# Spatial Language Understanding with Multimodal Graphs using Declarative Learning based Programming

**Parisa Kordjamshidi** and **Taher Rahgooy** and **Umar Manzoor**
Computer Science Department, Tulane University, New Orleans, USA
{pkordjam,trahgooy,umanzoor}@tulane.edu

## Abstract

This work is on a previously formalized semantic evaluation task of spatial role labeling (SpRL) that aims at extraction of formal spatial meaning from text. Here, we report the results of initial efforts towards exploiting visual information in the form of images to help spatial language understanding. We discuss the way of designing new models in the framework of declarative learning-based programming (DeLBP). The DeLBP framework facilitates combining modalities and representing various data in a unified graph. The learning and inference models exploit the structure of the unified graph as well as the global first order domain constraints beyond the data to predict the semantics which forms a structured meaning representation of the spatial context. Continuous representations are used to relate the various elements of the graph originating from different modalities. We improved over the state-of-the-art results on SpRL.

## 1 Introduction

Spatial language understanding is important in many real-world applications such as geographical information systems, robotics, and navigation when the robot has a camera on the head and receives instructions about grabbing objects and finding their locations, etc. One approach towards spatial language understanding is to map the natural language to a formal spatial representation appropriate for spatial reasoning. The previous research on spatial role labeling (Kordjamshidi et al., 2010, 2017b, 2012) and ISO-Space (Pustejovsky et al., 2011, 2015) aimed at formalizing such a problem and providing machine learning solutions to find such a mapping in a data-driven way (Kordjamshidi and Moens, 2015; Kordjamshidi et al., 2011). Such extractions are made from available textual resources. However, spatial semantics are the most relevant and useful information for visualization of the language and, consequently, accompanying visual information could help disambiguation and extraction of the spatial meaning from text. Recently, there has been a large community effort to prepare new resources for combining vision and language data (Krishna et al., 2017) though not explicitly focused on formal spatial semantic representations. The current tasks are mostly image centered such as image captioning, that is, generating image descriptions (Kiros et al., 2014; Karpathy and Li, 2014), image retrieval using textual descriptions, or visual question answering (Antol et al., 2015). In this work, we consider a different problem, that is, how images can help in the extraction of a structured spatial meaning representation from text. This task has been recently proposed as a CLEF pilot task[1], the data is publicly available and the task overview will be published (Kordjamshidi et al., 2017a). Our interest in formal meaning representation distinguishes our work from other vision and language tasks and the choice of the data since our future goal is to integrate explicit qualitative spatial reasoning models into learning and spatial language understanding.

The contribution of this paper is a) we report results on combining vision and language that extend and improve the spatial role labeling state-of-the-art models, b) we model the task in the framework of declarative learning based programming and show its expressiveness in representing such complex structured output tasks. DeLBP provides the possibility of seamless integration of heteroge-

[1] http://www.cs.tulane.edu/~pkordjam/mSpRL_CLEF_lab.htm

*Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing*, pages 33–43
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

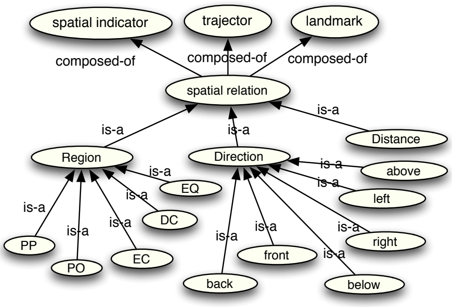

Figure 1: Given spatial ontology (Kordjamshidi and Moens, 2015)

neous data in addition to considering domain ontological and linguistic knowledge in learning and inference. To improve the state-of-the-art results in SpRL and exploiting the visual information we rely on *existing* techniques for continuous representations of image segments and text phrases, and measuring similarity to find the best alignments.

The challenging aspect of this work is that the formal representation of the textual spatial semantics is very different from the raw spatial information extracted from image segments using their geometrical relationships. To alleviate this problem the embeddings of phrases as well as the embeddings of the relations helped connecting the two modalities. This approach helped improving the state of the art results on spatial role labeling (Kordjamshidi et al., 2012) for recognizing spatial roles.

## 2 Problem Description

The goal is to extract spatial information from text while exploiting accompanying visual resources, that is, images. We briefly define the task which is based on a previous formalization of spatial role labeling (SpRL) (Kordjamshidi et al., 2011; Kordjamshidi and Moens, 2015). Given a piece of text, $S$, here a sentence, which is segmented into a number of phrases, the goal is to identify the phrases that carry spatial roles and classify them according to a given set of spatial concepts; identify the links between the roles and form spatial relations (triplets) and finally classify the spatial relations given a set of relation types. A more formal definition of the problem is given in Section 5, where we describe our computational model. The spatial concepts and relation types are depicted in Figure 1 which shows a light-weight spatial ontology. Figure 2 shows an example of an image and the related textual description. The **first level** of this task is to extract spatial roles including,

(a) **Spatial indicators (sp):** these are triggers indicating the existence of spatial information in a sentence;

(b) **Trajectors (tr):** these are the entities whose location are described;

(c) **Landmarks (lm):** these are the reference objects for describing the location of the trajectors.

