Spatial Role Labeling with Convolutional Neural Networks

Alexey Mazalov
University of Lisbon, IST and INESC-ID
Lisboa, Portugal
a_a_mazalov@mail.ru

Bruno Martins
University of Lisbon, IST and INESC-ID
Lisboa, Portugal
bruno.g.martins@ist.utl.pt

David Matos
University of Lisbon, IST and INESC-ID
Lisboa, Portugal
david.matos@inesc-id.pt

ABSTRACT

Many natural language processing applications require information about the spatial locations of objects referenced in text, or spatial relations between these objects in space. For example, the phrase *a book on the shelf* contains information about the location of the object *book*, corresponding to a trajector, with respect to the object *shelf*, which in turn corresponds to a landmark. Spatial role labeling concerns with the task of automatically processing textual sentences and identifying objects of spatial scenes and relations between them. In this paper, we describe the application of modern machine learning methods to extract spatial roles and their relations, specifically by adapting a pre-existing system based on a convolutional neural network architecture that has been recently proposed for the more general task of semantic role labeling. We report on experiments with datasets from the SemEval challenges on spatial role labeling, showing that our method can achieve results in line with the current state-of-the-art. We therefore argue that that spatial role labeling can leverage on recent developments in semantic role labeling, requiring only minimal adaptations.

CCS Concepts

•**Computing methodologies** → *Natural language processing; Machine learning;* •**Information systems** → *Spatial-temporal systems;*

Keywords

Spatial Semantics, Spatial Role Labeling, Natural Language Processing, Convolutional Neural Networks

1. INTRODUCTION

While unconstrained natural language is an intuitive and flexible way for humans to communicate, the automated understanding of linguistic inputs remains a particularly challenging problem, because diverse words and phrases must be mapped into structures that a machine can process, and because elements in those structures must be grounded into unambiguous interpretations. Computational models for understanding (geo)spatial language, in particular, are a cardinal issue in multiple disciplines, and they can support critical applications related to robotics, navigation, and natural language understanding in general.

Many different types of natural language constructs can be used to express relational structures of objects, spatial relations between them, and patterns of motion through space relative to some reference point. Understanding such spatial utterances can be cast as a statistical natural language processing problem. Analogous to semantic role labeling [3, 10, 5], the task of spatial role labeling [17] has been defined as that of automatically labeling words and phrases, in a sentence, with a set of spatial roles such as trajector, landmark, spatial indicator, distance, direction, etc. More specifically, the task involves identifying and classifying spatial arguments that are triggered by spatial expressions mentioned in a sentence and establishing relations between them with attributes, according to the theory of holistic spatial semantics [29] and with the intent of covering all aspects of spatial concepts, including both static and dynamic spatial relations. Properly addressing the spatial role labeling task has many important applications related to natural language understanding. In the past, spatial role labeling has for instance been used in applications related to biomedical text mining [16] or image understanding [24], and many other use cases have also been envisioned (e.g., mining spatial commonsense [27]).

Within the context of the 2012 and 2013 editions of the Semantic Evaluation (SemEval) initiative [14, 13], the Spatial Role Labeling evaluation task extended on previous work in the area [17]. The organizers introduced an annotation scheme and a set of labeled documents, which can facilitate the application of machine learning techniques. In this paper, we specifically report on a set of experiments concerning with spatial role labeling, for which we adapted a state-of-the-art learning-based system named `nlpnet`, originally developed for semantic role labeling [9]. Our experimental results confirmed that previous work focused on the semantic role labeling task can easily be adapted into this new problem domain, as we obtained results that are in line with those reported at the SemEval competitions, requiring only some minimal adaptations. Future work in the area can thus leverage on advances within the general area of semantic role labeling [19, 28].

The rest of this paper is organized as follows: Section 2 presents a formal definition for the spatial role labeling problem, following the guidelines from the task at SemEval-2013. Section 3 describes the proposed method, resulting from the adaptation of a system for the more general problem of semantic role labeling, based on convolutional neural networks. Section 4 describes the experimental validation of the adapted system, detailing the methodology and the obtained results. Section 5 presents a brief summary of previous research in the area. Finally, Section 6 presents our main conclusions, and discusses possible directions for future work.

