, potentially increasing generalization possibilities. Errors concerning incorrectly recognizing prepositions' spatial meaning can propagate and lead to incorrect recognition of spatial roles and relationships. Thus, the pipelined approach has difficulties competing with models that jointly learn the spatial meaning of a preposition and corresponding spatial roles of its arguments. Analyzing and comparing these settings provide inspiration for utilizing (other) resources for our task.

We present the first experimental study on learning to extract spatial information from unrestricted natural language. Our main **contributions** include the following:

—We introduce the novel *spatial role labeling* task, which extracts spatial relations from natural language.

[1] See also http://imageclef.org/photodata

—We present the first domain-independent English dataset with labeled data for spatial expressions, specifically designed for machine learning solutions.

—Based on linguistically oriented features, we evaluate *conditional random field* (CRF) algorithms and compare their suitability for the task.

—We demonstrate the injection of external data resources into the spatial role labeling task by exploiting sense-annotated prepositions from TPP and compare it to a one-step approach, limited to only using spatially annotated data.

—We provide extensive experiments to show that our approach produces good results for the spatial role labeling task.

—We extensively survey related approaches for spatial language understanding in cognitive science, linguistics and computer science.

—We pinpoint bottlenecks and outline future research directions.

Main structure of this article. This paper is structured as follows. In Section 2, we describe the spatial role labeling task and formally define it in Section 3.

In Section 4, we describe our approach, based on machine learning techniques, to learn the spatial role labeling task from an annotated dataset. This approach solves two main subproblems for which solutions are described subsequently. The first subproblem is identifying the *pivot* of spatial relations, for which we learn to predict prepositions' roles more specifically, as described in Section 4.1. The second subproblem is identifying possible *arguments* of the spatial relations, for which we learn to predict whether parts of a sentence can be classified as so-called *trajectors* or *landmarks*, as described in Section 4.2. Both subproblems tackle the overall goal of extracting spatial *relations* from text. In Section 4.3, we investigate another setting in which we classify all roles jointly, i.e., without separate classifications for spatial indicators and trajectors/landmarks. Section 4.4 reports which algorithms, based on probabilistic graphical models, are employed by both subproblems.

In Section 5, we present and discuss a series of experiments. After introducing the main structure and rationale of the experiments, we show results for several datasets and perform an additional feature analysis. We give results in quantitative form but also present a qualitative analysis to show the effectiveness of the approachs. To complement the error analysis and see how well the learned classifiers generalize new data, we evaluate them on several texts from different subject domains than the training domain. After the experiments, in Section 6, we discuss related lines of research on spatial information representation and extraction in cognitive science, linguistics and machine learning. Section 7 concludes this article and outlines prominent research directions in spatial language processing.

2. THE SPATIAL ROLE LABELING TASK

As discussed above, spatial information plays an important role in many applications [Galton 2009]. However, its automatic recognition in natural language expressions is undeveloped or, when addressed, limited to recognizing coarse-grained and brittle information added to predicates and mainly expressed by verbs.

To highlight some general aspects of spatial semantics, consider the following two sentences (taken from [Bateman et al. 2010]):

(1) *He left the institute an hour ago.*
(2) *He left the institute a year ago.*

In the first example the sentence semantics indicate that the person is no longer in the building, and the sentence is about physically leaving the building and going somewhere else. This change directly amounts to a physical and spatial relocation. The second sentence expresses a more fundamental change: the person has apparently quit his job at the institute. The second type of spatial change is more involved and less material. Another set of examples is as follows:

(3) *The computer is on the table and the mouse is to the left of it.*
(4) *The party leader could be considered at the far left of the political spectrum.*

The first sentence expresses two explicit physical relations about objects on a table. The second sentence uses a similar relation *"at the far left of"*, but its meaning is more conceptual. Only drawing this "political spectrum" on a piece of paper allows one to put the party leader on its left side.

These examples illustrate some of the challenges in spatial language understanding. Similar lexical items can provide different spatial meanings. Conversely, two different descriptions may have a similar semantic interpretation:

(5) *Looking over his right shoulder, he saw his dog sitting quietly.*
(6) *The dog sat quietly on the floor to his right.*

In the sentences (1) and (2), the spatial information is mainly expressed through a verb, whereas the other examples primarily use prepositions. Furthermore, some information is not explicitly represented in the words but can be *inferred* from common sense. For example, one can infer that the mouse is on the table in sentence (3). This sentence also includes a related inference step resolvable at the linguistic level. An *anaphora resolution* step attaches *it* to *the computer* before determining the spatial semantics. *It* could refer to *the table*, in which case the spatial semantics also differ.

Despite the variations in spatial information in natural language expressions, a sentence can essentially express *spatial relations* between objects. For example, the third sentence contains an *on*-relation between the computer and the table. Another relation is that the mouse is *to the left of* the computer. Such relations, denoted *on(computer,table)* and *toTheLeftOf(mouse,computer)*, form the starting point of any system that processes spatial information in natural language. In *on(computer,table)*, we can distinguish the different *spatial roles* of phrases in a sentence: *on* expresses a predicate (or, relation) and *computer* and *table* are arguments with their own roles. Our main concern in this article is *extracting* such *spatial relations*.

We define **spatial role labeling** as *the automatic labeling of words or phrases in sentences with a set of spatial roles*. The roles take part in one or more *spatial relations* expressed by the sentence. The sentence-level spatial analysis of texts characterizes spatial descriptions, such as determining the objects' spatial properties and locations to answer *"what/who"* and *"where"* questions. The *spatial indicator* (typically a preposition) establishes the type of spatial relation, and other constituents express the participants of the spatial relation (e.g., entities' locations). The following sentence is an example:

$$\textit{Give me the } [gray\ \ book]_{tr}\, [on]_{si}\, [the\ \ big\ \ table]_{lm}.$$

Our spatial role set consists of *trajector* (tr), landmark (lm) and *spatial indicator* (si) (and *none* otherwise) [Kelleher 2003; Zlatevl 2007; Kordjamshidi et al. 2010b]. The above sentence contains several subsequences labeled with these roles. They are as follows:

—**Trajector**: the entity whose (trans)location is of relevance. The *book* is the main entity of which location is specified in the sentence. The trajector can be static or dynamic, a person or an object, or even a whole event. Alternative terms used in the literature are *local/figure object*, *locatum*, *referent* or *target*.

—**Landmark**: the reference entity in relation to which the location or trajectory of the trajector's motion is specified. The main entity's location designator (the trajector, the *book*) is the *table*. Other terms for landmarks are *reference object*, *ground*, or *relatum*.

—**Spatial indicator**: the tokens that define constraints on the spatial properties, such as the trajector's location with respect to the landmark (e.g., *in*, *on*). A spatial indicator expresses a relation (or predicate) with the landmark and trajector as its arguments. Spatial indicators explain the types of spatial relations and are often prepositions but can also be verbs and nouns among other parts of speech. These indicators are the **pivot** of spatial relations.

Other conceptual aspects, such as **motion indicators**, indicate specific spatial motion information (usually specified in terms of *verbs*); **frame of reference** and the **path** of a motion are influencing concepts for spatial semantics and roles [Zlatevl 2007]. However, we restrict our focus to prepositions conveying spatial information.

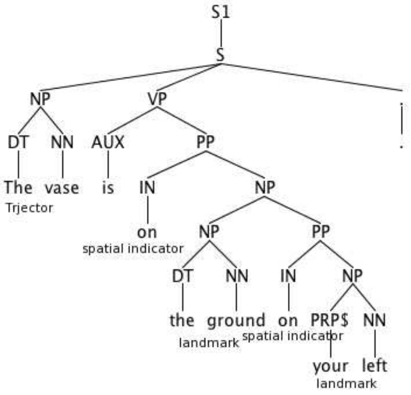

Fig. 1. Parse tree labeled with spatial roles.

Spatial role labeling is a special type of *semantic role labeling*, and, as with semantic roles, the spatial relations supported by the roles contribute to a sentence's semantic frame recognition [Màrquez et al. 2008]. In semantic frame labeling, a predicate is identified and disambiguated, and its role arguments are recognized. In spatial role labeling, the spatial indicator is identified (instead of the verb predicate) and disambiguated, and its semantic role arguments including the trajector and landmark, are found.

However, differences between these two tasks exist. In spatial role labeling, the roles are more specific regarding their semantics; there is no direct correspondence between the sentence's semantic structure based on traditional semantic frames (*patient*, *agent*) and the spatial semantics' structure. In the above example, FrameNet's *"Giving"* frame provides the **semantic type** *Locative_relation; the **Place** where the **Donor** gives the **Theme** to the **Recipient**.* The location refers to the place where the *give* is performed, and not the location of the book, mentioned in the prepositional phrases. Moreover, both the formal and informal (pragmatic) spatial expression meanings in natural language are highly dependent on lexical details, the ontological structure of spatial information spaces, and the embedding of extracted information into existing spatial knowledge.

Another difference between spatial role labeling and semantic role labeling is that no large annotated corpora were available from which spatial roles could be learned directly. New data resources were needed to apply machine learning techniques. In this respect, breaking the problem into parts and utilizing existing linguistic resources have the advantage of limiting the training examples that must be labeled. These external resources could improve the performance of the spatial role labeling task, which is evaluated in this paper.

General spatial relation extraction presents many challenges concerning task-specific ambiguities and difficulties. However, there is not always a direct mapping between a sentence's grammatical structure and its spatial semantic structure. This issue is more challenging in complex spatial expressions that convey several spatial relations. The simple example below shows that grammatical dependencies cannot always identify spatial dependencies and connections:

The vase is **on** *the ground* **on** *your left.*

The dependency tree relates the first appearance of *"on"* to the words *"vase"* and *"ground"*. This process produces a valid spatial relation connecting the right trajector to the right landmarks. If we systematically follow the grammatical clues and information, then the second appearance of *"on"* connects the *"ground"* and *"your left"*, producing a less meaningful spatial relation in terms of trajector, landmark and spatial indicator (*"ground on your left"*), Figure 1 shows the related parse tree. When confronted with more complex relations and nested noun phrases, deriving "spatially valid" relations is not straightforward and highly dependent on the *lexical meaning* of words. However, recognizing the right prepositional phrase (PP) attachment during syntactic parsing can improve the identification of spatial arguments.

Other linguistic phenomena, such as *spatial-focus-shift* and *ellipsis of trajector and landmark* [Li et al. 2007], make extraction more difficult. *Spatial motion detection* and *recognition of the frame of reference* are additional challenges that are not treated here.

3. PROBLEM DEFINITION

The spatial role labeling task finds *spatial relations* in natural language sentences, each of which includes a *spatial indicator* and its *arguments*. We assume that the sentence is a priori partitioned into a number of segments. The segments could be words, phrases or arbitrary subsequences of the sentence. More formally, let S be a sentence defined as a sequence of N segments:

$$S = \langle w_1, w_2, \ldots, w_N \rangle$$

We define a set of roles: $\mathrm{roles} = \{\mathrm{trajector}, \mathrm{landmark}, \mathrm{spatial\_indicator}, \mathrm{none}\}$, and each segment in the sentence can be assigned one or more of these roles. Each **spatial relation** in sentence S is a triple

$$\Big\langle\ w_{\mathrm{spatial\_indicator}},\ \ w_{\mathrm{trajector}},\ \ w_{\mathrm{landmark}}\ \Big\rangle$$

where $w_{\mathrm{spatial\_indicator}}$, $w_{\mathrm{trajector}}$ and w_{landmark} are three distinct segments of S, denoting the parts of S that represent the *spatial indicator* and its *trajector* and *landmark* arguments, respectively. For any spatial relation, the value of the trajector (or landmark) can be "undefined", meaning that no segment in S represents the trajector (or landmark). In those cases, we call the trajector (or landmark) *implicit*, as in the sentence *"Come over here"*, where the trajector *"you"* is only implicitly present.

Given a sentence S, the set of all **spatial indicators** of S is denoted I. It is induced by the indicator function I defined over all segments w of S:

$$I(w) = \begin{cases} 1, & \text{if } w \text{ is a spatial indicator} \\ 0, & \text{otherwise} \end{cases}$$

We assume that spatial indicators overlap with neither each other nor trajectors and landmarks. In other words, for any sentence S, if w and w' are two segments of S, then $I(w) = 1$ and $I(w') = 1$ imply that $w \cap w' = \emptyset$. Because trajectors and landmarks are spatial indicator *arguments*, we define two indicator functions relative to a given spatial indicator s in sentence S. The set of **trajectors** (**landmarks**) with respect to spatial indicator s is denoted T_s (L_s), induced by indicator functions T_s and L_s defined over all segments in S. For a spatial indicator s, its trajector and landmark cannot overlap with each other or s itself (though they can be undefined, as mentioned earlier).

Although we have defined spatial indicators, trajectors and landmarks as arbitrary segments of a sentence, we focus on single words, each as one segment. However, a phrase in the sentence commonly plays a role, and we thus assume that the *head word* of the phrase is the role-holder. A head word determines its phrase's syntactic type; analogously, it is a stem that determines the semantic category of its component's compound. The other elements of a phrase modify the head. For example, in *”the huge blue book”*, *”book”* is the head word, and *”huge”* and *”blue”* are modifiers. In our data, the labeling scheme reflects this fact and only assigns roles to head words and labels the remaining words (e.g., modifiers) as “none”. Hence, a sentence is hereafter assumed to be a sequence of words.