In the textual description of Figure 2, the location of *kids (trajector)* has been described with respect to the *stairs (landmark)* using the preposition *on (spatial indicator)*. This is example of some spatial roles that we aim to extract from the whole text. The **second level** of this task is to extract spatial relations.

(d) **Spatial relations (sr):** these indicate a link between the three above mentioned roles $(sp.tr.lm)$, forming spatial triplets.

(e) **Relation types:** these indicate the type of relations in terms of spatial calculi formalisms. Each relation can have multiple types.

For the above example we have the triplet $spatial\_relation(kids, on, stairs)$. Recognizing the spatial relations is very challenging because there could be several spatial roles in the sentence and the model should be able to recognize the right links. The formal type of this relation could be $EC$ that is externally connected. The previous research (Kordjamshidi and Moens, 2015) shows the extraction of triplets is the most challenging part of the task for this dataset, therefore we focus on (a)-(d) tasks in this paper. The hypothesis of this paper is that knowing the objects and their geometrical relationships in the companion image might help the inference for the extraction of roles as well as the relations from sentences. In our training dataset, the alignment between the text and image is very coarse-grained and merely the whole text is associated with the image, that is, no sentence alignment, no phrase alignment for segments, etc is available.

Each companion image $I$ contains a number of segments each of which is related to an object and the objects spatial relationships can be described qualitatively based on their geometrical structure of the image. In this paper, we assume the image segments are given and the image object annotations are based on a given object ontology. More-

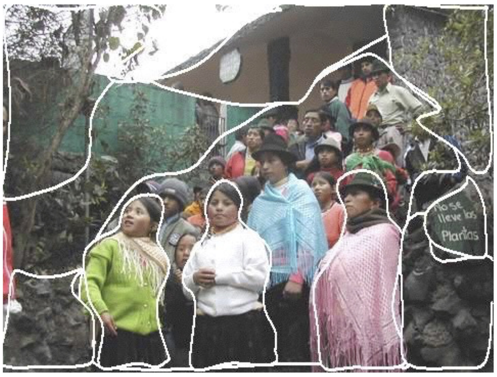

Figure 2: Image textual description:"About 20 kids in traditional clothing and hats waiting on stairs. A house and a green wall with gate in the background. A sign saying that plants can't be picked up on the right."

over, the relationships between objects in the images are assumed to be given. The spatial relationships are obtained by parsing the images and computing a number of relations based on geometrical relationships between the objects boundaries. This implies the spatial representation of the objects in the image is very different from the spatial ontology that we use to describe the spatial meaning from text; this issue makes combining information from images very challenging.

## 3 Declarative Modeling

To extend the SpRL task to a multimodal setting, we firstly, replicated the state-of-the-art models using the framework of declarative learning based programming (DeLBP) *Saul* (Kordjamshidi et al., 2015, 2016). The goal was to extend the previously designed configurations easily and facilitate the integration of various resources of data and knowledge into learning models. In DeLBP framework, we need to define the following building blocks for an application program,

(a) **DataModel:** Declaring a graph schema to represent the domain's concepts and their relationships. This is a first order graph called a *data-model* and can be populated with the actual data instances.

(b) **Learners:** Declaring basic learning models in terms of their inputs and outputs; where the inputs and outputs are properties of the *data-model*'s nodes.

(c) **Constraints:** Declaring constraints among output labels using first order logical expressions.

(d) **Application program:** Specifying the final end-to-end program that starts with reading the raw data into the declared *data-model* graph referred to as data population and then calls the learners and constrained learners for training, prediction and evaluation.

Each application program defines the configuration of an end-to-end model based on the above-mentioned components. In the following sections we describe these components and the way they are defined for multimodal spatial role labeling.

### 3.1 Data Model

A graph is used to explicitly represent the structure of the data. This graph is called the *data-model* and contains *typed* nodes, edges and properties. The node types are domain's basic data structures, called *base types*. The base types are mostly pre-established in *Saul* (Kordjamshidi et al., 2016), including base types for representing documents, sentences, phrases, etc, referred to as *linguistic units*. In this work, we also have added a set of preliminary image base types in *Saul* that could be extended to facilitate working on visual data task-independently in the future. The below code shows a data model schema including nodes of linguistic units and image segments. The *typed* nodes are declared as follows:

```
val documents = node[Document]
val sentences = node[Sentence]
val tokens = node[Token]
val phrases = node[Phrase]
val pairs = node[Relation]
val images = node[Image]
val segments = node[Segement]
val segmentPairs = node[SegmentRelation]
```

('val' is a Scala keyword to define variables; documents, sentences, etc, are the programmer-defined variables; 'node' is a Saul keyword to define typed graph nodes; Document, Sentence, etc, are the NLP and other base types built-in for Saul.)