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837706

2. SPATIAL ROLE LABELING

In natural language discourse, a spatial relation between two objects is usually implemented by a preposition (e.g., *in*, *on*, *at*, etc.) or a prepositional phrase (e.g., *on top of*, *inside of*, etc.). Earlier work on spatial role labeling leveraged the notions of trajectors, landmarks and spatial indicators, for instance as introduced by Kordjamshidi et al. [17] with basis on the theory of holistic spatial semantics [29]. SemEval-2012 introduced an evaluation challenge focusing on spatial role labeling, considering the annotation of static spatial relations according to the aforementioned roles [14]. At SemEval-2013, these roles were expanded to include motion indicators, paths, directions and distances, in order to capture fine-grained spatial semantics of static spatial relations, and also to accommodate dynamic spatial relations [13].

Static spatial relations are defined as relationships between still objects, where one object (i.e., the trajector) plays a central role in the spatial scene, and the second one (i.e., the landmark) plays a secondary role. These spatial relations can also be grouped into the domains of topological (i.e., region), directional or distance relations. In the proposed scheme for annotating spatial relations expressed in natural language sentences, static spatial relations are thus defined as a tuple that contains the spans of text corresponding to a trajector, a landmark and a spatial indicator that acts as the pivot for the spatial relation. The spatial indicator can encode properties such as region containment, direction (i.e., the spatial arrangement between a trajector and a landmark), or the distance between a trajector and a landmark. The following examples illustrate how static spatial relations should be annotated, for some particular sentences.

- **Sentence :** I saw an elephant in the zoo.

  Corresponding spatial tuple

  - *Trajector :* an elephant
  - *Landmark :* the zoo
  - *Spatial indicator :* in
  - *Spatial relation type :* region

- **Sentence :** There is a statue on top of the building.

  Corresponding spatial tuple

  - *Trajector :* a statue
  - *Landmark :* the building
  - *Spatial indicator :* on top of
  - *Spatial relation type :* direction

- **Sentence :** The tables are close to the poolside bar.

  Corresponding spatial tuple

  - *Trajector :* The tables
  - *Landmark :* the poolside bar
  - *Spatial indicator :* close
  - *Spatial relation type :* distance

Notice that a single sentence can be associated to multiple tuples encoding spatial relations, as shown in the next example.

- **Sentence :** A woman and a child are walking over the square.

  Corresponding spatial tuple 1

  - *Trajector :* A woman
  - *Landmark :* the square
  - *Spatial indicator :* walking over
  - *Spatial relation type :* direction

  Corresponding spatial tuple 2

  - *Trajector :* a child
  - *Landmark :* the square
  - *Spatial indicator :* walking over
  - *Spatial relation type :* direction

Dynamic spatial relations, on the other hand, involve motions, being annotated by relations that hold between a number of spatial roles, particularly involving an object (i.e., the trajector) which moves, starts, interrupts, resumes a motion, or is forcibly involved in a motion, and always involving also a phrase that constitutes a motion spatial indicator. A path spatial role can be used to annotate the path of the motion as the trajector is moving along, starting in, arriving in or traversing it, whereas spatial roles of distance and direction can be used to annotate textual spans encoding this type of information. Finally, a landmark spatial role can also be used to capture the spatial context of a motion. The following examples illustrate how dynamic spatial relations should be annotated, in the case of one particular sentence.

- **Sentence :** In Germany, while coming from Berlin, she stepped into the Grunewald forest and followed down a narrow trail.

  Corresponding spatial tuple 1

  - *Trajector :* she
  - *Landmark :* Germany
  - *Spatial indicator :* In
  - *Motion indicator :* coming
  - *Path :* from Berlin
  - *Spatial relation type :* direction

  Corresponding spatial tuple 2

  - *Trajector :* she
  - *Motion indicator :* stepped into
  - *Path :* the Grunewald forest
  - *Spatial relation type :* direction

  Corresponding spatial tuple 3

  - *Trajector :* she
  - *Motion indicator :* followed down
  - *Path :* a narrow trail
  - *Spatial relation type :* direction

Due to length constraints, we omit the complete formal specifications for the different spatial roles and relations. Detailed information in nonetheless available in the SemEval-2013 task description webpage[1]. A complete automated system for addressing

[1] http://www.cs.york.ac.uk/semeval-2013/task3/

the spatial role labeling task is required to (i) identify the markable spans of text for the different types of spatial annotations (i.e., trajector, landmark, spatial indicator, motion indicator, path, direction and distance), (ii) identify the tuples that connect the different spatial annotations into spatial relations, and (iii) perform the semantic classification of the spatial relation tuples into predefined relation types (i.e., region, direction or distance relationships).