Our ground-truth data include sequences, each of which contains exactly one (labeled) spatial indicator with all possible trajectors and landmarks. A sentence can thus provide multiple examples, up to the number of its contained spatial indicators. We formally define each sentence in the corpus as a sequence of words $\langle w_1, \ldots, w_n \rangle$. Let k be the number of prepositions in a sentence s; s then induces k examples $e_1 \ldots e_k$, where examples e_i and e_j have the same spatial indicator for no i and j. Each e_i ($i = 1 \ldots k$) is a sequence $\langle (w_1, l_1), \ldots, (w_n, l_n) \rangle$ in which each word w_i ($i = 1 \ldots n$) is tagged such that i) at most, one w_j gets a label $l_j = \mathrm{spatial\_indicator}$; ii) some words get a label $\mathrm{trajector}$ or $\mathrm{landmark}$, if they are a trajector or landmark of the spatial indicator w_j; and iii) the remaining words get a label none. If a preposition is not spatial, all words in the example are tagged with none. As an illustration, consider the following sentence, which gives two examples:

A	girl	and	a	boy	are	sitting	at
none	**trajector**	**none**	**none**	**trajector**	**none**	**none**	**sp.indicator**
none	**none**	**none**	**none**	**none**	**none**	**none**	**none**

the	desk	in	the	classroom.
none	**landmark**	**none**	**none**	**none**
none	**trajector**	**sp.indicator**	**none**	**landmark**

The sentence is labeled twice, each time with a different indicator. Using our indicator functions, we have

$$I = \{\text{at,in}\} \qquad T_{\text{at}} = \{\text{girl,boy}\} \qquad \text{and} \qquad L_{\text{at}} = \{\text{desk}\}$$
$$T_{\text{in}} = \{\text{desk}\} \qquad \text{and} \qquad L_{\text{in}} = \{\text{classroom}\}$$

The spatial relations for this sentence are the triples produced by the following (we only account for head words in the role-playing phrases):

$$\{\text{at}\} \times \{\text{girl}, \text{boy}\} \times \{\text{desk}\} = \big\{\langle \text{at}, \text{girl}, \text{desk}\rangle, \langle \text{at}, \text{boy}, \text{desk}\rangle\big\}$$
$$\{\text{in}\} \times \{\text{desk}\} \times \{\text{classroom}\} = \big\{\langle \text{in}, \text{desk}, \text{classroom}\rangle\big\}$$

An example with an implicit trajector is the following sentence:

Go	under	the	bridge
none	**spatial_indicator**	**none**	**landmark**

In this case, we derive the spatial relation using

$$I = \{\text{under}\} \qquad \text{and} \qquad T_{\text{under}} = \emptyset \qquad \text{and} \qquad L_{\text{under}} = \{\text{bridge}\}$$

which results in $\langle \text{under}, \text{undefined}, \text{bridge}\rangle$ as the corresponding spatial relation.

This article takes a given corpus of sentences tagged with spatial indicators, trajectors and landmarks, giving a multitude of sequence examples, and constructs (i.e., learns) an automated spatial relation extraction method that can be employed successfully on unseen data.

4. APPROACH

The problem definition leads to a similar problem as semantic role labeling (SRL), where words are classified based on a known *predicate* (a verb). In spatial role labeling, the spatial indicator is the pivot (i.e., predicate) of the spatial relation. A spatial indicator can be from various lexical word classes, although the most dominant form is the *preposition*. In SRL, one can start from a verb and find roles related to it, but in spatial role labeling, one must first find the *sense* of the pivot (i.e., the preposition). Sometimes, a proposition has a *spatial sense*, but that same preposition might not have a spatial sense in a different context.

In our approach, the set of roles is $\{\mathrm{trajector}, \mathrm{landmark}, \mathrm{spatial\_indicator}, \mathrm{none}\}$, and we use an additional term $\mathrm{undefined}$ to highlight the existence of *implicit* trajectors or landmarks; $\mathrm{undefined}$ does not appear in the annotated data, nor is it learned or predicted by our classifiers. It solely serves as a place-holder for missing elements if the three components of a spatial relation cannot be explicitly found in a sentence (Algorithm 1 provides further explanation). The set of all spatial relations in a sentence S, denoted SR, is defined thus (where s, t, l are head words in S):

$$\mathrm{SR} = \big\{\langle w, w', w''\rangle \mid w \in I,\ w' \in T_w,\ w'' \in L_w\big\}$$

In this definition, three functions should be estimated. First, the function I is needed; it takes a word in the sentence as an input and estimates whether it is a spatial indicator. We employ a general *probabilistic* classifier; for spatial indicators, we learn a function $\hat{I}$ representing the probability that a word is spatial, given some features about sentence S. To get the (deterministic) indicator function I, we compute (using $r = \{\mathrm{spatial}, \mathrm{nonspatial}\}$)

$$I(w) = \begin{cases} 1, & \text{if } \mathrm{spatial} = \arg\max_{x \in r} \hat{I}(x \mid w, f(w, S)) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

optimized over training data, where $f(w, S)$ denotes a set of features derived from sentence S and word w.

Indicating which words in the sentence have the trajector or landmark role requires two other functions, *given that we know that some word s is a spatial indicator*. Because the parameters for both trajectors and landmarks are the same (i.e., the spatial indicator), we can combine them into a multi-class classification problem that classifies words in a sentence (i.e., head words) into $r' = \{\mathrm{trajector}, \mathrm{landmark}, \mathrm{none}\}$. We call this function $\hat{R}$, and it takes a spatial indicator and tags words with these roles. We use a probabilistic classifier here, and to obtain deterministic classifications for landmarks and trajectors, we first compute

$$r_{w,s} = \arg\max_{x \in r'} \hat{R}(x \mid w, s, f(w, s, S)) \tag{2}$$

where w is a word in sentence S, s is a spatial indicator, $f(w, s, S)$ denotes a set of features defined over the word w, the spatial indicator s, and the sentence S. This process maximizes a probability function given a set of features. The details of this function are described in the next section. We continue with

$$L_s(w) = \begin{cases} 1, & \text{if } r_{w,s} = \mathrm{landmark} \\ 0, & \text{otherwise} \end{cases} \qquad T_s(w) = \begin{cases} 1, & \text{if } r_{w,s} = \mathrm{trajector} \\ 0, & \text{otherwise} \end{cases}$$

From Equations 1 and 2, we see that a natural pipelined task decomposition presents itself. We can first find words that potentially carry a spatial sense ($I(s) = 1$), and we then find the corresponding trajectors and landmarks for each pivot.

The general structure of our pipeline approach consists of the following steps, outlined in subsequent sections:

—**Finding spatial indicators:** The first task consists of labeling parts of an input sentence S that play the spatial pivot role or finding the preposition with spatial sense. Section 4.1 describes this step, which utilizes TPP data to learn the labeling task. As we see below, we reduce this step to finding potential spatial indicators by only considering a sentence's prepositions.

—**Finding spatial arguments:** The second task consists of classifying parts of an input sentence S that play the landmark or trajector roles, *given a (spatial) pivot*. We employ two annotated datasets (CLEF and GUM (Maptask)) and describe it in Section 4.2.

In an additional **relation extraction** phase, we assemble the results of the previous two steps to form spatial relation triplets with spatial indicators and their trajector and landmark arguments (see also Algorithm 1). This step is straightforward and involves no learning. We also investigate an alternative approach in which we tackle both steps jointly:

—**Finding spatial indicators and their arguments jointly:** In this task, we do not use a separate preposition disambiguation step but instead learn to tag all words in a sentence jointly. The examples in the dataset are used to train a single classifier that assigns the spatial indicator, trajector, and landmark roles simultaneously. Classifications can therefore correlate without using additional data resources (e.g., TPP). Section 4.3 describes this approach.

The remainder of this section describes the features and algorithms we designed and implemented for the spatial relation recognition task.

4.1 Learning Spatial Indicators

Various lexical categories (e.g., verbs, adjectives) can express spatial information, but prepositions primarily do so [Baldwin et al. 2009]. However, because prepositions often have different senses [Tratz and Hovy 2009; Litkowski and Hargraves 2007], we wish to *recognize whether* they convey a spatial sense. The sense of prepositions can be disambiguated by machine learning methods, as a large corpus exists for it. We consider prepositions because of their importance and the feasibility of the disambiguation task.

According to the aforementioned formalization, the set I contains only prepositions and $I(w) = 1$ holds only for prepositions with spatial sense. We aim to promote the use of a specific training scheme for preposition sense disambiguation and not perform other linguistic techniques to recognize them. The *locatives* recognized by SRL might be a solution, but this is often not true. The following two examples stem from the preposition disambiguation dataset (TPP) [Litkowski and Hargraves 2007].

(i) *He saw Owen redden with pleasure and laughed flinging an arm* ***about*** *his shoulders . . .*

(ii) *This project compares assumptions incorporated into social policies* ***about*** *these obligations . . .*

Prep	**POS**	**DepRel**	**SRL**	**sense**
about(i)	IN	NMOD	Arg1	spatial
about(ii)	IN	NMOD	Arg1	topic

Table I. Assigned labels by a POS tagger, dependency tree and SRL to *"about"* with two senses.

Table I shows the labels assigned by a part-of-speech (POS) tagger, a dependency parser, and SRL to the preposition *"about"*. The parse tree, the dependency tree and even the semantic role labeler could not distinguish between two senses of the preposition *"about"*. We therefore propose to *learn* these senses from a corpus labeled with senses (TPP) provided for the preposition disambiguation task (SemEval07) [Litkowski and Hargraves 2007], featuring the category $SpatialSense$ among others.

More specifically, the component $\hat{I}$ performs this preposition disambiguation task in Equation 1. It uses the following linguistically motivated features and the preposition contextual features that we aim to classify:

- The **preposition** itself
- By exploiting the dependency parser:
  - The words directly dependent on the preposition (*head1*)
  - The words on which the preposition is directly dependent (*head2*)
- For the predicates which have a dependency relation with the preposition:
  - All words that are arguments of the predicate other than the preposition are added using a semantic role labeler

For all extracted words satisfying the above conditions, the following features are also included:

- The **lemma**

- The **part-of-speech** tag (POS)
- The type of **dependency relation** (DPRL)
- The **semantic role labels** and, for predicates, the sense of the predicate (if assigned)

We present a sentence containing a preposition and the extracted features as an example.

He saw Owen redden with pleasure, and laughed , flinging an arm about his shoulders ...

$$\begin{array}{ll}
\{Preposition(''about''), & Preposition\_POS(''IN''), \\
Preposition\_DPRL(''NMOD''), & Preposition\_isarg(''A1\_arm.01''), \\
head1(''arm''), & head1Lemma(''arm''), \\
head1\_POS(''NN''), & head1\_DPRL(''OBJ''), \\
head1\_sense(''arm.01''), & head1\_isarg(''A1\_flinging.01''), \\
head2(''shoulders''), & head2\_lemma(''shoulder''), \\
head2\_POS(''NNS''), & head2\_DPRL(''PMOD''), \\
head2\_isarg(shoulders.01), & head2\_sense(''shoulders.01'')\}
\end{array}$$

To identify the spatial prepositions, we use the TPP data provided for the preposition disambiguation task, SemEval07 [Litkowski and Hargraves 2007]. We extract the features from the training and test data and use a *maximum entropy* and a *Naive Bayes* classifier to disambiguate the prepositions' sense. This process results in a binary classification of a preposition's spatial or nonspatial sense.

4.2 Trajector and Landmark Classification

As explained in Section 4, a multi-class classifier $\hat{R}$ must be trained to map each word w onto a class label from the set $\{\text{trajector}, \text{landmark}, \text{none}\}$, given a spatial indicator s. Because spatial indicator features are used to classify the roles of words in the sentence, the spatial indicator must be known before classifying trajectors and landmarks. Hence, we utilize the first step of preposition sense disambiguation, described in the previous section, to recognize the spatial indicators first, after which its arguments (trajectors and landmarks) can be classified. The generic feature set used in Equation 2 can now be defined in more detail using three different sorts. The first set of features relates to the word that we aim to classify ($f_1(w)$), the second includes the features of the spatial indicator of which the word may be an argument ($f_2(s)$), and the third contains the features that relate the word to the sentence's indicator ($f_3(w, s)$). SRL inspired these features, but they center on the spatial indicator. As mentioned, features are defined for head words.

- Features of a word w — $f_1(w)$:
  - The **word (form)** of w.
  - The **part-of-speech tag**.
  - The **dependency to the syntactic head in the dependency tree**.
  - The **semantic role**.
  - The **subcategorization** of the word (sister-nodes of its parent node in the tree).
- Features of the spatial indicator s — $f_2(s)$:
  - The **spatial indicator** word (form).
  - The **subcategorization** of s.
- Relational features of w w.r.t. s — $f_3(w, s)$:
  - The **path** in the parse tree from the w to the s.
  - The binary **linear position** of w with respect to the s (e.g., *before* or not).