Given the base types, domain sensors can be used to populate raw data into the *data-model*. Sensors are black box functions that operate on base types and can generate instances of nodes, properties and edges. An edge that connects documents to sentences using a sensor called 'documentToSentenceMatching' is defined as:

```
val documentTosentence = edge(documents,
    sentences)
documentTosentence.addSensor(
    documentToSentenceMatching_)
```

('edge' is a Saul keyword, addSensor is a Saul function)

The properties are assigned to the graph nodes only and defined based on the existing domain property sensors. The following example receives a phrase and returns the dependency relation label of its head:

```
val headDependencyRelation = property(
    phrases){x =>  getDependencyRelation
    (getHeadword(x))}
```

('property' is a keyword in Saul, getDependencyRelation and getHeadword are two NLP sensors applied on words and phrases respectively.)

## 3.2 Learners

The learners are basically a set of classic classifiers each of which is related to a target variable in the output space. The output variables are a subset of elements represented in the ontology of Figure 1. The previous work shows the challenging element of the ontology is the extraction of spatial triplets. Therefore, in this work our goal is to improve the extraction of the roles and spatial triplets. Each classifier/learner is applied on a typed node which is defined in the *data-model*. For example, a *trajector role classifier* is applied on the `phrase` nodes and defined as follows:

```
object TrajectorRoleClassifier extends
    Learnable(phrases) {
    def label = trajectorRole
    override lazy val classifier = new
    SparseNetworkLearner
    override def feature = using(
    headDependencyRelation,...)}
```

('label' is a Saul function to define the output of the classifier, 'trajectorRole' is a name of a property in the datamodel to be predicted. 'feature' is a Saul function to define the list of properties of the datamodel to be used as input features.)

All other learners are defined similarly and they can use different types of *data-model* properties as 'feature's or as 'label'. In our proposed model, only the role and pair classifiers are used and triplets of relations are generated based on the results of the pair classifiers afterwards.

### 3.2.1 Role and Relation Properties

Spatial Roles are applied on phrases and most of the features are used based on the previous works (Kordjamshidi and Moens, 2015), however the previous work on this data is mostly token-based; we have extended the features to phrase-based and added some more features. We use linguistically motivated features such as lexical form of the words in the phrases, lemmas, pos-tags, dependency relations, subcategorization, etc. These features are used sometimes based on the head-word of the phrases and sometimes by concatenation of the same features for all the tokens in a phrase. The relations are, in-fact, a pair of phrases and the pair features are based on the features of the phrases. The relational features between two phrases include their path, distance, before/after information. In addition to the previously used features, here, we add phrase and image embeddings described in the next section. The details of the linguistic features are omitted due to the lack of space and since the code is publicly available.

### 3.2.2 Image and Text Embeddings

Using continuous representations has several advantages in our models. One important aspect is compensating for the lack of lexical information due to the lack of training data for this problem. Another aspect is the mapping between image segments and the phrases occurring in the textual descriptions and establishing a connection between the two modalities. The experiments show these components improve the generalization capability of our trained models. Since our dataset is very small, our best embeddings were the commonly used word2vec (Pennington et al., 2014) model trained over google's gigaword+wikipedia corpora.

**Text Embeddings.** We generate the embeddings for candidate roles. More specifically, for each phrase we find its syntactic head and then we use the vector representation of the syntactic head as a feature of the phrase. This is added to the rest of linguistically motivated features.

**Image Embeddings.** For the image side we rely on a number of assumptions given the type of image corpora available for our task. As mentioned in Section 1, the input images are assumed to be segmented and the segments have been labeled according to a given ontology of concepts. For example, the ontology for a specific object like *Bush* can be `entity->landscape-nature->vegetation->trees->bush`. Given the image segments, the spatial relations between segments are automatically extracted in a pairwise exhaustive manner using the geometrical properties of the segments (Escalante et al., 2010). These relations are limited to relationships such as besides, disjoint, below, above, x-aligned, and y-aligned. In this work, we employed the pre-processed images that were publicly available[2]. Since the segment label ontology is independent from the textual descriptions, finding the alignment between the segments and the words/phrases in the text is very challenging. To alleviate this problem, we exploit the embeddings of the image segment labels using the same representations that is used for words in the text. We measure the similarity between the segment label embeddings and word embeddings to help the fine-grained alignments between the image segments and text phrases. To clarify, we tried the following variations: we compute the word embeddings of image segment labels and words in the text candidate phrases, then we find the most similar object in the image to each candidate phrase. We use the embedding of the most similar object as a feature of the phrase. Another variation that we tried is to exploit the embeddings of the image segment ontologies. The vector representation of each segment label is computed by averaging over the representation of all the ontological concepts related to that segment.