3. THE PROPOSED METHOD

The spatial role labeling task definition, given in the previous section, leads to a similar problem as that of semantic role labeling, where words are classified based on a known predicate (i.e., a verb). In spatial role labeling, the spatial indicator is the pivot (i.e., the predicate, usually corresponding to a preposition or to a prepositional phrase) of the spatial relation. In this paper, we argue that existing state-of-the-art systems for semantic role labeling can easily be adapted to address the spatial role labeling task.

Concretely, we have that Semantic Role Labeling (SRL) is nowadays a popular and well-studied natural language processing (NLP) task, which consists of identifying and labeling semantic arguments (i.e., expressions defining time, location, manner, etc.) of verb predicates. The state-of-the-art for the task, in English and when evaluating results on PropBank data [20], corresponds to a F1 score of approximately 80%. Most SRL systems make use of a vast number of linguistically-informed features, including parts-of-speech tags, paths in the syntactic tree of the input sentence, named entities, and many others. However, recent proposals have instead addressed the task through multilayer neural network architectures, relying also on unsupervised word embeddings as features [4, 9].

In this paper, we have specifically adapted the `nlpnet`[2] system [9], which in turn uses similar ideas to those found in the well-known SENNA system for SRL [4]. Both these systems rely on a convolutional neural network based on a multilayer perceptron (MLP) architecture, which to some degree avoids the usage of explicit linguistic knowledge. Instead of engineered features, `nlpnet` mostly leverages vectorial representations for the words (i.e., word embeddings trained in an unsupervised fashion [21]), as input to the system. In results reported for English SLR, this architecture achieves near-state-of-the-art performance, while greatly reducing execution times and allowing for a standalone system.

Our adapted version of the `nlpnet` system tags each word in a sentence, in relation to a particular spatial indicator, with an a-priori score for each possible label (i.e., labels for all kinds of spatial roles and none), and it then uses the Viterbi algorithm to find the best tag path (i.e., after computing a-priori scores for all tokens, the final answer is found through a dynamic programming algorithm that uses these scores and a transition matrix between labels, also learned during model training). Each spatial relation is thus an instance of a spatial indicator and its roles. The system starts by identifying the spatial indicators, through the usage of a simple MLP network (i.e., without a convolution layer), which is fed with a 7 word window as input to tag the middle word. This network converts the tokens into feature vectors (i.e., tokens are initially represented through their word embeddings), and adjusts these features during training. After identifying the spatial indicators, `nlpnet` allows us to address the task of identifying the corresponding roles through two different alternatives: we can consider two different classification steps, respectively for identifying arguments and for associating them to a spatial indicator, or we can alternatively consider a single classification step for addressing both tasks simultaneously.

The one-step approach in `nlpnet` uses the same architecture as that from the SENNA system, except that a-priori scores are given in the form of IOB tags, instead of following the IOBES encoding used in SENNA (i.e., since the different roles are composed of syntactic phrases rather than isolated words, we require the usage of an encoding scheme such as IOB or IOBES, in order to delimit the boundaries of the spatial roles). The IOB tags, together with labels for the different spatial roles, allow for the identification of the spans of text that constitute the different roles that are associated to the spatial indicator. In the two-step approach, the first network is mostly equal to the one from the one-step approach, with the difference that its output consists of the a-priori score for each of the five IOBES tags, without combining them with labels for the different spatial roles. The network of the second step classifies the spatial roles, rather than the word tokens, although its input is still one token at a time. In order to handle varying size sentences, the networks employed by `nlpnet` perform a temporal convolution, as explained further below. In both approaches, the different neural network connections, the word representations, and the transition matrix, are all adjusted during training, via gradient descent.