—The number of nodes on the path between s and w normalized by dividing over the number of all nodes in the parse tree (to obtain an integer value it is reversed and rounded afterwards):

$$\mathrm{distance} = \frac{\text{\#Nodes on the path between } s \text{ and } w}{\text{\#Nodes in the parse tree}}$$

Take the following sentence as an example.

”The vase is on the ground on your left.”

Here, the input features for classification of *”vase”* w.r.t. the first *”on”* are:

$$\underbrace{\text{”vase”, ”NN”, ”SBJ”, ”A0”, NP-VP}}_{f_1(w)}\ \underbrace{\textbf{”on”}\text{, ”NP”}}_{f_2(w)}\ \underbrace{\text{NN} \uparrow \text{NP} \uparrow \text{S} \downarrow \text{VP} \downarrow \text{PP} \downarrow \text{IN, ”true”, ”3”}}_{f_3(w,s)}$$

A semantic role labeler is typically trained on a large external dataset. Using assigned semantic roles as features brings in additional knowledge, which may not be present in the dataset used to train the spatial role labeler. This issue encourages the use of the semantic roles as features.

The task is now a multi-class classification problem in which each word, represented by a feature vector, is separately classified, assuming that these classifications are independent. We use such a model in our initial experiments. In subsequent models, words are also described by their features, but the class to which they are assigned depends not only on their own values but also on the other feature vector values and relations among the various classes. The obtained class of a word may constrain the class of the next word.

We therefore employ several *conditional random field* (CRF) models. In these models, a sentence is a sequence of observations (i.e., words), $\langle w_1, \ldots, w_N \rangle$, which can be represented using a probabilistic graphical model. Each observation can be described in terms of the described feature vectors, and the model outputs a label for each word in the sequence.

After recognizing the trajector and landmark given a spatial indicator, we have all the relation elements. Relation extraction is performed in a straightforward way, by assembling all extracted spatial indicators, trajectors, and landmarks and combining them into spatial relation triplets. Algorithm 1 shows the entire process, based on preposition disambiguation and trajector/landmark classification.

4.3 Learning Spatial Relations without a Priori Spatial Indicator Classification

The spatial role labeling task can be seen as a *joint classification task*: to predict each triplet of segments as *being in the indicator-trajector-landmark* relation or not. In the previous section, we outlined a pipelining method for spatial role labeling, where a preposition (i.e., spatial indicator) is classified as spatial or nonspatial and the trajector and landmark are then sought for the obtained spatial indicators. Our focus on prepositions added one constraint to this task; the indicator should be a preposition (a realistic bias in English). The main purpose of this pipeline approach is to exploit a large external data source (TPP) for spatial sense disambiguation.

Combining two steps of the pipeline provides another option for learning spatial relations. We could omit the first step of using a dedicated classifier for spatial sense recognition, and learn to assign all spatial roles jointly, i.e., tagging words with $\mathrm{trajector}$,

Algorithm 1 Spatial-Relation-Extraction(S : sentence) **returns** relations SR

```
 1: {preposition disambiguation}
 2: for all $w \in S$ do
 3:   Estimate $\hat{I}(w)$ by training a probabilistic classifier and
 4:   construct the set $I$ of all spatial indicators of the sentence $S$.
 5: for all $s \in I$ do
 6:   {trajector and landmark classification}
 7:   for all $w \in S$ do
 8:     Estimate a probabilistic multi-class classifier $\hat{R}$ and
 9:     construct the sets $T_s$ and $L_s$ according to the assigned labels.
10:   if $T_s = \emptyset$ then $T_s \leftarrow \{undefined\}$
11:   if $L_s = \emptyset$ then $L_s \leftarrow \{undefined\}$
12:   {relation extraction}
13:   $\mathrm{SR} \leftarrow \mathrm{SR} \bigcup \big\{ \langle s, t, l \rangle \mid t \in T_s, l \in L_s \big\}$
14: return $\mathrm{SR}$
```

$\mathrm{landmark}$, $\mathrm{spatial\_indicator}$ or none, based on a training dataset. To train the classifier, we can employ a procedure and examples as in the pipeline setting, but the classifier must then learn one more label ($\mathrm{spatial\_indicator}$). To test and evaluate the classifier on a new (unlabeled) sentence S, we see that S can contain several prepositions with spatial sense and many trajectors and landmarks, whereas the classifier can only assign a single label to each word. The solution we use here is to, again, generate multiple examples from S, where each example contains a designated pivot with specific features extracted for that word (e.g., path features from words to the pivot). For each example, the words are classified using these features. One must theoretically generate as many examples as there are words in S; in our practice, it suffices to do this procedure only for pivots that are prepositions. The main advantage of this setting is that the learning algorithm gets the freedom to classify trajectors, landmarks and indicators in the context of one another.

In the relation extraction step, we perform the same general steps as in Algorithm 1, differing primarily in that we take all prepositions as possible spatial indicators in the preposition disambiguation phase (lines 1–4) and that the classifier $\hat{R}$ now uses all roles, including $\mathrm{spatial\_indicator}$. This fact allows multiple words to be classified as spatial indicators in one sequence and could in principle allow the extraction of spurious relations. However, due to the learning bias (i.e., each example contains only one targeted preposition), we discovered that spurious relations are rarely extracted. While on the one hand, the joint setting enables a learning algorithm to use the information in the data without depending on external data resources, on the other hand, there is a hazard of becoming specialized to the spatial preposition distribution in the available data. The experimental results section empirically investigates this trade-off.

4.4 Algorithms

A *conditional random field* (CRF) is a state-of-the art model for context-dependent classification. A CRF is an *undirected graphical model* or *Markov random field*, conditioned on a set of observations X to predict a set of output variables Y. We define $G = (V, E)$ as an undirected graph (with vertices V and edges E) such that a node $v \in V$ corresponds to each random variable and $V = X \cup Y$. We denote an assignment to X by x, an assignment to a set $A \subset X$ by x_A, and similarly for Y. If each random variable $y \in Y$ obeys the Markov property with respect to G, then (Y, X) is a conditional random field. This model represents a probability distribution over a large number of random variables by a product of local functions that each depend on a small subset of variables. This *factorization* of the global probability distribution makes learning and inference feasible.

A CRF generally defines a probability distribution $p(y|x)$ as follows:

$$p(y|x) = \frac{1}{Z(x)} \prod\nolimits_A \Psi_A(x_A, y_A)$$

in which $\Psi_A(x_A, y_A)$ is a *potential function*, where $\Psi_A : V^n \rightarrow \Re^+$ and $Z(x)$ is the normalization factor:

$$Z = \sum_y \prod\nolimits_A \Psi_A(x_A, y_A) \quad \text{and} \quad \Psi_A(x_A, y_A) = \exp\left\{ \sum_{k=1}^{K(A)} \lambda_{Ak} f_{Ak}(x_A, y_A) \right\}$$

Finally the conditional probability is the following:

$$p(y|x) = \frac{1}{Z(x)} \prod_{\Psi_A \in G} \exp\left\{ \sum_{k=1}^{K(A)} \lambda_{Ak} f_{Ak}(x_A, y_A) \right\}$$

For the CRF experiments we use Mallet[2] and GRMM:[3]

—**Linear-chain CRF.** The structure of graph G is theoretically arbitrary; however, when modeling sequences (in our case, words of a sentence), the simplest graph is a linear-chain CRF in the form of a (often first-order) Markov chain [Lafferty et al. 2001; Sutton and MacCallum 2006]. In this setting, the spatial role label of a word in the sentence depends on the label of word in the previous position. Considering sequential relationships can increase the learning model's accuracy. The conditional probability $p(x|y)$ is

$$\frac{1}{Z(x)} \prod\nolimits_{t=1}^{N} \Psi_t(y_{t-1}, y_t, x)$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ is a sequence or other structural set of observations and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_K)$ is the corresponding set of labels assigned to X. In the spatial role labeling task, X ranges over the words of a sentence, while Y ranges over the classes trajector (tr), landmark (lm), spatial indicator (si in the joint setting) or none of these ($none$). $\Psi_t(y_{t-1}, y_t, x)$ is a *potential function*, which is a real-valued function that captures the degree to which the assignment y_t to the output variable fits the transition from y_{t-1} and X. The potentials typically factorize according to a set of features $F = \{f_k\}$ such that $\Psi(y_{t-1}, y_t, x) = \exp\{\sum_{k=1}^{K} \lambda f_k(y_{t-1}, y_t, x)\}$.

The linear chain CRF setting of Mallet uses a forward-backward algorithm to compute the marginal distributions and the Viterbi algorithm to compute the most probable sequence label assignment. For our task, allowing transitions unobserved in the training

[2] http://mallet.cs.umass.edu/download.php

[3] http://mallet.cs.umass.edu/grmm/index.php

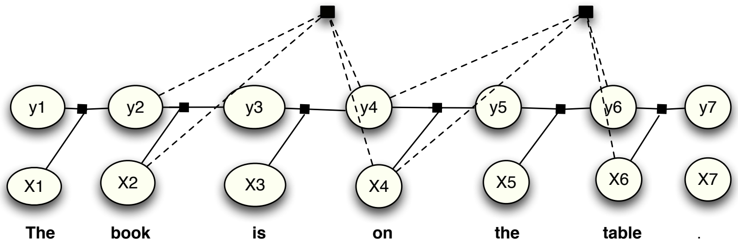

Fig. 2. Graphical representation of CRF with preposition template. Prepositions are connected to the candidate trajectors and candidate landmarks i.e noun phrases. Factors occur as black squares.

data during the inference and prediction phases adds more flexibility to the model, particularly when there are few training examples. This setting is called *fully-connected* in the Mallet tool, and we use it in our experiments. We refer to this setting as "linear chain CRF(FC)".

—**General CRF with preposition template.** In many relation extraction tasks, certain long-distance dependencies between entities play an important role. In our task, prepositions primarily play a spatial indicator role, while trajectors and landmarks are noun phrases. There could be many words in between the roles in the sentence that have no particular role and are assigned the *none* label. In light of this fact, we apply a version of a *skip-chain* CRF [Sutton and MacCallum 2006] to account for the probabilistic dependencies between distant labels. These dependencies are represented by augmenting the linear-chain CRF with factors dependent on the labels of the sentence's pivot preposition and noun phrases. The features on skip edges can incorporate information from the context of both endpoints, so the strong evidence of one endpoint can influence the label at the other endpoint. In our skip-chain CRF model, we exploit two clique templates: one is the normal sequential part (connecting neighboring words) and the other connects pivot prepositions to candidate trajectors and landmarks. Following the related work [Sutton and MacCallum 2006], the set of all pairs of positions for which there are skip edges (i.e., between prepositions and nouns) is represented as $PN = \{(u, v)\}$; the probability of label sequence y given input x is

$$p_\theta(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{N} \Psi_t(y_t, y_{t-1}, x) \prod_{(u,v)\in PN} \Psi_{uv}(y_u, y_v, x)$$

where Ψ_t are factors for sequential relations and Ψ_{uv} are factors over skip edges. We define the factors as $\Psi_t(y_t, y_{t-1}, x) = \exp\{\sum \lambda_{1k} f_{1k}(y_t, y_{t-1}, x, t)\}$ and $\Psi_{uv}(y_u, y_v, x) = \exp\{\sum \lambda_{2k} f_{2k}(y_u, y_v, x, u, v)\}$, where $\theta_1 = \{\lambda_{1k}\}_{k=1}^{K_1}$ are the parameters of the linear-chain template and $\{f_{1k}\}$ is the related set of feature functions or sufficient statistics. Similarly, $\theta_2 = \{\lambda_{2k}\}_{k=1}^{K2}$ are the parameters of the preposition template, and $\{f_{2k}\}$ is its related set of feature functions or sufficient statistics. The full set of model parameters is $\theta = \{\theta_1, \theta_2\}$. We use *loopy belief propagation* as the approximate inference algorithm in our experiments.

We compare the results of the CRFs with two baseline approaches:

—**MaxEnt (baseline) model.** As a baseline learning model, we classify the words of a sentence independently using a standard maximum entropy classifier.

—**Simple baseline.** To encourage the use of machine learning, a simple baseline is employed: given a spatial preposition, the first head word *before* the preposition is taken as the trajector and the *head word* after the preposition as the landmark. There is no learning from data in this setting, but the dependency tree is exploited to discover dependent headwords.

5. EXPERIMENTAL STUDY

In this section, we report on a series of experiments to evaluate various components of the spatial role labeling and relation extraction tasks.

5.1 Structure and Goals of the Experimental Setup

We present our leading research questions and identify the sections where we experimentally answer those questions.

—*Which data resources are available, or can be generated, to learn the spatial role labeling task from data?*
We answer this question in Section 5.2. In our experiments, we clearly want to solve the spatial role labeling task for unrestricted natural language input. However, we are limited by the amount of available data for machine learning. We describe our novel data resources and summarize their statistics.

—*How can we detect the spatial sense of prepositions using available resources?*
We answer this question in Section 5.3. We first investigate whether other resources (e.g., *locatives* obtained from SRL) can help and what benefits lie in directly learning the spatial sense from a large external and available data source (TPP).

—*If we assume that the spatial sense of a preposition is known or learned beforehand, how can we learn its corresponding trajectors and landmarks from data?*
In Section 5.4, we present various classifiers that take a given (spatial) preposition as input, and label the corresponding arguments (landmark and trajector) of the predicate the preposition represents.