### 3.3 Global Constraints

The key point of considering global correlations in our extraction model is formalizing a number of global constraints and exploiting those in learning and inference. The constraints are declared using first order logical expressions, for example, the constraint, ”if there exists a trajector or a landmark in the sentence then an indicator should also exist in the sentence” , we call it integrity constraint and it is expressed as follows:

```
((sentences(s)~>phraseEdge)._exists{x:
    Phrase=>(TrajectorRoleClassifier on
    x is "Trajector") or (
    LandmarkRoleClassifier on x is "
    Landmark"}))==>((sentences(s)~>
    phraseEdge)._exists{y:Phrase=>
    IndicatorRoleClassifier on y is "
    Indicator"})
```

The domain knowledge is inspired from this work (Kordjamshidi and Moens, 2015).[3] The first order constraints are automatically converted to linear algebraic constraints for our underlying computational models.

## 4 Application program

Using the building blocks of a DeLBP including a *data-model*, *learners* and *constraints*, we are able to design various end-to-end configurations for learning and inference. The first step for an application program is to populate the annotated corpus in the graph schema, that is, our declared *data-model*. To simplify the procedure of populating the graph with linguistic annotations, we have established a generic XML reader that is able to read the annotated corpora from XML into the Saul *data-model* and provide us a populated graph. The nodes related to the linguistic units (i.e. sentence, phrase, etc) are populated with the annotations as their properties. The population can be done in various ways, for example, `SpRLDatamodel.documents.populate (xmlReader.documentList())` reads the content of DOCUMENT tag or its pre-defined[4] equivalent into `documents` nodes in the *data-model*. Populating documents can lead to populating all other types of nodes such as sentences, tokens, etc if the necessary sensors and edges are specified beforehand. Saul functions and *data-model* primitives can be used to make graph traversal queries to access any information that we need from either image or text for candidate selection, feature extraction.

The feature extraction includes segmentation of the text and candidate generation for roles and pair relations. Not all tokens are candidates for playing trajector roles, most certainly verbs will not play this role. After populating the data into the graph we program the training strategy. We have the possibility of training for each concept independently, that is, each declared classifier can call the `learn` , for example, `trajectorClassifier.learn()`. However, the independently trained classifiers can exploit the global constraints like the one we defined in Section 3.3 and be involved in a global inference jointly with other role and relation classifiers. Such a model is referred to as L+I (Punyakanok et al., 2008). Moreover, the parameters of the declared classifiers can be trained jointly and for this purpose we need to call `joinTrain` and pass the list of classifiers and the constraints to be used together. We use L+I models in this paper due to the efficiency of the training.

## 5 Computational Model

The problem we address in this paper is formulated as a structured prediction problem as the out-

[2] http://www.imageclef.org/photodata

[3] constraints code is available on GitHub.

[4] The programmer is able to specify the tags that are related to the base types before reading the xml.

put contains a number of spatial roles and relationships that obey some global constraints. In learning models for structured output prediction, given a set of $N$ input-output pairs of training examples $E = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} : i = 1..N\}$, we learn an objective function $g(x, y; W)$ which is a linear *discriminant* function defined over the combined feature representation of the inputs and outputs denoted by $f(x, y)$ (Ioannis Tsochantaridis and Altun, 2006):

$$g(x, y; W) = \langle W, f(x, y) \rangle. \tag{1}$$

$W$ denotes a weight vector and $\langle , \rangle$ denotes a dot product between two vectors. A popular discriminative training approach is to minimize the following convex upper bound of the loss function over the training data:

$$l(W) = \sum_{i=1}^{N} \max_{y \in \mathcal{Y}} (g(x^i, y; W) - g(x^i, y^i; W) \quad + \Delta(y^i, y)),$$

the inner maximization is called loss-augmented inference and finds the so called most violated constraints/outputs ($y$) per training example. This is the base of inference-based-training models (IBT). However, the inference over structures can be limited to the prediction time which is known as learning plus inference (L+I) models. L+I uses the independently trained models (this is known as piece-wise training as well (Sutton and McCallum, 2009)) and has shown to be very efficient and competitive compared to IBT models in various tasks (Punyakanok et al., 2005). Given this general formalization of the problem we can easily consider both configurations of L+I and IBT using a declarative representation of our inference problem as briefly discussed in Section 4. We define our structured model in terms of first order constraints and classifiers.

Here in *Saul*'s generic setting, inputs $\boldsymbol{x}$ and outputs $\boldsymbol{y}$ are sub-graphs of the *data-model* and each learning model can use parts and substructures of this graph. In other words, $\boldsymbol{x}$ is a set of nodes $\{x_1, \ldots, x_K\}$ and each node has a *type* $p$. Each $x_k \in \boldsymbol{x}$ is described by a set of properties; this set of properties will be converted to a feature vector $\phi_p$. Given the multimodal setting of our problem, $x_i$ 's can represent segments of an image or various linguistic units of a text, such as a phrase (*atomic node*) or a pair of phrases (*composed node*), and each type is described by its own properties (e.g. a phrase by its headword, the pair by the distance of the two headwords, an image segment by the vector representation of its concept). We refer to the text-related nodes and image-related nodes differently as $x_T$ and $x_I$, respectively. The goal is to map this pair to a set of spatial objects and spatial relationships, that is $f : (x_T, x_I) \mapsto y$.