Notice that in classical SLR, each single sentence may involve one or more predicates, although each having a single semantic interpretation, eventually with a different number of arguments. However, in the case of spatial role labeling, each spatial indicator (i.e., the equivalent of a predicate) can be associated to multiple spatial relations. In the current version of our adapted system, we use all possible spatial relations that are associated to a given spatial indicator as training instances. However, during inference, we only output a single relation per spatial indicator. In future work, this limitation can perhaps be addressed by considering an additional number of arguments for each spatial indicator.

In the classification models from `nlpnet` that are responsible for identifying roles, at each instant t, the t-th token in the sentence is fed to an MLP neural network, together with a window of neighbours. The input layer maps them to feature vectors and forwards them to a convolution layer. After applying a weighted sum, each convolution neuron stores the result of the t-th input. This process is repeated for all tokens in the sentence, and after that, each convolution neuron outputs the highest value found among all tokens. A sigmoid function is applied, and the results are further fed to a common MLP layer, which extracts higher level latent features.

In the one-step approach, in order to indicate which is the word being labeled (i.e., the target word) and with respect to which spatial indicator (i.e., spatial indicators act as the pivot for each spatial relation), word vectors are augmented with distance features. The relative distance from the t-th word to the target and to the predicate are calculated as the difference between their positions in the sentence. For each distance, up to a given threshold, there is a feature vector. These vectors are initialized randomly and adjusted during training. Note that, without distance features, the input for all words in a sentence would be exactly the same. In the case of the two-step approach, instead of the distance to the target word, we must find the distance to the target spatial role. Tokens inside the target are assigned distance zero, while tokens before it are assigned the distance to the first token of the spatial role, and tokens after the target are assigned the distance to the last token. Distance to spatial indicators can be computed in the same way as in the case of the one-step approach.

The authors of SENNA showed that, for classical SLR, initializing the word representations through a separate neural language model yields an improvement of more than 3 points in F1 score, compared with random initializations [4]. The only desideratum in these word representations is that words with similar meaning and usage should have similar vectors, considering their Euclidian

[2] http://nilc.icmc.usp.br/nlpnet/

distance or the cosine similarity. Proper initialization can help the network to detect which words should be treated similarly. Additionally, pre-trained word embeddings can reduce the impact of words that were not seen in the training data, as long as they have feature vectors. In our experiments, we used GloVe word embeddings as the initialization [21], respectively by relying on the pre-trained embeddings available from GloVe's website[3] (i.e., we used 300-dimentional word embeddings trained from 6B tokens of text, collected from Wikipedia and from the Gigaword corpus).

GloVe embeddings are based on global corpus statistics. We assume a collection of words w and their contexts c, and denote the collection of observed word-context pairs as D. The collection D is commonly obtained by taking a corpus and defining the contexts of word w as the words surrounding it in an given window of text. Usually, the vocabulary of contexts is equal to the vocabulary of words. A function $\text{count}(w, c)$ denotes the number of times the pair (w, c) appears in D. GloVe seeks to represent each word w and each context c as d-dimensional vectors $\mathbf{w}$ and $\mathbf{c}$ such that:

$$\mathbf{w} \cdot \mathbf{c} + b_w + b_c = \log(\text{count}(w, c)) \quad \forall (w, c) \in D \qquad (1)$$

In the previous equation, the b_w and b_c scalars are word/context-specific biases, and are also parameters to be learned, in addition to the embeddings $\mathbf{w}$ and $\mathbf{c}$. GloVe's objective is explicitly defined as a factorization of the log-count matrix, shifted by the entire vocabularies' bias terms, as shown below:

$$M^{\log(\text{count}(w,c))} \approx W \cdot C^{\top} + \mathbf{b}^w + \mathbf{b}^c \qquad (2)$$

In the formula, $\mathbf{b}^w$ is a row vector with dimensionality corresponding to the number of different terms, and $\mathbf{b}^c$ is column vector with dimensionality corresponding to the number of different contexts. The model is fit to minimize a weighted least square loss, giving more weight to frequent (w, c) pairs. After model fitting, one can take the representation of a word w to be the embedding resulting from $\mathbf{w}+\mathbf{c}$, where $\mathbf{c}$ is the row corresponding to w in $C^{\top}$. Additional details are given in the paper by Pennington et al. [21].