—*What benefits lie in the sequential nature of finding the spatial sense of a preposition and then finding trajectors and landmarks (the so-called pipeline technique)?*
In Section 5.4.1, we first decouple the two problems and focus solely on the situation in which the spatial sense is known perfectly. In Section 5.4.2, we investigate two different situations where we fully automate the task, i.e., we use the preposition disambiguation output as input for spatial relation recognition. Without ground-truth data on the spatial sense of the prepositions, some landmarks or trajectors cannot be found because this spatial sense is classified incorrectly. We investigate a setting in which unknown prepositions are classified as spatial by default and another in which they are nonspatial by default.

—*What benefits lie in jointly recognizing spatial indicators, trajectors and landmarks, and how can long-distance dependencies help in this setting?*
In Section 5.4.3, we investigate an approach in which we learn to tag words with an extended label set that includes spatial indicators. This process side-steps preposition disambiguation as a separate phase; thus, classifications depend only on the information in one training dataset.

—*How do different pipelining methods affect the accuracy of the whole-relation extraction?*
In Section 5.5, we perform experiments in which we measure the accuracy of different pipelining techniques on the whole-relation extraction (thus finding the correct spatial indicators, i.e., prepositions *and* their correct landmarks and trajectors).

—*What is the effect of the used features on the extraction task?*
Section 5.6 discusses the effects of leave-one-out feature analysis.

—*What is the cross-domain performance of the approach on an unrestricted natural language text that contains both spatial and nonspatial information?*
In Section 5.7, we apply our system to several small, general, and unrestricted natural language texts to evaluate performance on data outside the training domain.

—*What are the main sources of errors in our approach?*
In Section 5.8, we investigate the errors made in 50 sentences of our dataset. We can distinguish five general categories of errors, including nested spatial relations and spatial focus shift. The errors caused by different model characteristics and different data domain characteristics are investigated in two separate subsections.

5.2 Dataset Description

For our experimental analysis, we use several manually annotated datasets. We describe their characteristics and usefulness for our study in this section. Statistics for the corpora are presented in Table II.

—**TPP dataset** For the preposition disambiguation task, we employ the standard test and training data provided by the SemEval-2007 challenge [Litkowski and Hargraves 2007]. It contains 34 separate XML files, one for each preposition, totaling over 25,000 instances with 16,557 training and 8,096 test example sentences; each sentence contains one example of the respective preposition.

—**GUM (Maptask) dataset** Because the spatial role labeling task is newly defined, there is no annotated English corpus available. However, the GUM (General Upper Model) evaluation data [Bateman et al. 2007], comprising a subset of a well-known corpus for spatial language is a useful dataset. It has been used to validate the expressivity of spatial relations in the GUM ontology. Currently, the dataset contains more than 300 English examples and 300 German examples. We used 100 English samples in this corpus that are originally from the Maptask corpus. The GUM-annotation for this sentence is an example:

”The destination is beneath the start.”

is:

SpatialLocating (locatum ”destination”, process ”being”, placement GL1 (relatum ”start”, hasSpatialModality UnderProjectionExternal)).

Here, *relatum* and *locatum* are alternative terms for landmark and trajector. *Spatial modality* is the spatial relation mentioned in the specific spatial ontology. The corpus contains 65 trajectors and 69 landmarks appearing in 112 spatial relations. Each sentence produces spatially labeled sequences in the number of its prepositions: 122 sequences for GUM (Maptask). Although complete phrases are annotated in this dataset, we only use a phrase’s headword with trajector (tr) and landmark (lm) labels and their spatial indicator (si). Using this small corpus to evaluate our approach for a very domain-specific corpus, including only instructions and guidance for finding the way on a map, is beneficial.

	CLEF	**GUM** (Maptask)	**Fables**	**DCP**
#Sentences	686	100	289	250
#Sequences	1430	122	864	809
#Spatial Relations	869	112	121	222
#Trajectors	839	65	106	199
#LandMarks	741	69	95	188
#Spatial Prepositions	735	112	121	222
#nonSpatial Prepositions	695	10	743	587

Table II. Data statistics.

- **CLEF dataset** Because the available dataset is small, an additional dataset[4] was annotated, based on textual descriptions of 400 images of the IAPR TC-12 Image dataset [Grubinger et al. 2006], hereafter the *CLEF dataset*. This dataset generated an additional 686 English sentences with 869 spatial relations. The CLEF dataset contains images taken by tourists with descriptions in several languages. The text describes objects with their absolute and relative positions in the image. It is therefore a rich resource for spatial information. However, the descriptions are not always limited to spatial descriptions and are thus less domain-specific and contain free image explanations.

We have annotated the textual descriptions with spatial roles of trajector (tr), landmark (lm) and their corresponding spatial indicator (si). Roles are assigned to the headwords of the phrases only. Two annotators provided annotations (325 sentences) and we investigated the *inter-annotator agreement* [Carletta 1996]. The *Kappa value* is 0.896 with a 95% confidence interval (0.882–0.910).

As mentioned above, we only consider prepositions as spatial indicators. This restriction is natural in English texts and especially for our data. Ignoring lexical categories other than prepositions has a trivial influence on our experiments with this corpus. Three exceptional cases exist in CLEF, where the words *crossing*, *supporting* and *away* are tagged as spatial indicators and this is the case for seven sentences in GUM (maptask) dataset. Furthermore, in compound verbs such as *"surrounded by"*, the preposition, here *"by"*, is annotated as the indicator although it is attached to the verb. However, for mapping to the spatial relation semantics, having the correct *pp-attachment* is an important feature, though beyond the scope of this paper.

In addition to the datasets mentioned above, two other corpora from different domains are annotated for evaluation purposes. These are described below.

- **DCP dataset** The dataset contains a random selection from the website of *The Degree Confluence Project*.[5] This project seeks to map all possible latitude-longitude intersections on Earth and have people who visit these intersections provide written narratives of the visit. The main textual parts of randomly selected pages are manually copied, and up to 250 sentences are annotated. Approximately 30% of the prepositions are spatial. This

[4] The datasets will be made publicly available.
[5] http://confluence.org/

percentage represents the proportion of spatial clauses in the text. These webpages are similar to travelers' weblogs but include more precise geographical information. The richness of this data enables broader applicability for future applications. Compared to CLEF, this dataset includes less spatial information, and the type of text is narrative rather than descriptive. It also contains more free (unrestricted) text. Moreover, the spatiotemporal information contained in this data has recently been used to extract discourse relations [Howald and Katz 2011].

—**Fables dataset** This dataset contains 59 randomly selected fable stories[6], which have been used for data-driven story generation [McIntyre and Lapata 2009]. The dataset contains a wide scope of vocabulary and only 15% of the prepositions are spatial, making it the most difficult corpus for our system. We annotated 289 sentences from this corpus for cross-domain experiments.

The datasets are preprocessed as follows. We generate parse trees for the sentences using the Charniak parser[7] [Charniak and Johnson 2005], and the LTH[8] tool [Johansson and Nugues 2007] produces the semantic roles and several other features in CoNLL-2008 output format.[9]

5.3 Preposition Disambiguation

Because this study concerns recognizing spatial prepositions, we investigated how accurately semantic role labeling (SRL) recognizes and labels the locatives in the TPP corpus before performing preposition sense disambiguation. We measure SRL's accuracy in labeling spatial prepositions with LOC (location) or DIR (direction). The results show that the *precision* is good. Whenever SRL recognizes the spatial sense, it is mainly correct; however, there are many cases in which SRL does not recognize spatial senses, rendering a lower recall and consequently a lower accuracy (Table III). This experiment provides an argument for the necessity of sense disambiguation even when recognizing only spatial prepositions. TPP contains 8,781 spatial prepositions and 14,681 nonspatial prepositions. The 99% confidence interval for the accuracy and F1-measure of both MaxEnt and Naive Bayes is $(0.875 - 0.89)$ and $(0.868 - 0.88)$, respectively. The reported results show the mentioned classifiers' performances in a multi-class classification setting with respect to the class of spatial prepositions.

System	**Precision**	**Recall**	**F1**	**Accuracy**
SRL(locatives)	0.83	0.49	0.53	0.59
Naive Bayes	0.86	0.92	0.88	0.88
MaxEnt	0.88	0.91	0.88	0.88

Table III. Accuracy of the detection of spatial or nonspatial preposition sense, relying on detected locatives when labeling semantic roles (SRL), using a Naive Bayes and maximum entropy classifier (MaxEnt). The results are given for the TPP dataset and averaged over 10 folds.

[6] http://homepages.inf.ed.ac.uk/s0233364/McIntyreLapata09/
[7] http://www.cfilt.iitb.ac.in/ anupama/charniak.php
[8] http://barbar.cs.lth.se:8081/
[9] http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=conll2008:format

In the preposition disambiguation experiments, we evaluate the recognition of *coarse-grained* senses on the preposition SemEval-2007 data [Litkowski and Hargraves 2007]. Coarse-grained senses include 20 general classes of preposition senses, such as spatial, temporal, causal, and membership.

System	**Accuracy**
Proposed-features(MaxEnt)	0.874
Proposed-features(NB)	0.86
MELB-YB(Best in SemEval-2007)	0.861
BOW(MaxEnt)	0.81
FreqSense	0.649
FirstSense	0.61

Table IV. Accuracy of coarse grained disambiguation (TPP).

Table IV gives the accuracy of a 10-fold cross-validation using a maximum entropy classifier and a Naive Bayes classifier. This table shows the results of the best system in the SemEval-2007 challenge for this coarse-grained sense disambiguation, the accuracy of applying bag of words (BOW), using the most frequent (FreqSense) and first (FirstSense) senses as baselines. The difference between our system and the best system from SemEval-2007 is statistically significant with a 95% confidence level ($p < 0.05$). Table III gives the evaluation considering only the prepositions' spatial sense, as mentioned before, compared to SRL's recognition. Table V gives results for some frequently used prepositions (e.g., *in*, *on*, *after*, *before*).

Preposition	**Naive Bayes**			**MaxEnt**			**SRL**		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
on	0.733	0.963	0.832	0.788	0.950	0.861	0.707	0.399	0.510
after	0.500	0.900	0.643	0.540	0.700	0.609	0.000	0.000	0.000
in	0.660	0.920	0.769	0.697	0.882	0.779	0.558	0.906	0.691
before	0.670	0.857	0.750	0.800	0.570	0.666	0.500	0.428	0.461

Table V. Accuracy of the detection of spatial or nonspatial preposition senses for some frequently used prepositions in the TPP dataset.

Although other work [Tratz and Hovy 2009] on preposition sense disambiguation outperforms results of the SemEval-2007 challenge too, the authors report only on the results of *fine-grained* sense disambiguation, which was not required for spatial sense recognition in our setup.

As the TPP data are a benchmark problem, we use a similar evaluation setting for comparison purpose and do not further experiment with different training regimes (in train/test splits). The current preposition disambiguation results are a promising start for spatial sense recognition and spatial relation extraction. After the evaluation process, the final preposition sense classifiers were constructed using the whole TPP dataset. We implemented 34 classifiers for the prepositions. For some prepositions in CLEF, e.g., *"opposite"*, no classifier exists. This issue occurred in 35 of 1,430 cases. Table VI shows the preposition disambiguation performance on GUM (Maptask) and CLEF. GUM (Maptask) is more domain-specific and contains more spatial prepositions (112/122), including a larger percentage (24/122) of prepositions that are not found in the TPP corpus and thus not recognized as spatial prepositions. This fact leads to a lower recall for spatial preposition recognition in this corpus in comparison to CLEF. We use the disambiguated prepositions in this step in the pipeline of spatial role labeling.

Corpus	Precision	Recall	F1	#Unrecognized PPs
CLEF	0.858	0.818	0.84	35
GUM (Maptask)	0.97	0.71	0.82	24

Table VI. Performance of preposition disambiguation trained on TPP and tested on CLEF and GUM (Maptask).

5.4 Extraction of Trajector and Landmark

The classification of trajectors and landmarks is not an isolated classification of words, but a classification of relations between a word and spatial pivot. This statement is the underlying assumption for relation extraction in the experiments described below. We show results for different settings: i) using ground-truth preposition disambiguation; ii) using a pipeline approach in which the preposition disambiguation is learned from external data; and iii) using a joint classification mode in which spatial indicators, trajectors and landmarks are learned and classified together.

5.4.1 *Using Ground Truth for Preposition Disambiguation.* To extract the trajectors and landmarks related by a spatial pivot, we first use the disambiguated ground-truth pivots. We implemented two different classification settings. In one setting, we classify each word based on its related extracted features described in section 4.2 and using maximum entropy classifier. This process generates a multi-class classification setting in which each word is classified as trajector, landmark or none. In the second setting, we classify each word using probabilistic graphical models, particularly CRFs, considering its context (the sentence) and employing the same linguistic input features as the first setting. Tables VII and VIII show the precision, recall and F1 measures for each tag using 10-fold cross-validation on the CLEF and GUM (Maptask) datasets.