The output $\boldsymbol{y}$ is represented by a set of *labels* $\boldsymbol{l} = \{l_1, \ldots, l_P\}$ each of which is a *property* of a node. The labels can have semantic relationships. In our model the set of labels is $l = \{tr, lm, sp, sp.tr, sp.lm, sp.tr.lm\}$. Note that these labels are applied merely to the parts of the text, $tr$, $lm$ and $sp$ are applied on the phrase of a sentence, $sp.tr$ and $sp.lm$ are applied on pairs of phrases in the sentence, and finally $sp.tr.lm$ is applied on triplets of phrases. According to the terminology used in (Kordjamshidi and Moens, 2015), the labels of atomic components of the text (here phrases) are referred to as *single-labels* and the labels that are applied to composed components of the input such as pairs or triplets are referred to as *linked-labels*. These labels help to represent $y$ with a set of indicator functions that indicate which *segments* of the sentence play a specific spatial role and which are involved in relations. The labels are defined with a graph query that extracts a property from the *data-model*. The $l_p(x_k)$ or shorter $l_{pk}$ denotes an indicator function indicating whether component $x_k$ has the label $l_p$. For example, $sp(on)$ shows whether $on$ plays a spatial role and $sp.tr(on, kids)$ shows whether $kids$ is a trajector of $on$. As expected, the form of the output is dependent on the input since we are dealing with a structured output prediction problem. In our problem setting the spatial roles and relations are still assigned to the components of the text and the connections, similarities and embeddings from image are used as additional information for improving the extractions from text.

The main objective $g$ is written in terms of the instantiations of the feature functions, labels and their related blocks of weights $\boldsymbol{w}_p$ in $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_P]$,

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) &= \sum_{l_p \in \boldsymbol{l}} \sum_{x_k \in C_{l_p}} \langle \boldsymbol{w}_p, f_p(x_k, l_p) \rangle \qquad (2) \\ &= \sum_{l_p \in \boldsymbol{l}} \sum_{x_k \in C_{l_p}} \langle \boldsymbol{w}_p, \phi_p(x_k) \rangle \, l_{pk} \\ &= \sum_{l_p \in \boldsymbol{l}} \left\langle \boldsymbol{w}_p, \sum_{x_k \in C_{l_p}} (\phi_p(x_k) l_{pk}) \right\rangle, \end{aligned}$$

where $f_p(\boldsymbol{x}_k, l_p)$ are the local joint feature vector for each candidate $\boldsymbol{x}_k$. This feature vector is computed by scalar multiplication of the input feature vector of $\boldsymbol{x}_k$ (i.e. $\phi_p(\boldsymbol{x}_k)$), and the output label $l_{pk}$.

Given this objective, we can view the inference task as a *combinatorial constrained optimization* given the polynomial $g$ which is written in terms of labels, subject to the constraints that describe the relationships between the labels (either single or linked labels). For example, the *is-a* relationships can be defined as the following constraint, $(l(\boldsymbol{x}_c)\ is\ 1) \Rightarrow (l'(x_c)\ is\ 1)$, where $l$ and $l'$ are two distinct labels that are applicable on the node with the same type of $\boldsymbol{x}_c$. These constraints are added as a part of *Saul*'s objective, so we have the following objective form, which is in fact a constrained conditional model (Chang et al., 2012), $g = \langle \boldsymbol{w}, f(\boldsymbol{x}, \boldsymbol{y})\rangle - \langle \rho, c(\boldsymbol{x}, \boldsymbol{y})\rangle$, where $c$ is the constraint function and $\rho$ is the vector of penalties for violating each constraint. This representation corresponds to an integer linear program, and thus can be used to encode any MAP problem. Specifically, the $g$ function is written as the sum of local joint feature functions which are the counterparts of the probabilistic factors:

$$g(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \sum_{l_p \in \boldsymbol{l}} \sum_{\boldsymbol{x}_k \in \{\tau\}} \langle \boldsymbol{w}_p, f_p(\boldsymbol{x}_k, l_{pk})\rangle \\ + \sum_{m=1}^{|C|} \rho_m c_m(\boldsymbol{x}, \boldsymbol{y}), \tag{3}$$

where $C$ is a set of global constraints that can hold among various types of nodes. $g$ can represent a general scoring function rather than the one corresponding to the likelihood of an assignment. Note that this objective is automatically generated based on the high level specifications of learners and constraints as described in Section 3.

## 6 Experimental Results

In this section, we experimentally show the influence of our new features, constraints, phrase embeddings and image embeddings and compare them with the previous research.