Besides the embeddings for each word type, and besides the vectors encoding distances, the representations used in `nlpnet` can also include vectors for other discrete attributes (e.g., parts-of-speech tags and chunking information). In our experiments, we used the pattern.en[4] Python module for English text processing, relying on it to tokenize the datasets containing the spatial role labeling annotations, for performing parts-of-speech (POS) tagging (i.e., annotating words according to morphological and syntactic categories from the Penn Treebank II tagset), for lemmatization and for phrase chunking (i.e., identifying noun phrases and verb phrases). Currently, the `nlpnet` system only allows one to use POS tags as additional linguistic features.

The training of the neural networks involved in `nlpnet` is done by back-propagation in order to maximize the likelihood over training sentences. Details on the training procedure are given in the paper by Fonseca and Rosa [9], and also on the original paper describing the SENNA system [4].

4. EXPERIMENTAL RESULTS

In our experiments, we used the datasets previously made available in the context of the SemEval-2013 shared task on Spatial Role Labeling [13]. The datasets for this shared task comprised two different corpora, namely a subset of the IAPR TC-12 image benchmark corpus[5] containing annotations for static spatial relations, and a corpus of documents gathered from the Confluence project[6], containing annotations for both static and dynamic spatial relations. The IAPR TC-12 corpus contains 613 textual documents that, in total, include 1213 sentences. The texts describe photos taken by tourists, presenting objects in a scene together with their absolute and relative positions in the image. A total of 600 sentences, describing 765 spatial relations, were selected for model training (i.e., about 50% of the entire corpus), and the remaining 613 sentences, with a total of 940 spatial relations, are used for evaluation. The Confluence corpus contains user-generated contents corresponding to descriptions of locations situated at specific latitude and longitude intersections in the world. The entire corpus contains 1789 sentences (i.e., about 40,000 tokens), with 1422 sentences (i.e., 2105 spatial relations, either static or dynamic) for training and 367 sentences (i.e., a total of 598 spatial relations) for evaluation.

The original XML encoding for the SemEval-2013 datasets was converted into the format used in the CoNLL-2005 shared task focusing on Semantic Role Labeling (SRL), which is directly supported by `nlpnet` and by most other open-source tools addressing the SLR task. We detected several encoding errors and inconsistencies in the SemEval-2013 datasets (e.g., attributes with empty values in the XML datasets, or inconsistent references to particular spans of text within the XML attributes), which we corrected for the converted datasets used in our experiments. Nonetheless, apart from these corrections, we used the same data splits for model training and testing, as those used by the participants at SemEval.

Notice that spatial role labeling systems are required to (i) identify the spans of text for the different types of spatial annotations (i.e., trajector, landmark, spatial indicator, motion indicator, path, direction and distance), (ii) identify the tuples that connect the different spatial annotations into spatial relations, and (iii) perform the semantic classification of the spatial relation tuples into predefined relation types (i.e., region, direction or distance relationships). In the context of SemEval, system outputs were evaluated against the gold annotations, and the task organizers considered different evaluation scenarios, leveraging both on strict or related criteria. Participants were allowed to consider partial versions of the complete task (e.g., participants could focus on identifying spatial relations, without considering their semantic classification).

For task (i), the system annotations are spatial roles, and each role was considered correct if it had a minimal overlap of one character with a gold annotation, and if it matched the role type of the gold annotation (i.e. SemEval followed a relaxed evaluation criteria for evaluating systems participating in task (i)). For tasks (ii), the system annotations are spatial relation tuples, and each tuple is considered correct if it is of the same length as the gold annotation (i.e., if it has the same number of spatial roles), and if each spatial role in the system tuple matches each role in the gold tuple. A spatial role estimated by a system is considered correct if it matches a gold reference when having the same character offsets and types (i.e., a strict evaluation criteria, although the organizers also considered a relaxed scenario in which a minimal overlap of one character was enough to consider estimated spatial roles to be correct). Finally, for task (iii), a system annotation of a spatial relation is considered correct if the spatial relation tuple is correct under the evaluation of task (ii), and if the relation type of the system relation is the same as the relation type of the gold relation. Systems were evaluated for each of the tasks in terms of precision, recall and F1 scores, according to the aforementioned notions of correct results.