Method	Trajector			Landmark		
	Pr	Rec	F1	Pr	Rec	F1
MaxEnt(baseline)	0.775	0.744	0.758	0.916	0.853	0.881
Linear-chain CRF	0.870	0.744	0.801	0.950	0.869	0.907
Linear chain CRF(FC)	0.905	0.792	0.844	0.953	0.879	0.914
Simple baseline	0.269	0.413	0.326	0.456	0.784	0.576

Table VII. Extraction of trajector/landmark roles in the CLEF dataset relying on the ground-truth preposition sense; 10-fold cross-validation.

The results show that context-dependent classification models outperform the maximum entropy model and that the differences are statistically significant for $p < 0.05$, where the fully connected CRF model gives the best results. Using the fully connected setting of the simple tagger yields statistically significant improvements in trajector classification in CLEF and landmark classification in GUM (Maptask).

Method	Trajector			Landmark		
	Pr	Rec	F1	Pr	Rec	F1
MaxEnt(baseline)	0.862	0.931	0.891	0.776	0.762	0.750
Linear-chain CRF	0.990	0.959	0.973	0.916	0.918	0.915
Linear chain CRF(FC)	1.000	0.969	0.983	0.947	1.000	0.971
Simple baseline	0.008	0.015	0.011	0.337	0.500	0.402

Table VIII. Extraction of trajector/landmark roles in the GUM (Maptask) dataset relying on the ground-truth preposition sense; 10-fold cross-validation.

5.4.2 *Pipeline Setting – Exploiting Preposition Disambiguation.* In this experiment, we fully automate the tasks of recognizing spatial roles and the corresponding spatial relations. The preposition disambiguation and the extraction of trajector/landmark tasks are connected and followed by the whole-relation-extraction. The preposition classifier is trained on the TPP dataset. The landmark/trajector/none classifier is trained on the subset of GUM and also the CLEF dataset.

In this setting, various options are examined during the test phase. Each preposition in a sentence is given to the relevant classifier from the 34 TPP-classifiers. If it does not match a TPP preposition, it is an *unknown* preposition and treated in two distinct ways: i) nonspatial (first row in Tables IX, X) or ii) spatial (second row in the tables). If the preposition is recognized as spatial, the process of the trajector/landmark extraction is performed; otherwise, all words in the sentence are labeled as none with respect to that preposition. We compare these settings to the one in which every preposition is blindly assumed to be a spatial indicator. These results help to assess the effect of preposition disambiguation.

Training and test instances are drawn from sentences in the respective datasets. For each preposition recognized in the sentence, a distinct instance of the sentence is created. In training instances, only the landmark(s) and trajector(s) (if any) in a spatial relationship with the pivot of the instance are annotated. In test instances, trajector(s) and landmark(s) (if any) in a spatial relationship with the pivot of the instance are automatically labeled.

Method	Trajector			Landmark		
	Pr	Rec	F1	Pr	Rec	F1
Pip(unrec PP nonSp)	0.886	0.654	0.752	0.914	0.714	0.801
Pip(unrec PP Sp)	0.889	0.685	0.773	0.916	0.741	0.819
All PP's spatial	0.870	0.792	0.828	0.904	0.878	0.891
Ground truth PP's	0.905	0.792	0.844	0.953	0.879	0.914
Joint Learning	0.884	0.668	0.759	0.919	0.712	0.802
Joint Learning+PPtemplate	0.988	0.998	0.980	0.866	0.892	0.843

Table IX. Extraction of trajector/landmark on CLEF dataset, comparing pipeline, ground-truth and joint learning by 10-fold cross-validation.

The experimental results in Table IX show that exploiting the linguistic features of the correct spatial preposition in the CLEF corpus improves the trajector and landmark extraction performance compared to pipelining, as expected. The difference is statistically significant $(p < 0.05)$. However, in the complete extraction problem, i.e., with unknown spatial indicators, assuming all prepositions to be spatial yields the highest recall, as it allows the trajector/landmark classifier to find related arguments. The pipeline model (assuming unrecognized prepositions as spatial), receiving input from the preposition disambiguation module, improves precision but lowers recall. Investigating the errors indicates that no trajectors and landmarks are generally extracted when nonspatial prepositions are recognized as spatial and the words are correctly classified as none. However, having a spatial preposition wrongly classified as nonspatial prohibits trajector and landmark extraction, causing a drop in recall.

Method	**Trajector**			**Landmark**		
	Pr	Rec	F1	Pr	Rec	F1
Pip(unrec PP nonSp)	1.000	0.510	0.660	0.930	0.460	0.580
Pip(unrec PP Sp)	1.000	0.701	0.801	0.937	0.660	0.752
All PP’s spatial	1.000	0.969	0.983	0.947	1.000	0.971
Ground truth PP’s	1.000	0.969	0.983	0.947	1.000	0.971
Joint Learning	1.000	0.956	0.976	0.920	0.956	0.934
Joint Learning+PPtemplate	0.934	0.945	0.936	0.720	0.760	0.727

Table X. Extraction of trajector/landmark on GUM (Maptask) dataset, comparing pipeline, ground-truth and joint learning by 10-fold cross-validation.

In the GUM (Maptask) corpus, inputting the correct preposition does not make a significant difference compared to “all spatial”; moreover, pipelining yields lower recall. GUM (Maptask)’s statistics show that more than 93% of the prepositions are spatial and errors in preposition disambiguation prohibit the extraction of related trajectors and landmarks, resulting in a sharp drop in recall with no significant variation in precision.

5.4.3 *Joint Learning Setting.* According to section 4.3, each training instance in this setting contains at most one preposition labeled as a spatial indicator and annotations for only the landmark(s) and trajector(s) (if any) of that spatial indicator. Each sentence gives several instances, up to the number of prepositions it contains. In test instances, the $\mathrm{trajector}$, $\mathrm{landmark}$, $\mathrm{spatial\_indicator}$ and none labels are automatically assigned to each word in the sentence based on the input features for a given preposition.

Because spatial indicators are classified jointly with other spatial roles, some of the errors caused by the pipelining can be removed. However, as Table IX shows on the CLEF dataset, the recall of best pipeline system (unrec PP Sp), is slightly higher than joint learning in trajector and landmark classification, and the improvement is statistically significant $(p < 0.1)$.

Adding long distance dependencies to joint learning through the preposition template greatly improves performance on CLEF dataset, particularly in trajector classification. In contrast, a sharp decrease in landmark classification occurs in GUM (Maptask). The difference in language characteristics in these datasets affects these results, which calls for further investigation. In Section 5.8, an error analysis categorizes the types of errors that can occur in the spatial role labeling task and the errors of two models (with and without a template) are compared using a test subsample.

For GUM (Maptask), Table X shows that assuming all prepositions to be spatial outperforms other settings, including joint learning. The previous experiments show joint learning outperforming pipelining, though the pipeline setting uses the external resource

TPP. Cross-domain differences and sentence types in TPP, CLEF, and GUM (Maptask) datasets account for this discrepancy. This issue will be discussed later in this paper.

Method	**WR** (GUM)			**WR** (CLEF)		
	Pr	Rec	F1	Pr	Rec	F1
Pip (unrec PP nonSp)	0.874	0.534	0.663	0.653	0.605	0.628
Pip (unrec PP Sp)	0.894	0.722	0.799	0.547	0.627	0.584
All PP's spatial	0.870	0.948	0.907	0.391	0.722	0.507
Ground truth PP's	0.948	0.948	0.948	0.704	0.723	0.714
Joint learning	0.888	0.904	0.896	0.704	0.737	0.720
Joint learning+PPtemplate	0.672	0.703	0.684	0.830	0.830	0.830

Table XI. Extraction of whole relations (WRs) on GUM (Maptask) /CLEF, comparing pipeline, ground-truth and joint learning using 10-fold cross-validation.

5.5 Whole Relation Extraction

A correct relation is a relation in which all three components, i.e., a spatial indicator and its corresponding trajector and landmark, are correctly recognized. Every wrong assumption about a spatial indicator initiates a new wrong spatial relation. In this case, the precision and recall of the whole relation are as follows.

TP = the number of correctly produced spatial relations
FP = the number of incorrectly produced spatial relations
FN = the number of spatial relations that are incorrectly not produced

Recall and precision are thus:

$$\text{recall} = \frac{TP}{TP+FN} \text{ and } \text{precision} = \frac{TP}{TP+FP}$$

When the preposition is incorrectly classified as spatial, the number of FPs increases, leading to lower precision. If the preposition is incorrectly classified as nonspatial, the number of FNs increases, leading to lower recall.

As observed in Table XI, based on this experiment, assuming all prepositions as spatial is generally impractical. The low performance on CLEF indicates that the relation extraction by this assumption is not robust for unrestricted language, though this setting works well for trajector and landmark extraction on GUM (Maptask). Employing ground-truth prepositions provided the best results for GUM (Maptask), though we observed no significant difference compared to joint learning for relation extraction in CLEF. To explain how the joint learning setting can, in this particular case, perform as well as the ground-truth setting, we must examine the input and output features of the models. In the ground-truth setting, the (correct) spatial indicators function as input, and the classifier learns to label trajectors and landmarks, but not spatial indicators. In the joint learning setting, the model learns to utilize the correlations between trajectors, landmarks *and* spatial indicators and outputs labels for all of them, so it considers the transitions between spatial indicators and other labels. The settings thus differ in input/output. Future work may consider integrating the settings more tightly by stacking the classifiers and incorporating the output of a TPP-trained classifier into the joint model or by employing a joint learning setting in which the spatial indicator values are 'clamped' and used as hard constraints, fully using both the TPP data and learning joint probabilities over the spatial relation triplets.

In the two pipeline settings, assuming prepositions to be spatial (for unrecognized prepositions with TPP classifiers) shows better results in GUM (Maptask) but worse results in CLEF. This finding is reasonable due to the prior distribution of spatial prepositions in GUM (Maptask) and CLEF, as discussed above. The joint learning setting gives the best results for whole relation extraction on CLEF. This setting is ideal for spatial role labeling problem when there are sufficient training examples, which is not always the case. The pipeline setting performs better in some trajector and landmark classifications, which signals the significance of exploiting the TPP resource. Our final experiments on texts from different domains in Section 5.7 highlight the importance of the TPP resource.

Data	Features	Trajector			Landmark			WholeRelation		
		Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
CLEF	All features	0.905	0.792	0.844	0.953	0.879	0.914	0.704	0.723	0.714
	-dis	0.889	0.792	0.836	0.956	0.879	0.915	0.697	0.717	0.707
	-SRL	0.893	0.795	0.840	0.961	0.876	0.916	0.701	0.717	0.709
	-dis-wordsubcat	0.883	0.770	0.822	0.954	0.871	0.911	0.680	0.693	0.687
GUM	All features	1.000	0.969	0.983	0.947	1.000	0.971	0.940	1.000	0.971
	-dis	1.000	0.969	0.983	0.920	0.987	0.951	0.932	0.947	0.940
	-SRL	0.987	0.956	0.969	0.917	0.924	0.916	0.874	0.894	0.884
	-dis-wordsubcat	0.983	0.969	0.975	0.920	0.946	0.929	0.906	0.921	0.913

Table XII. The effect of applying distance (dis), word subcategorization (wordsubcat) and SRL feature for trajector and landmark extraction, using ground-truth preposition senses. The baseline uses all features.

5.6 Experimental Feature Analysis

As mentioned in section 4, SRL inspired most of the employed input features. However, we also used the distance, word subcategorization and semantic roles. In the results reported above, we use all of the features described in that section. By investigating the features' impacts and omitting them one by one, we determined that almost all features contribute positively to the performance. The path feature contribution was marginal, especially for GUM (Maptask). Because GUM (Maptask) is a small corpus and the path feature has too many unique values in our dataset, its discriminative power is limited here. The complex path feature generally can produce some overfitting or inserts noise into the model, due to incorrect prepositional phrase (PP) attachments, for example. The distance between the preposition and its arguments is thus a valuable feature that helps determine whether a word is an argument of a preposition. The experiments with and without this feature show a positive impact on both datasets; an overall gain of approximately $1\% - 3\%$ for both GUM (Maptask) and CLEF is statistically significant $(p < 0.1)$. To understand the effect of our additional features, we use ground-truth preposition senses, and Table XII shows the results.

Exploiting more discriminative structural features may compensate for the lack of lexical information, we therefore evaluate adding the subcategorization of a target word using the aforementioned definition. The last row in Table XII for each dataset shows the performance using neither distance nor sub-categorization. The quantitative effect of the SRL feature is represented in the same table. The table clearly shows the positive influence of this feature on GUM (Maptask), but it contributes less for CLEF. GUM (Maptask) contains directional instructions with few compound locative descriptions, so there are more direct relations between semantic roles, including AM-DIR, AM-LOC and being a landmark as well as between a “patient” role and being a trajector.