**Data.** We use the SemEval-2012 shared tasks data (Kordjamshidi et al., 2012) that consists of textual descriptions of 613 images originally selected from the IAPR TC-12 dataset (Grubinger et al., 2006), provided by the *CLEF* organization. In the previous works only the text part of this data has been used in various shared task settings (Kordjamshidi et al., 2012; Oleksandr Kolomiyets and Bethard, 2013; Pustejovsky et al., 2015) and with a variation in the annotation schemes. This data includes about 1213 sentence containing 20,095 words with 1706 annotated relations. We preferred this data compared to more recent related corpora (Pustejovsky et al., 2015; Oleksandr Kolomiyets and Bethard, 2013) for two main reasons. First is the availability of the aligned images and the second is the static nature of the most spatial descriptions.

**Implementation.** As mentioned before, we used *Saul* (Kordjamshidi et al., 2015, 2016) framework that allows flexible relational feature extraction as well as declarative formulation of the global inference. We extend *Saul*'s basic data structures and sensors to be able to work with multimodal data and to populate raw as well as annotated text easily into a *Saul* multimodal *data-model*. The code is available in Github.[5] We face the following challenges when solving this problem: the training data is very small; the annotation schemes for the text and images are very different and they have been annotated independently; the image annotations regarding the spatial relations include very naively generated exhaustively pairwise relations which are not very relevant to what human describes by viewing the images. We try to address these challenges by feature engineering, exploiting global constraints and using continuous representations for text and image segments. We report the results of the following models in Table 1.

**BM:** This is our baseline model built with extensive feature engineering as described in Section 3.2.1. We train independent classifiers for the roles and relations classification in this model;

**BM+C:** This is the BM that uses global constraints to impose, for example, the integrity and consistency of the assignments of the roles and relation labels at the sentence level.

**BM+C+E:** To deal with the lack of lexical information, the features of roles and relations are augmented by w2vec word embeddings, the results of this model without using constraints (BM+E) are reported too;

**BM+E+I+C:** In this model in addition to text embeddings, we augment the text phrase fea-

[5] `https://github.com/HetML/SpRL`

|  | Trajector |  |  | Landmark |  |  | Spatial indicator |  |  | Spatial triplet |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 |
| BM | 56.72 | 69.57 | 62.49 | 72.97 | 86.21 | 79.05 | 94.76 | 97.74 | 96.22 | 75.18 | 45.47 | 56.67 |
| BM+C | 65.56 | 69.91 | 67.66 | 77.74 | 87.78 | 82.46 | 94.83 | 96.86 | 95.83 | 75.21 | 48.46 | 58.94 |
| BM+E | 55.87 | 77.35 | 64.88 | 71.47 | 89.18 | 79.35 | 94.76 | 97.74 | 96.22 | 66.50 | 57.30 | 61.56 |
| BM+E+C | 64.40 | 76.77 | 70.04 | 76.99 | 89.35 | 82.71 | 94.85 | 97.48 | 96.15 | 68.34 | 57.93 | **62.71** |
| BM+E+I | 56.53 | 79.29 | 66.00 | 71.78 | 87.44 | 78.84 | 94.76 | 97.74 | **96.22** | 64.12 | 57.08 | 60.39 |
| BM+E+I+C | 64.49 | 77.92 | **70.57** | 77.66 | 89.18 | **83.02** | 94.87 | 97.61 | 96.22 | 66.46 | 57.61 | 61.72 |
| BM+E+C-10f | 78.49 | 77.67 | 78.03 | 86.43 | 88.93 | 87.62 | 91.70 | 94.71 | 93.17 | 80.85 | 60.23 | 68.95 |
| SOP2015-10f | - | - | - | - | - | - | 90.5 | 84 | 86.9 | 67.3 | 57.3 | 61.7 |
| SemEval-2012 | 78.2 | 64.6 | 70.7 | 89.4 | 68.0 | 77.2 | 94.0 | 73.2 | 82.3 | 61.0 | 54.0 | 57.3 |

Table 1: Experimental results on SemEval-2012 data including images. BM: Baseline Model, C: Constraints, E: Text Embeddings, I: Image Embeddings.

tures with the embeddings of the most similar image segments. The version without constraints is denoted as BM+E+I.

**SemEval-2012:** This model is the best performing model of SemEval-2012 (Roberts and Harabagiu, 2012). It generates the candidate triplets and classifies them as spatial/not-spatial. It does an extensive feature extraction for the triplets. The roles then are simply inferred from the relations. The results are reported with the same train/test split.

**SOP2015-10f:** This model is an structured output prediction model that does a global inference on the whole ontology including the prediction of relations and relation types (Kordjamshidi and Moens, 2015).

The experimental results in Table 1 show that adding constraints to our baseline and other model variations consistently improves the classification of trajectors and landmarks dramatically although it slightly decreases the F1 of spatial indicators in some cases. Adding word embeddings (BM+C+E) shows a significant improvement on roles and spatial relations. The results on BM+E+I+C show that image embeddings improves trajectors and landmarks compared to BM+E+C, though the results of triples are slightly dropped (62.71 → 61.72).