[3] http://nlp.stanford.edu/projects/glove/

[4] http://www.clips.ua.ac.be/pages/pattern-en

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

[6] http://confluence.org

Table 1: Results for the complete identification and classification of spatial relations, for both corpora from SemEval-2013.

Golden indicators	IAPR TC-12			Confluence		
	P	R	F1	P	R	F1
1-step approach	72.7	72.0	72.3	48.2	45.6	46.9
2-step approach	72.4	68.6	70.5	44.3	45.2	44.8

Inferred indicators	IAPR TC-12			Confluence		
	P	R	F1	P	R	F1
1-step approach	73.2	67.5	70.2	47.8	44.9	46.3
2-step approach	72.8	64.1	68.2	44.1	44.7	44.4

In this paper, we do not report on results according to all the criteria considered for SemEval-2015. We focus on evaluating the complete spatial role labeling task, and therefore we mostly rely on a strict criteria, in which we measure precision and recall in terms of exactly identifying spatial relations as defined in the golden annotations, with all their corresponding spatial roles.

Table 1 presents the obtained results in terms of the complete and exact identification of spatial relations, both over the IAPR TC-12 corpus (i.e., involving only static spatial relations) and over the Confluence corpus (i.e., involving both static and dynamic spatial relations). Table 1 also presents results when considering a scenario where the spatial indications are correctly provided as input to the system (i.e., the results on the top two rows), versus the realistic scenario in which spatial indicators also need to be inferred from the text. The obtained results attest to the adequacy of the proposed procedure, as we obtained scores that are in line with those reported at previous studies in the area [17, 14, 13].

Additional experiments are currently underway, involving (i) ablation tests in which only word embeddings are considered as features (i.e., without considering POS tags), (ii) tests with different types of word embeddings, and with embeddings of different dimensionality, and (iii) tests with other datasets annotated for spatial roles (e.g., large datasets of annotated image descriptions [24]).

5. RELATED WORK

Extracting semantic relations from text is at the core of any application involving text understanding. In the past years, computational semantics has received a significant boost within the natural language processing (NLP) community, although extracting all semantic relations in text, even if considering only single sentences, is still an elusive goal. Most existing approaches target either a single relation, e.g., PART-WHOLE [11], or relations that hold between arguments following some syntactic construction. Among the latter kind, the task of verbal semantic role labeling focuses on extracting semantic links exclusively between verbs and their arguments. PropBank is a popular corpus for this task [20], and tools to extract verbal semantic roles have been proposed for years [10, 3]. In this paper, we build on this large body of previous work, proposing to adapt an existing state-of-the-art tool for verbal semantic role labeling, to the less studied task of spatial role labeling.

While both cognitive and formal linguistics have examined the meaning of motion verbs and spatial prepositions (e.g. the the cognitive framework introduced by Langacker states that elementary spatial concepts can be characterised by locative relations between a potentially mobile object, called the trajector, and a static reference object called the landmark [18]), earlier approaches in the area of computational spatial semantics have not yield precise computable representations that are expressive enough for natural languages [25]. The spatial role labeling task, as considered in this paper, was introduced in 2010 by Kordjamshidi et al. [15, 17], and latter explored in the context of joint evaluation challenges on SemEval [14, 13]. Kordjamshidi et al. proposed an annotation scheme, and also an automated procedure for annotating static spatial relations, initially without considering their semantic classification, based on three different modules. The first module corresponds to classifier (e.g., a maximum entropy model) that takes a word (i.e., a preposition) in the sentence as an input, represents this word through a set of linguistically-motivated features (i.e., the preposition itself, the words that are directly associated to the preposition, according to the results of a dependency parser, their lemmas and POS tags, etc.) and estimates whether the word is a spatial indicator. The second module is multi-class classifier, based on the formalism of conditional random fields, that takes a spatial indicator, together with the complete sequence of words that form the input sentence, and tags the words according to the different spatial roles (i.e., trajector and landmark). The features used in this second model are mostly inspired on those from previous studies in semantic role labeling (i.e., features capturing different aspects of each word to be classified, of the spatial indicator of which the word may be an argument, and of the relation between these two), although the authors also found that additional features based on distances, word sub-categorizations and semantic roles, could positively impact the results. The final module, which is relatively straightforward and involves no learning, assembles the results of the previous two steps to form spatial relation tuples. The authors also investigated an alternative approach, in which the different steps were addressed jointly (i.e., using a single learned model to tag all words in a sentence jointly), which nonetheless achieved slightly worse results. In experiments with two datasets similar to those from the SemEval competitons (i.e., over textual descriptions for 400 images of the IAPR TC-12 image dataset, and over a set of sentences from the Confluence project), the authors respectively report on F1 scores of 0.714 and 0.475, for the complete task of labeling static spatial roles, although without considering their semantic classification (i.e., into region, direction or distance relationships).