5.7 Cross-domain Evaluation

Although our methodology for extracting spatial semantics is domain independent, the general problem still depends on lexical features. A model trained in one domain and later employed in another often performs poorly due to feature distribution changes. Additionally, classifiers trained on the new domain alone may suffer from too few training samples. Other applications of machine learning methods share this problem [Jiang et al. 2008], particularly the natural language processing area. Because one of our contributions is preparing a corpus for spatial information extraction, we annotated several types of textual data for domains in which spatial information extraction could be an important issue. As explained in previous sections, our main data set is CLEF, which contains many spatial descriptions but is still balanced with nonspatial information. The GUM (Maptask) corpus is much smaller, domain-specific and biased to spatial descriptions; learning from the same corpus in a cross-validation setting produces good results. We based our experiments on these two data sets to show that our problem, spatial role labeling, is feasible and that specific learning algorithms and representations are effective. In this section, we discuss experiments that explore *transfer* capabilities from one dataset (domain) to another and test the advantage of using external data resources (e.g., TPP) in that process.

Corpus	**Precision**	**Recall**	**F1**	**#Unrecognized PPs**
Fables (TPP)	0.444	0.657	0.530	13
Fables(SRL-locatives)	0.495	0.420	0.454	-
DCP (TPP)	0.584	0.687	0.631	29
DCP(SRL-locatives)	0.226	0.423	0.295	-

Table XIII. Preposition disambiguation performance trained on TPP and tested on Fables and DCP.

Our first experiment concerns the intrinsic cross-domain nature of employing TPP data. As previously done for CLEF and GUM (Maptask), we evaluate preposition sense disambiguation performance on the new datasets Fables and DCP. This classifier is also used in the pipeline setting in subsequent experiments. The results in Table XIII indicate that the preposition spatial sense recognition is harder in these data sets than in CLEF and GUM (Maptask).

However, for Fables and DCP datasets, the TPP-based model outperforms SRL in spatial preposition recognition. The results also show that the SRL system is more accurate for Fables than DCP. The more frequent use of compound verbs in Fables may account for this phenomenon, as the prepositions are mostly attached to verb phrases.

In a second set of experiments concerning trajector and landmark extraction, we applied the settings described in previous sections to Fables and DCP. The models were trained on CLEF and tested on these data sets. For the joint learning setting, we applied the learned classifier (for spatial indicators, landmarks and trajectors) to the unlabeled Fables/DCP data. For the pipeline setting, the classifiers trained on TPP find the spatial indicators, after which we apply a classifier trained on CLEF (for trajectors and landmarks) to the

Method	Fables			DCP		
	Trajector	Landmark	Indicator	Trajector	Landmark	Indicator
Pip(unrec PP nonSp)	0.080	0.397	0.530	0.208	0.435	0.631
Pip(unrec PP Sp)	0.100	0.424	0.348	0.232	0.463	0.554
All PP's spatial	0.181	0.342	0.245	0.293	0.447	0.431
Ground truth PP's	0.231	0.620	——	0.338	0.590	——
Joint Learning	0.113	0.378	0.45	0.223	0.432	0.614
Joint Learning+PPtemplate	0.101	0.292	0.333	0.163	0.319	0.409

Table XIV. F1-measure of cross-domain evaluation; the classifiers learned on CLEF and tested on fable stories and DCP data.

unlabeled Fables/DCP data. Table XIV reports the results, to summarize the tables only F1-measure is presented. Confidence intervals (90%) for the last column are (0.428–0.532) and (0.423–0.527), and all others have a lower variance. The table shows that the pipeline setup outperforms joint learning, demonstrating the benefits of the model trained on TPP and the value of exploiting TPP in this experiment. The outperforming of the pipeline is statistically significant.

For trajectors and landmarks, the first unsurprising result is that the ground-truth prepositions, trajectors and landmarks can be classified more accurately in both data sets. The whole relation extraction (not shown) proved more difficult here. Once more, in Section 5.8 we chose a sample of the errors to obtain a clearer analysis on cross-domain evaluation.

Dataset	Best method	Trajector	Landmark	Indicator	whole_rel
Fable stories	Joint Learning	0.544	0.569	0.638	0.481
Confluence	Joint Learning	0.518	0.595	0.685	0.475

Table XV. F1-measure of 10-fold cross-validation; the best method has the maximum F1-measure averaged over all roles.

For completeness, we evaluate how well our techniques work on the additional datasets without training on CLEF using standard 10-fold cross validation in a third experiment (see Table XV). To summarize the tables, we only present the F1-measures of the most outperforming models, where we find that joint learning is the best setting for both datasets. Considering the previous cross-domain experiment, this result is reasonable. The joint learning setting shows higher performance with 10-fold cross validation because the training and test sets have similar (lexical) feature distribution. Overall, the evaluation of these two datasets performed worse than our main datasets, CLEF and GUM (Maptask), because of the broad vocabulary range in these additional datasets and the lower proportion of spatial expressions. This situation requires more training examples to obtain acceptable accuracy. In Section 5.8, a brief discussion on the errors of this experiment is given too.

5.8 Error Analysis

The experiments using the GUM (Maptask) and CLEF datasets clearly indicate that dependencies between observed nodes in the CRF model are advantageous for spatial role labeling. Most errors are classical information extraction errors. The lack of a huge training corpus with sufficient word occurrences results in invalid argument assignments concerning spatial semantics. Cross-domain experiments on Fables and DCP are more affected by this lack. In the pipeline setting, errors are primarily propagated from one phase to another. The more elaborate solution of jointly classifying prepositions and trajector/landmarks should, theoretically, provide a better solution. However, this setting suffers even more from the lack of lexical information but shows promising results in general. This setting could be the best platform with the injection of the partially labeled external TPP resource. Many words in a sentence have ambiguous meanings, which also causes errors, as in other semantic annotation tasks. In particular, errors may occur more often in sentences with more than one relation due to the issues mentioned above. In the following subsections we consider three subsamples of sentences, two from test folds of CLEF and one from the Fables dataset, to investigate the error types and the ways that model characteristics and data characteristics cause certain errors.

5.8.1 *Error types.* To understand the nature of the errors (i.e., other than those from pipelining), we manually inspected over 10% of the errors, 50 wrongly labeled sequences from the largest data set CLEF. We selected the setting with a given ground-truth preposition to analyze problematic issues in classifying trajector and landmark roles and relation extraction. Table XVI categorizes the errors based on their cause and gives the percentage of each category in the random sample.

Class	**Description**	**Percentage**
1	A role element is classified as none	48%
2	Nesting spatial relations	24%
3	Spatial focus shift	10%
4	Irregularity in the grammar	10%
5	Errors in the annotated data	8%

Table XVI. Error classes.

—**Class 1**. One frequent error assigns none labels to words that play spatial roles. This error originates from two sources: the lack of lexical information and the high prior probability of the none class compared to role-holder words, leading to lower recall of both trajectors and landmarks. The latter generally causes errors in experiments on the CLEF dataset. In the sentence below, *"woman"* is wrongly classified as none, which the latter issue causes.
Example: *A* $[woman]_{Tr}$ *holding a plastic bag* $[on]_{Ind}$ *the left.*
However, the first cause (i.e., the lack of lexical information) generally affects errors in the cross-domain experiments in section 5.7.

—**Class 2**. These errors are caused when the sentence expresses spatial relations that are more complex. In these cases, multiple trajectors are assigned to a preposition. In nested relations, the spatial relation has the transitivity property, so the assigned roles are semantically correct. However, we avoid spatial reasoning in the hand-labeled data, and these relations have not been annotated. The transitivity property of the relation depends only on the context, type of relation and its trajector and landmark entities. Injecting these more complex inputs makes the learning more difficult for the machine learning model, particularly when it lacks training data. These additional role assignments are

classification errors and cause lower precision particularly in trajector labeling in our dataset.
Example: *A dark-haired girl in a white T-shirt is sitting at a* $[desk]_{Tr}$ $[in]_{Ind}$ *a* $[classroom]_{Lm}$.
With respect to the second *"in"*, only *"desk"*is the annotated trajector, though the classifier also classifies *"girl"* as a trajector. This assignment is semantically correct, but as described above, it does not match the ground-truth annotations.

—**Class 3**. This type of error concerns cases in which the transitivity property does not hold. A preposition trajector cannot semantically be a trajector of the next preposition in the sentence, but the landmark of the first relation is often the trajector of the next. In other words, the spatial descriptions' focus changes from one trajector to another. In these cases, a wrong trajector is assigned to a preposition and is related to a wrong landmark.
Example: *More kids sitting at their desks and a* $[blackboard]_{Tr}$ $[in]_{Ind}$ *the* $[background]_{Lm}$.
Depending on the context, one can infer that only the *"blackboard"* is in the *"background"* and the desks are not. Hence, *"background"* is a wrong landmark for both *"kids"* and *"desks"*.

—**Class 4**. The sentences' grammar causes this type of error, primarily the phenomenon of *semantic ellipsis*.
Example: *A king size bed with the night* $[table]_{Tr}$ $[on]_{Ind}$ *the* $[side]_{Lm}$.
Here, *"bed"* is classified as the trajector of *"on"*, while "table "is actually the trajector. In fact, *"side"* should be labeled as the landmark that actually refers to the side of "bed".

—**Class 5**. The annotator, not the classifier, causes these errors. This fact implies that accuracy can, to some extent, vary.

5.8.2 *Error Analysis Cross Folds and Models (CLEF and GUM).* Adding the preposition template had inconsistent impacts on the performance of CRF's on different datasets. Particularly, this impact was greatly positive on trajector classification in CLEF and negative on landmark classification in GUM. This inconsistency encouraged to take a small subset of testing examples and compare the errors of two models (with and without a template) to address the effects of adding the templates on each type of error.

In the CLEF dataset, several sentences contain nouns and prepositions between the pivot-preposition and its related trajector. The sequential joint learning makes errors due to assigning "none" to these long distance trajectors. The template performs the first correction to handle these long distance trajectors properly in the skip-chain CRF. To quantify, 65% (11 of 17) of the errors in the checked subsample (100 instances) are in this category, leading to lower recall in the linear-chain CRF. Those errors belong to class 1, and most are corrected by the skip-chain CRF model. The following sentence is an example:

Example: *A dark-skinned, dark-haired* $[boy]_{Tr}$ *with a gray shirt is standing in a room* $[in]_{Ind}$ *front of a* $[wall]_{Lm}$ *made of red bricks.*
The linear-chain model labels *"boy"* as "none" with respect to *"in" (front of)*, which the skip-chain model corrects it to "trajector".

The second type of error includes cases in which two trajector labels are assigned despite there being only one actual trajector. The previous subsection classifies and explains these errors as classes 2 and 3. In this subsample, we see that the long distance noun is the actual trajector in 3 of 17 such cases. In 3 other cases, the noun immediately before the preposition is the actual trajector. These errors, totaling 6 of 17 (36%), lead to a decrease in both recall and precision.

Example: *There is a wooden commode and a mirror on the left, a wooden bedside table with a table lamp next to the bed and a huge* $[fan]_{Tr}$ *on the wall* $[above]_{Ind}$ *the* $[bed]_{Lm}$. The linear-chain tagger labels both *"wall"* and *"fan"* as trajectors with respect to *"above"*, while the general skip-chain CRF correctly tags only *"fan"* as the trajector.

The only error made by the skip-chain CRF concerning trajectors in our subsample is the example below, in which the trajector *"boy"* is assigned a "none" label with respect to the second *"in"*, in the sentence:

Example: *A dark-skinned, dark-haired* $[boy]_{Tr}$ *in a very colorful pullover is standing in between two desks* $[in]_{Ind}$ *the* $[classroom]_{Lm}$.
This error is not typical but merely arbitrary, as there are similar cases in the test data that the skip-chain CRF model correctly classifies.

Furthermore, the improved model outperforms even the ground-truth in trajector classification. This finding is unexpected but not contradictory. In the ground-truth and pipeline settings, the correlations between indicators and other role labels are not considered, while joint learning uses this extra feature in the form of hidden variable(s). The template clearly increases the probability of assigning role labels (i.e., trajector/landmark) instead of a none label, with the additional probabilistic factor connecting distant nouns to the pivot-preposition; this process corrects the long distance words labeled as none and increases recall of both trajectors and landmarks. This feature removes one cause of class 1 errors. Because landmarks are usually in prepositional phrases and close to the pivot preposition, modeling long distance dependencies contributes less than for trajectors. However, it still increases recall of landmarks. It may, however, introduce additional false positive landmarks, as in the following example:

Example: $[Tourists]_{Tr}$ *are standing* $[in]_{Ind}$ *the* $[classroom]_{Lm}$ *of a school in front of the blackboard.*
Here, both *"school"* and *"classroom"* are labeled as landmarks of the first *"in"*. The F-measure, therefore, has less improvement in landmarks than in trajectors.

In contrast to CLEF, sentences are short in GUM (Maptask), and modeling long-distance dependencies does not improve recall. Some cases lack trajectors because sentences contain directional instructions in which *"you"* is the implicit trajector. The skip-chain CRF thus only does equally well or slightly worse in trajector classification. Fitting the more complex model to the small amount of data in GUM (Maptask) lowers both the recall and precision of landmarks. Additional investigation of one fold of the errors in the skip-chain CRF of GUM (Maptask) shows that many landmarks are annotated as none because both occurring a specific noun as a landmark in the training data and the combination of a landmark with a specific preposition are important to the model. However, the linear chain CRF is less strict and annotates them correctly. The additional probabilistic factor makes the model tend to overfit the data, strengthening the effects of the lack of training data and lexical information. The incorrect "none" labels here assigned more primarily due to the lack of training data than due to the frequently occurring "none" labels in sentences. The additional template can, therefore, also introduce class 1 errors, but for a different reason than mentioned above.