Our results exceed the state of the art models reported in SemEval-2012 (Kordjamshidi et al., 2012). The SemEval-2012 best model uses same train/test split as ours (Roberts and Harabagiu, 2012). The results of the best performing model in (Kordjamshidi and Moens, 2015), SOP2015-10f, are lower than our best model in this work. Although that model uses structured training but here the embeddings make a significant improvement. While SOP2015-10f performance results on triples, spatial indicators, pairs of trajector and landmarks with indicators have been reported, there is no reports on trajecotrs and landmarks prediction accuracy as designated independent roles –those are left empty in the table. There are some differences in our evaluation and the previous systems evaluations.The SOP2015-10f is evaluated by 10-fold cross validation rather than the train/test split. To be able to compare, we report the 10-fold cross validation results of our best model BM+E+C and refer to it as BM+E+C-10f in Table 1 which is outperforming other models. Note that the folds are chosen randomly and might be different from the previous evaluation setting. Another difference is that our evaluation is done phrase-based and overlapping phrases are counted as true predictions. The SemEval-2012 and SOP2015-10f models operate on classifying tokens/words which are the headwords of the annotated roles. However, our identified phrases cover the headwords of role (trajectors and landmarks) phrases with 100% and for spatial indicators 98% which keeps the comparisons fair yet.

Our results exceed the stat-of-the-art models significantly. Both word and image embeddings help expanding our semantic dimensions for spatial objects but interestingly the spatial indicators can not be improved using embeddings. Since the indicators are mostly prepositions, it seems capturing the semantic dimensions of prepositions using continuous vectors is harder than other lexical categories such as nouns and verbs. This is even worse when we use images since the terminology of the relations in the images is very different from the way the relations are expressed in the language using prepositions. Though the improvement on objects can improve the relations but it will be interesting to investigate how the semantics of the relations can be captured using richer representations for spatial prepositions. A possible direction for our work could have been to train deep models that map the images to the formal semantic representations of the text's content, however for training such models using only 2013 sentences related to about 600 images will not be feasible. The existing large corpora which contain image and text, do not contain formal semantic annotation with the textual description. Dealing with this problems remains as our future work.

## 7 Related Research

This work can be related to many research works from various perspectives. However, for the sake of both clarity and conciseness, we limit our exploration in this section to two research directions. First body of related work is about the specific SpRL task that we are solving. This direction is aiming at obtaining a generic formal spatial meaning representation from text. The second body of the work is about combining vision and language which itself has a large research community around it recently and has turned to a hot topic.

Several research efforts in the past few years aimed at defining a framework for the extraction of spatial information form natural language. These efforts start from defining linguistic annotation schemes (Pustejovsky and Moszkowicz, 2008; Kordjamshidi et al., 2010; Pustejovsky and Moszkowicz, 2012; jeet Mani, 2009), annotating data and defining tasks (Kordjamshidi et al., 2012; Oleksandr Kolomiyets and Bethard, 2013; Pustejovsky et al., 2015) to operate on the annotated corpora and learn extraction models. However, there exists, yet, a large gap between the current models and the ones that can perform reasonably well in practice for real world applications in various domains. Though we follow that line of work, we aim at exploiting the visual data in improving spatial extraction models. We exploit the visual information accompanying the text which is mostly available nowadays. We aim at text understanding while assuming that the text highlights the most important information that might be confirmed by the image. Our goal is to use the image to recognize and disambiguate the spatial information from text.

Our work is very related to the research done by computer vision community and in the intersection of vision and language. There are many progressive research works on generating image captions (Karpathy and Li, 2014), retrieving images and visual question answering (Antol et al., 2015). However the center of attention has been understanding images. Here, our aim is to exploit the images for text understanding.This task is as challenging as the former ones or even more challenging because among the many possible objects and relationships in the image a very small subset of those are important and have been expressed in the text. Therefore the available visual corpora are not exactly the type of the data that can be used to train supervised models for our task though it could provide some indirect supervision particularly for having a unified semantic representation of spatial objects (Ludwig et al., 2016).

This work can be improved by exploiting external models and corpora (Pustejovsky and Yocum, 2014) but this will remain for our future investigation. Our task can benefit from the research performed on reference resolution that targets identifying the objects in the image that are mentioned in the text (Schlangen et al., 2016). Having a high-quality alignment by training explicit models for resolving references should help recognizing the spatial objects mentioned in the text and the type of spatial relations according to the image. Explicit reference resolution between modalities in dialogue systems are also inspiring (Fang et al., 2014). In the mentioned reference a graph representation of the scene is gradually made by machine based on the grasped static visual information and the representation is corrected and completed dynamically as the dialogue between the machine and human is going on. However, in this work there is no learning component and there is no spatially annotated data to be used for our goal of formal spatial meaning representation for a generic text.