The SemEval-2012 task on spatial role labeling had a single participant [26], which compared different feature sets within a system that employed a high recall heuristic (i.e., take all noun phrase heads) for recognizing objects capable of participating in a spatial relation, together with a lexicon of spatial indicators. All possible combinations of these arguments are then considered by a binary support vector machine classifier, that leverages a rich set of features in order to make a joint decision. Within the context of the SemEval challenge, it is also important to notice that spatial indications were not limited to single-token spatial prepositions (e.g., *in* or *over*), as they could also correspond to other types of spatial phrases (e.g., *in front of* or *on the left*). The single participant achieved a best F1 score of 0.573 for the full identification of static spatial relations, and an F1 score of 0.566 when additionally considering the classification of the triplet's semantic type.

In SemEval-2013, the task was extended in order to consider also dynamic spatial roles, although none of the participants (i.e., again only one) addressed the complete version of the task. The best results for the complete recognition of static spatial relations corresponded to an F1-score of 0.358, and they were obtained with a system that identified the spatial roles though sequence-based word classification, specifically by leveraging SVM^{hmm} models, and that used a separate SVM classifier for verifying the candidate spatial relations that may hold between the recognized roles, in this case leveraging a convolution kernel that measures the similarity between syntactic structures [1].

The low number of participants in the SemEval challenges that focused on spatial role labeling lead us to conclude that, although this is an interesting task with many practical applications, it is also highly complex (i.e., the development from scratch of a system for spatial role labeling is a complex endeavour, often involving also the usage of different NLP annotation layers). In this paper, we argue that spatial role labeling can build on recent developments on the general area of semantic parsing. We show that the adaptation of a state-of-the-art system for semantic role labeling, mostly using word embeddings as features and leveraging an approach based on convolutional neural networks [9], is relatively straightforward.

6. CONCLUSIONS

The ability to understand spatial language can enable a variety of new applications, supporting systems that can respond to verbal directions, map travel narratives, render spatial scenes from text, etc. In this paper, building largely upon the resources of previous joint evaluation efforts at SemEval, we proposed to address the task of spatial role labeling through an adapted version of a state-of-the-art semantic role labeling system. Our experimental results are in line with those from previous studies in the area, showing that progress within the spatial role labeling task can effectively leverage on advances for the general NLP area of semantic role labeling.

Despite the interesting results, there are also many possible avenues for future work. Research within the NLP community continues to explore different word embedding models and, in future experiments, it would be interesting to test recent alternatives, as a replacement to the GloVe embeddings used in our experiments [8, 6, 7]. Recent NLP studies have also proposed alternative procedures for addressing the task of semantic role labeling, for instance leveraging tensor-based approaches to induce a compact feature representation for words and their relations [19], or leveraging on different types of neural network architectures [28]. For future work, it would be interesting to compare the approach based on the `nlpnet` system [9], that was employed here, against other state-of-the-art approaches proposed for semantic role labeling.

Recently, Blanco and Vempala repored on a study that a considered a semantic annotation task that is different from spatial role labeling as defined in SemEval-2013, specifically by complementing PropBank-style semantic role representations with additional spatial knowledge, and considering also temporal spans of validity and the certainty of the associations [2]. These latter two aspects are particularly useful because (i) spatial information for most objects changes over time, and (ii) humans sometimes can only state that an object is probably located somewhere. Future work in the area will likely also involve more expressive semantic annotations, together with the spatio-temporal anchoring of the different spatial roles [12]. The SpaceEval task, introduced in the context of SemEval-2015, already extended the spatial role labeling task in several dimensions [22], by adopting the more advanced annotation specification from ISOspace [23].