5.8.3 *Error Analysis Cross Domains and Models (DCP and Fables)* . The lower performance of cross-domain evaluation and also 10-fold cross validation on Fables and DCP encouraged an investigation on the incorrectly classified sentences in these datasets. A sample is selected from Fables's test errors because it shows more problematic than DCP.

Most of the errors belong to class 1. The high prior probability of the "none" labels in the sequence of words is the main cause. Adding the preposition template in the skip-chain CRF model increases the errors of this type. The increased complexity of this model and the limited training data typically cause *overfitting*, i.e., the model adapts to the training data characteristics too strongly and does not generalize properly. This type of error is more problematic for trajector classification, whereas the landmarks are frequently in prepositional phrases and close to the indicators. Syntactical information thus helps achieve higher recall for there. If the indicator has been identified correctly, landmarks are more easily recognized than trajectors.

For trajector classification, due to the variety of trajectors syntactical features, lexical information are more discriminative and useful for the model. In the example sentence below, in which gold labels are indexed, the trajector is incorrectly classified as "none" because the word *eagle* does not occur as a trajector in the CLEF dataset:

Example: *An* $[Eagle]_{Tr}$ *sat perched* $[on]_{Ind}$ *a lofty* $[rock]_{Lm}$*, keeping a sharp look-out for prey.*

The next example is another case in which none of the roles are recalled and all are labeled as "none". Because the type and context of the texts differ from Fables to CLEF, contextual features are ineffective.

Example:*A* $[huntsman]_{Tr}$*, concealed* $[in]_{Ind}$ *a* $[cleft]_{Lm}$ *of the mountain and on the watch for game.*

Conversely, exploiting grammatical features introduces more false positives and decreases precision for landmarks. The following sentence is an example of this phenomenon:

Example:*One touch from you and I should be broken* in *pieces.*

The model wrongly classifies *in* as an indicator and *pieces* as a landmark while *in* has no spatial sense. For this example, the semantic role labeler labels *in* as AM-LOC, which is also incorrect. Despite dissimilarities in the sentences' vocabulary and context, there are several cases where all roles have been labeled correctly. Their similarity to typical spatial description grammatical structures in CLEF accounts for this and the below sentence is an example:

Example:*There were two* $[Cocks]_{Tr}$ $[in]_{Ind}$ *the same* $[farmyard]_{Lm}$*, and they fought to decide who should be master.*

We also briefly study the errors made by the system in 10-fold cross-validation inside these datasets. The trajector and landmark classification precision is nearly 100% for both datasets, but recall is very low, signifying that the major problem is again insufficient evidence for assigning the roles, i.e., a lack of training data and particularly a lack of positive examples. If we compare the overall number of prepositions to the number of spatial prepositions, there are many more non-spatial prepositions per sentence in the Fables and DCP compared to GUM and CLEF, which leads to a stronger bias toward assigning none labels. Having an unbalanced dataset (with respect to positive and negative examples) is a typical challenge for relation extraction with machine learning.

Overall, the error analysis in these experiments indicates the main issues for successful transfer of models across different domains. It also suggests ways to improve spatial role labeling systems in the future.

Because labeling data to train a model in each domain of interest is inefficient, we have shown one way in which to use existing resources to alleviate the annotation labor. Experiments in different domains present difficulties in cross-domain transferability and indicate that learned classifiers become biased to the distribution of features and words in the training dataset. However, exploiting more general resources, such as TPP, can help reducing this bias. Other future directions include feature expansion and using latent word models [Deschacht and Moens 2009] to broaden the vocabulary range recognizable by the model. In fact, more abstract features and categories of sentence components can be obtained either by exploiting linguistic resources or from various corpora in an unsupervised setting. For example, acquiring the feature *animal* from *eagle* or more abstract properties, such as *physical object* about entities, could increase both recall and precision of classifying trajectors and landmarks. Moreover, using partially labeled data in the joint learning setting would gradually decrease the requirements for manually prepared, annotated data. Performance could improve by coupling these directions with underlying machine learning models. The last section of the paper presents more general future directions and research potentials.

6. RELATED WORK

Several research areas have studied spatial information, both practically and theoretically. The various perspectives include *cognitive*, *linguistic* and *computational* aspects. Useful application areas include *geographic information systems* (GIS), *navigation*, *natural language processing*, *robotics* and *computer vision* tasks. Most applied approaches concerning spatial information are domain-specific and only provide some feasible solution for specific tasks and domains. However, such narrow domains inevitably lead to less realistic settings and inflexible solutions and are unsuitable for real situations involving multi-disciplinary tasks. Studying various aspects of spatial information processing and clarifying connections to the research described in this article are still important, with the goal of exploiting state-of-the-art approaches in less domain-specific contexts. In this section, we try to draw a summarized descriptive image of related works.

In the literature, various formal representation languages and reasoning systems exist to handle spatial information [Galton 2009]. Moreover, researchers have proposed spatial calculi models, such as qualitative spatial calculi, and elaborated compositional extensions for spatial reasoning and inference [Stock 1997; Renz and Nebel 2007; Cohn and Renz 2008]. For most computational aspects of spatial information, linguistic issues have not been given much attention, and non-linguistic formalizations are often pursued. This is the case for fundamental works on both formalizing and applying spatial relations. Bateman et al. [Bateman et al. 2010] also argue this shortage. These formal models are inherently based on the logics of human spatial cognition, independent of linguistic constructs, and the way humans express such relations in natural language. Spatial cognition studies could make the connection between the two aspects more transparent because they investigate abstract spatial concepts that, on the one hand, are expressed in language and, on the other hand, are formulated in spatial calculi, particularly in qualitative spatial calculi models.

When the theory is brought into practice and spatial computational models are exploited, a specific set of spatial concepts is presumed and formalized to make the computations feasible in each application. When natural language processing is required, a simplified and domain-specific setting is employed. The lack of domain-independent applications is caused by the flexible, complex and unmanageable linguistic constructs that lead to under- or over-specificity when mapping natural language to formal representations. Bateman [Bateman 2010] extensively discusses this issue and argues the necessity of using two semantic levels, explaining linguistically oriented spatial ontologies, such as GUM [Bateman et al. 2007]. This type of ontology facilitates mapping between natural language and spatial calculi [Hois and Kutz 2008b; Bateman et al. 2007; Hois and Kutz 2008a].

In application-oriented works, the type of works in which we are interested, investigating the cognitive aspects is more important when targeting understanding *unrestricted* natural language. Few research works exist that consider both this problem and the abstraction of spatial concepts in their systems [Ross et al. 2005; Kollar et al. 2010]. We therefore describe the related works that inspired us to employ the selected spatial elements for this work and to identify other research efforts in this area.

In this work, we pay particular attention to spatial prepositions. From a cognitive-linguistic point of view and in related spatial language research, spatial prepositions, their semantics' variation, and grounding their perceived meaning have been thoroughly investigated [Herskovits 1986]. In the visual context, applying computational spatial preposition models to a visually situated dialog system is investigated [Kelleher and Costello 2009]. Lockwood et al. [Lockwood et al. 2006; Lockwood et al. 2008] describe a model for learning to classify visual scenes according to the spatial preposition depicted. They use SEQL, an existing model for analogical generalization, to construct relational descriptions from stimuli input, such as hand-drawn sketches, and their suggested model can distinguish between *in*, *on*, *above*, *below*, and *left* after being trained on simple sketches exemplifying each preposition. These efforts are valuable but remain too limited to ground unrestricted spatial natural language perception. A spatial preposition only portrays a symbol whose meaning should be extracted from its context, using other spatial elements in the language. In our work, the automatic mapping to the prepositions' meaning is performed by exploiting the first level of mapping the language to spatial roles. This plays an important role in the semantic representation of spatial prepositions in a domain-independent way.

The preposition disambiguation task employed in this paper has been introduced before as a benchmark task [Litkowski and Hargraves 2007; Tratz and Hovy 2009]. However, though we show improved results on the general (coarse-grained) task, our focus on spatial prepositions in spatial relation extraction is novel. The importance of prepositions in meaning conveyance has been extensively investigated [Baldwin et al. 2009], and prepositions' dominant role in language semantics has been experimentally proven. This fact also explains why prepositional sense disambiguation has recently received much attention in semantic text analysis [O'Hara and Wiebe 2009; Dahlmeier et al. 2009]. Exploiting preposition disambiguation in this work shows the benefits of this computational linguistic task in spatial relation extraction.

After processing the prepositions and their senses in a computational linguistic task, we formulated the extraction of the trajector, landmark, and spatial indicator roles as the second step in our spatial role labeling task. Related works have noticed these primitive spatial elements in both visual contexts and processing locative phrases [Barclay and Galton 2008]. Extracting these elements from language has been noticed in few applications containing multimodal environments or in tasks that are occupied with visual information and visualization. There are few works that focus on the linguistic aspect, with notable exceptions [Li et al. 2007; Li et al. 2006] (for the Chinese language). These focus on extracting similar trajector and landmark elements to visualize fable stories. However, their approach is limited to a binary classification label for the trajector role. The landmark is extracted using limited background knowledge instead of a machine learning approach.

Kollar et al. in a recent work [Kollar et al. 2010] presented a system for interaction between humans and robots. The robot follows natural language directions by extracting a sequence of spatial description clauses (SDCs) from the linguistic input and infers the most probable path through the environment given information only about environmental geometry and detected visible objects. Their spatial description clauses contain elements including figure (trajector), verb, spatial relation and landmark. Grounding the flexible spatial language of directions in perception is interesting, but it essentially assumes that directional instructions are given, which renders it to be domain-specific understanding.

None of these works formalized the complete task of domain-independent spatial role labeling using machine learning nor did they pay attention to linguistically motivated features. The idea of defining spatial role labeling is inspired by the more general task of semantic role labeling [Màrquez et al. 2008]. Section 2 discusses some differences that make our task novel. The effects of structured and relational features in this task and the lexical information encouraged the use of relational machine learning methods, such as CRFs. Kollar et al. [Kollar et al. 2010], also used a CRF model (to extract SDCs) but with different settings and feature functions. Because of their capability, general CRFs can model long-distance dependencies, and skip-chain CRFs can be intuitively useful for certain information extraction tasks. Researchers have applied skip-chain CRF's for named entity recognition [Sutton and MacCallum 2006]. However, they have not been used in spatial information extraction. We show that modeling dependencies in the CRF framework benefits the spatial role labeling task.

Apart from cognitive/linguistic models, we point to additional related works that exploit machine learning models in their restricted spatial settings. For example, Reinbergerr [Reinbergerr 2005] presents an unsupervised method to extract spatial relations at the preposition level from text corpora and use the output as preprocessed material to build a virtual environment. They use a shallow parser and select functional relations from which they can extract spatial information. That work manually evaluates the adequacy of the extracted relations. Another work transforms a textual description of a spatial scene in a sequence of prepositions into a graph with objects, annotating local reference systems as nodes and relations as arcs [Claus et al. 1998; Wiebrock et al. 2000]. Inference is realized by multiplying transformation matrices, constraint propagation and verification using machine learning techniques. By assigning values to the parameters and using heuristics for object placement, a visualization of the described spatial layout is generated from the graph. They also consider a limited set of predefined relations.

Spatial relations are also important in semantic image analysis. In one work, eight fuzzy directional relations, such as *right*, *left* and *above*, are supported [Papadopoulos et al. 2006]. All relations are evaluated for each pair of objects in the image. That work presents a learning approach, coupling Support Vector Machines (SVMs) and a Genetic Algorithm (GA), for knowledge-assisted domain-specific semantic image analysis. There are also many challenges in video analysis, handling spatial and temporal relations and extracting those relations from video; an interesting work proposes a framework for learning object and event categories from video [Sridhar et al. 2008]. The work exploits graphical models, and spatio-temporal patterns in the video are represented using an activity graph.

To interpret spatial language for following navigational directions, a system is presented that does not use semantic annotation but instead learns from human demonstration on the Maptask corpus [Vogel and Jurafsky 2010]. In this work, a reinforcement learning setting derives the correspondence between the instruction language and path features. On the same corpus, an earlier work first manually maps the spatial language to conceptual NIUs (navigational information units) [Levit and Roy 2006]. The combination of NIUs is then automatically interpreted as a spatial path using dynamic programming. The linguistic part of NIU extraction is ignored there. The same authors in a recent work start from natural language and map it to SDCs [Kollar et al. 2010].

Moreover, several systems extract information directly from text and determine spatial relationships between objects in a 3D scene to generate such scenes from these textual descriptions. These systems consider the semantic models of spatial relations and their computational implementation. However, they are restricted to simple narratives, often invented by the authors, and do not consider a real corpus. For applying machine learning usually a limited number of relations is defined to keep the problem tractable. A more general overview of older vision and language systems can be found in [Kelleher 2003].