In this work we take a small step and investigate the ways to integrate information from both modalities for our textual extraction target. Our results are compared to the previous work (Kordjamshidi and Moens, 2015) that exploit the text part of the same spatially annotated corpora and improve the results when exploiting the accompanying images.

## 8 Conclusion

In this paper, we deal with the problem of spatial role labeling which targets at mapping natural language text to a formal spatial meaning representation. We use the information from accompanying segmented images to improve the spatial role extractions. Although, there are many recent research on combining vision and language, none of them consider obtaining a formal spatial meaning representation as a target while we do and our approach will be helpful for adding explicit reasoning component to the learning models in the future. We manifest the expressivity of declarative learning based programming paradigm for designing global models for this task. We put both the image and text related to a scene in a unified *datamodel* graph and use them as structured learning examples. We extract features by traversing the graph and using the continuous representations to connect the image segment nodes to the nodes related to the text phrases. We exploit the continuous representation to align the similar concepts in the two modalities. We exploit global first order constraints for global inference over roles and relations. Our models improve the state of the art results on previous spatial role labeling models.

## References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine Learning* 88(3):399–431.

Hugo Jair Escalante, Carlos A. Hernndez, Jesus A. Gonzalez, A. Lpez-Lpez, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseor, and Michael Grubinger. 2010. The segmented and annotated {IAPR} tc-12 benchmark. *Computer Vision and Image Understanding* 114(4):419 – 428. Special issue on Image and Video Retrieval Evaluation. https://doi.org/http://doi.org/10.1016/j.cviu.2009.03.008.

Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2014. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'14, pages 1544–1550.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. pages 13–23.

Thorsten Joachims Thomas Hofmann Ioannis Tsochantaridis and Yasemin Altun. 2006. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6(2):1453–1484.

jeet Mani. 2009. SpatialML: annotation scheme for marking spatial expression in natural language. Technical Report Version 3.0, The MITRE Corporation.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR* abs/1412.2306.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR* abs/1411.2539.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*. volume 2, pages 365–373.

Parisa Kordjamshidi, Daniel Khashabi, Christos Christodoulopoulos, Bhargav Mangipudi, Sameer Singh, and Dan Roth. 2016. Better call saul: Flexible programming for learning and inference in nlp. In *Proc. of the International Conference on Computational Linguistics (COLING)*.

Parisa Kordjamshidi and Marie-Francine Moens. 2015. Global machine learning for spatial ontology population. *Web Semant.* 30(C):3–21. https://doi.org/10.1016/j.websem.2014.06.001.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017a. CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview. In Julio Gonzalo Liadh Kelly Lorraine Goeuriot Thomas Mandl Linda Cappellato Nicola Ferro Gareth J. F. Jones, Samus Lawless, editor, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. Springer, volume 10456 of *LNCS*.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2010. Spatial role labeling: task definition and annotation scheme. In Nicoletta Calzolari, Choukri Khalid, and Maegaard Bente, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. pages 413–420.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing* 8:1–36.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2017b. Spatial role labeling annotation scheme. In N. Ide James Pustejovsky, editor, *Handbook of Linguistic Annotation*, Springer Verlag.

Parisa Kordjamshidi, Hao Wu, and Dan Roth. 2015. Saul: Towards declarative learning based programming. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* .

Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. 2016. Deep embedding for spatial role labeling. *CoRR* abs/1603.08474.

Parisa Kordjamshidi Marie-Francine Moens Oleksandr Kolomiyets and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA, pages 255–262.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Vasin Punyakanok, Dan Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 34(2).

Vasin Punyakanok, Dan Roth, W. Tau Yih, and D. Zimak. 2005. Learning and inference over constrained output. In *IJCAI'05*. pages 1124–1129.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. SemEval-2015 task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 9th international workshop on semantic evaluation (SemEval 2015), Denver, Colorado, 4-5 June 2015*. ACL, pages 884–894.

James Pustejovsky, J. Moszkowicz, and M. Verhagen. 2011. ISO-Space: The annotation of spatial information in language. In *ACL-ISO International Workshop on Semantic Annotation (ISA'6)*.

James Pustejovsky and J. L. Moszkowicz. 2008. Integrating motion predicate classes with spatial and temporal annotations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008: Companion volume D, Posters and Demonstrations*. pages 95–98.

James Pustejovsky and J. L. Moszkowicz. 2012. The role of model testing in standards development: The case of ISO-space. In *Proceedings of LREC'12*. European Language Resources Association (ELRA), pages 3060–3063.

James Pustejovsky and Zachary Yocum. 2014. Image annotation with iso-space: Distinguishing content from structure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Kirk Roberts and S.M. Harabagiu. 2012. UTD-SpRL: A joint approach to spatial role labeling. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval'12)*. pages 419–424.

David Schlangen, Sina Zarrie, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Charles Sutton and Andrew McCallum. 2009. Piecewise training for structured prediction. *Machine Learning* 77:165–194. https://doi.org/10.1007/s10994-009-5112-z.