Acknowledgments

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), through project grants with references EXCL/EEI-ESS/0257/2012 (DataStorm research line of excellency), EXPL/EEI-ESS/0427/2013 (KD-LBSN), and UID/CEC/50021/2013 (INESC-ID's associate laboratory multi-annual funding).

7. REFERENCES

[1] E. Bastianelli, D. Croce, R. Basili, and D. Nardi. UNITOR-HMM-TK: Structured kernel-based learning for spatial role labeling. In *Proceedings of the International Workshop on Semantic Evaluation*, 2013.

[2] E. Blanco and A. Vempala. Inferring temporally-anchored spatial knowledge from semantic roles. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2015.

[3] X. Carreras and L. Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Conference on Computational Natural Language Learning*, 2005.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2011.

[5] D. Das, D. Chen, A. F. T. Martins, N. Schneider, and N. A. Smith. Frame-semantic parsing. *Computational Linguistics*, 40(1), 2014.

[6] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2014.

[7] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. Smith. Sparse overcomplete word vector representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2015.

[8] Y. L. J. X. Fei Sun, Jiafeng Guo and X. Cheng. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2015.

[9] E. R. Fonseca and J. L. G. Rosa. A two-step convolutional neural network approach for semantic role labeling. In *Proceedings of the International Joint Conference on Neural Networks*, 2013.

[10] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3), 2002.

[11] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1), 2006.

[12] M. V. James Pustejovsky, Jessica Moszkowicz. A linguistically grounded annotation language for spatial information. *ATALA: Association pour la Traitment Automatique des Langues*, 53(2), 2013.

[13] O. Kolomiyets, P. Kordjamshidi, M.-F. Moens, and S. Bethard. Semeval-2013 task 3: Spatial role labeling. In *Proceedings of the International Workshop on Semantic Evaluation*, 2013.

[14] P. Kordjamshidi, S. Bethard, and M.-F. Moens. Semeval-2012 task 3: Spatial role labeling. In *Proceedings of the International Workshop on Semantic Evaluation*, 2012.

[15] P. Kordjamshidi, M. V. Otterlo, and M.-F. Moens. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[16] P. Kordjamshidi, D. Roth, and M. Moens. Structured learning for spatial information extraction from biomedical text: bacteria biotopes. *BMC Bioinformatics*, 16:129, 2015.

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

[18] R. Langacker. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press, 1987.

[19] T. Lei, Y. Zhang, L. Márquez, A. Moschitti, and R. Barzilay. High-order low-rank tensors for semantic role labeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2015.

[20] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 2005.

[21] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[22] J. Pustejovsky, P. Kordjamshidi, M.-F. Moens, A. Levine, S. Dworman, and Z. Yocum. Semeval-2015 task 8: Spaceeval. In *Proceedings of the International Workshop on Semantic Evaluation*, 2015.

[23] J. Pustejovsky, J. L. Moszkowicz, and M. Verhagen. Iso-space: The annotation of spatial information in language. In *Proceedings of the ACL-ISO International Workshop on Semantic Annotation*, 2011.

[24] A. Ramisa, J. Wang, Y. Lu, E. Dellandrea, F. Moreno-Noguer, and R. Gaizauskas. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015.

[25] T. P. Regier. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. PhD thesis, University of California at Berkeley, 1992.

[26] K. Roberts and S. Harabagiu. UTD-SpRL: A joint approach to spatial role labeling. In *Proceedings of the International Workshop on Semantic Evaluation*, 2012.

[27] N. Tandon, G. Weikum, G. d. Melo, and A. De. Lights, camera, action: Knowledge extraction from movie scripts. In *Proceedings of the International Conference on World Wide Web*, 2015.

[28] J. Zhou and W. Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2015.

[29] J. Zlatev. *The Oxford Handbook of Cognitive Linguistics*, chapter Spatial semantics. Oxford University Press, 2007.