To our knowledge, the current main obstacles to employing machine learning and using this effective approach are the lack of agreement on a unique semantic model for spatial information, the diversity of formal spatial relations, and the consequent lack of annotated data from which machine learning can learn and extract spatial relations. Some research works focus on annotating spatial descriptions in natural languages, specifically for geographical information systems, such as SpatialML [Mani et al. 2008]. To obtain an appropriate annotation scheme, one must investigate and design linguistic and also spatial ontologies to cover the necessary information and to maintain the practical feasibility of automatically annotating unobserved data. However, there is no systematic research on using existing annotations for learning to extract spatial information. Our work here is a first effort in this direction.

Hence, we place our work as an application-oriented investigation that considers abstract spatial elements and the way they are generally expressed in natural language. This abstraction enables processing unrestricted natural language for mapping to formal spatial relations. We define a novel framework for spatial relation extraction by machine learning approaches at the linguistic level. We use a subset of GUM evaluation data for our initial experiments. To our knowledge, no other reported experimental studies have used the same data with machine learning. We also utilize the CLEF corpus in this work, which is partially annotated based on our proposed scheme in [Kordjamshidi et al. 2010b]. We suggest this scheme to cover both primitive concepts of spatial semantic and a mapping to formal spatial ontologies for machine learning approaches. The suggested spatial ontology is small compared to the extensive and expressive ontology of GUM, however mapping between natural language and a formal ontology while exploiting probabilistic relational machine learning approaches is discussed in [Kordjamshidi et al. 2010a].

7. CONCLUSIONS AND FUTURE DIRECTIONS

This paper has mainly defined the *spatial role labeling* task and spatial information extraction from unrestricted natural language using machine learning techniques. We have presented a novel task (spatial role labeling), a set of techniques, and annotated data resources for spatial language understanding. Much of our method's success stems from well-engineered features and state-of-the-art machine learning techniques. The context-dependent classification methods employed, including various CRFs, were particularly useful. We have successfully applied spatial role labeling to texts from the GUM (General Upper Model spatial ontology)-evaluation and CLEF, IAPR TC-12 Image Benchmark datasets. Our extensive experiments have shown that learning the spatial role labeling task is feasible, and the results on individual datasets are promising. We have also tested our learned classifiers on different domains shown that if a classifier is trained on one dataset and tested on another with a different, or less restricted, vocabulary, performance drops accordingly. One part of our methodology, the pipeline, has performed particularly well in such cases. We have experimentally shown that utilizing large, external data resources, such as TPP, is useful when coping with these difficulties, as is splitting the task in two parts: preposition disambiguation and spatial role classification.

Several research directions can extend this work. First, obtaining more data and incorporating additional linguistic elements dealing with focus shift and motion detection are both useful directions. Moreover, using different types of spatial indicators can enhance the range of language constructs that can be handled. An example of the last is the sentence *“The table is supporting the book.”*, in which a verb is the spatial indicator.

Second, our main efforts will focus on using more powerful probabilistic machine learning systems that can address richly structured knowledge representations. The field of probabilistic logic learning [De Raedt et al. 2008] has developed many powerful probabilistic learning techniques, many of which have algorithms that can be applied in our setting. We have performed initial experiments with CRF extensions that can use the (logical) relational knowledge representation, TildeCRF [Gutmann and Kersting 2006]. To fully use such systems, the components in the spatial role labeling task should be defined relationally. The potential to inject background knowledge and probabilistic linguistic constraints in a logical form encourages a move beyond the purely propositional representations usually employed in computational linguistics. Extracted information, i.e., spatial relations, are generally best handled by representation and learning techniques that can explicitly deal with structured data in terms of objects and relations. In this direction, we will exploit various resources which can help to jointly learn smaller related tasks [Andrew et al. 2004]. Additionally, semi-supervised learning and expanding a small data set using latent words and related techniques to cope with the lack of lexical patterns in the training data are a promising direction [Deschacht and Moens 2009].

Third, a related direction of research is using formal knowledge representations to represent spatial information. Using additional reasoning mechanisms, can aid in the extraction process in addition to performing reasoning-after-solution on the extracted spatial information. We recently made the initial steps towards this direction in [Kordjamshidi et al. 2010a; Kordjamshidi et al. 2011].

REFERENCES

ANDREW, G., GRENAGER, T., AND MANNING, C. D. 2004. Verb sense and subcategorization: Using joint inference to improve performance on complementary task. In *Empirical Methods in Natural Language Processing(EMNLP)*. 150–157.

BALDWIN, T., KORDONI, V., AND VILLAVICENCIO, A. 2009. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics* No. 2, 119–149.

BARCLAY, M. AND GALTON, A. 2008. An influence model for reference object selection in spatially locative phrases. In *Spatial Cognition VI: Learning, Reasoning and Talking about Space*, C. Freksa, N. S. Newcombe, P. Gärdenfors, and S. Wölfl, Eds. Number 5241 in Lecture Notes in Artificial Intelligence. Springer, 216–232.

BATEMAN, J., TENBRINK, T., AND FARRAR, S. 2007. The role of conceptual and linguistic ontologies in discourse. *Discourse Processes 44*, 3, 175–213.

BATEMAN, J. A. 2010. Language and space: a two-level semantic approach based on principles of ontological engineering. *International Journal of Speech Technology 13,* 1, 29–48.

BATEMAN, J. A., HOIS, J., ROSS, R., AND TENBRINK, T. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence 174,* 14, 1027–1071.

CARLETTA, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics 22,* 2, 249–254.

CHARNIAK, E. AND JOHNSON, M. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. Association for Computational Linguistics, 173–180.

CLAUS, B., EYFERTH, K., GIPS, C., HRNIG, R., SCHMID, U., WIEBROCK, S., AND WYSOTZKI, F. 1998. Reference frames for spatial inference in text understanding. In *Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*. Springer-Verlag, 241–266.

COHN, A. G. AND RENZ, J. 2008. Chapter 13 qualitative spatial representation and reasoning. In *Handbook of Knowledge Representation*, V. L. Frank van Harmelen and B. Porter, Eds. Foundations of Artificial Intelligence, vol. 3. Elsevier, 551 – 596.

DAHLMEIER, D., NG, H. T., AND SCHULTZ, T. 2009. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Empirical Methods in Natural Language Processing(EMNLP)*. 450–458.

DESCHACHT, K. AND MOENS, M. F. 2009. Semi-supervised semantic role labeling using the latent words language model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*. Association for Computational Linguistics, 21–29.

GALTON, A. 2009. Spatial and temporal knowledge representation. *Journal of Earth Science Informatics 2,* 3, 169–187.

GRUBINGER, M., CLOUGH, P., MÜLLER, H., AND DESELAERS, T. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation (LREC)*.

HERSKOVITS, A. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge, MA.

HOIS, J. AND KUTZ, O. 2008a. Counterparts in language and spacesimilarity and s-connection. In *Proceedings of the 2008 Conference on Formal Ontology in Information Systems: Proceedings of the Fifth International Conference (FOIS 2008)*. 266–279.

HOIS, J. AND KUTZ, O. 2008b. Natural language meets spatial calculi. In *Spatial Cognition*, C. Freksa, N. S. Newcombe, P. Gärdenfors, and S. Wölfl, Eds. Lecture Notes in Computer Science, vol. 5248. Springer, 266–282.

HOWALD, B. AND KATZ, E. 2011. On the explicit and implicit spatiotemporal architecture of narratives of personal experience. In *Spatial Information Theory*, M. Egenhofer, N. Giudice, R. Moratz, and M. Worboys, Eds. Lecture Notes in Computer Science, vol. 6899. Springer Berlin / Heidelberg, 434–454.

JIANG, W., ZAVESKY, E., CHANG, S. F., AND LOUI, A. C. 2008. Cross-domain learning methods for high-level visual concept classification. In *15th IEEE International conference on image processing (ICIP)*. 161–164.

JOHANSSON, R. AND NUGUES, P. 2007. LTH: semantic structure extraction using nonprojective dependency trees. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval '07. 227–230.

KELLEHER, J. D. 2003. A perceptually based computational framework for the interpretation of spatial language. Ph.D. thesis, School of Computing Dublin City University.

KELLEHER, J. D. AND COSTELLO, F. J. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics 35,* 2, 271–306.

KOLLAR, T., TELLEX, S., ROY, D., AND ROY, N. 2010. Toward understanding natural language directions. In *Proceeding of the 5th ACM/IEEE International Conference on Human-robot Interaction*. HRI '10. ACM, 259–266.

KORDJAMSHIDI, P., HOIS, J., VAN OTTERLO, M., AND MOENS, M.-F. 2011. Machine learning for interpretation of spatial natural language in terms of qsr. Poster Presentation at the 10th International Conference on Spatial Information Theory (COSIT'11).

KORDJAMSHIDI, P., VAN OTTERLO, M., AND MOENS, M. F. 2010a. From language towards formal spatial calculi. In *Workshop on Computational Models of Spatial Language Interpretation (CoSLI 2010, at Spatial Cognition 2010)*.

KORDJAMSHIDI, P., VAN OTTERLO, M., AND MOENS, M. F. 2010b. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. European Language Resources Association (ELRA).

LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. Morgan Kaufmann Publishers Inc., 282–289.

LEVIT, M. AND ROY, D. 2006. Interpretation of spatial language in a map navigation task. *IEEE Transactions on Systems, Man and Cybernetics 37,* 3, 667–679.

LI, H., ZHAO, T., LI, S., AND HAN, Y. 2006. The extraction of spatial relationships from text based on hybrid method. *International Conference on Information Acquisition*, 284–289.

LI, H., ZHAO, T., LI, S., AND ZHAO, J. 2007. The extraction of trajectories from real texts based on linear classification. In *Proceedings of NODALIDA 2007 Conference*. 121–127.

LITKOWSKI, K. AND HARGRAVES, O. 2007. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval '07. Association for Computational Linguistics, 24–29.

LOCKWOOD, K., FORBUS, K., HALSTEAD, D. T., AND USHER, J. 2006. Automatic categorization of spatial prepositions. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society. Stressa*. 1705–1710.

LOCKWOOD, K., LOVETT, A., AND FORBUS, K. 2008. Automatic classification of containment and support spatial relations in English and Dutch. In *Spatial Cognition VI: Learning, Reasoning and Talking about Space*, C. Freksa, N. S. Newcombe, P. Gärdenfors, and S. Wölfl, Eds. Number 5241 in Lecture Notes in Artificial Intelligence. Springer, 283–294.

MANI, I., HITZEMAN, J., RICHER, J., HARRIS, D., QUIMBY, R., AND WELLNER, B. 2008. SpatialML: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, Eds. European Language Resources Association (ELRA).

MÀRQUEZ, L., CARRERAS, X., LITKOWSKI, K. C., AND STEVENSON, S. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics 34,* 2, 145–159.

MCINTYRE, N. AND LAPATA, M. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 217–225.

O'HARA, T. AND WIEBE, J. 2009. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics 35,* 2, 151–184.

PAPADOPOULOS, G. T., MEZARIS, V., DASIOPOULOU, S., AND KOMPATSIARIS, I. 2006. Semantic image analysis using a learning approach and spatial context. In *SAMT*, Y. S. Avrithis, Y. Kompatsiaris, S. Staab, and N. E. O'Connor, Eds. Lecture Notes in Computer Science, vol. 4306. Springer, 199–211.

REINBERGERR, M. 2005. Automatic extraction of spatial relations. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. 331–337.

RENZ, J. AND NEBEL, B. 2007. Qualitative spatial reasoning using constraint calculi. In *Handbook of Spatial Logics*, M. Aiello, I. Pratt-Hartmann, and J. van Benthem, Eds. Springer, 161–215.

ROSS, R., SHI, H., VIERHUFF, T., KRIEG-BRÜCKNER, B., AND BATEMAN, J. 2005. Towards dialogue based shared control of navigating robots. In *Proceedings of Spatial Cognition IV: Reasoning, Action, Interaction*. 478–499.

SRIDHAR, M., COHN, A. G., AND HOGG, D. C. 2008. Learning functional object-categories from a relational spatio-temporal representation. In *ECAI*, M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. M. Avouris, Eds. Frontiers in Artificial Intelligence and Applications, vol. 178. IOS Press, 606–610.

STOCK, O., Ed. 1997. *Spatial and Temporal Reasoning*. Kluwer.

SUTTON, C. AND MACCALLUM, A. 2006. Introduction to conditional random fields for relational learning. In *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press.

TAPPAN, D. A. 2004. Knowledge-based spatial reasoning for automated scene generation from text descriptions. Ph.D. thesis, New Mexico State University Las Cruces, New Mexico.

TRATZ, S. AND HOVY, D. 2009. Disambiguation of preposition sense using linguistically motivated features. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*. NAACL '09. Association for Computational Linguistics, 96–100.

VOGEL, A. AND JURAFSKY, D. 2010. Learning to follow navigational directions. In *ACL '10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 806–814.

WIEBROCK, S., WITTENBURG, L., SCHMID, U., AND WYSOTZKI, F. 2000. Inference and visualization of spatial relations. In *Spatial Cognition II*, C. Freksa, C. Habel, W. Brauer, and K. Wender, Eds. Lecture Notes in Computer Science, vol. 1849. Springer Berlin / Heidelberg, 212–224.

ZLATEVL, J. 2007. Spatial semantics. *In Hubert Cuyckens and Dirk Geeraerts (eds.) The Oxford Handbook of Cognitive Linguistics, Chapter 13*, 318–350.