26th Annual KES Conference, KES 2022, Verona, Italy, 7-9 September 2022

Mixing Word Embeddings and RoBERTa for Spatial Role Labeling

Alaeddine Moussa
alaeddine.moussa@etu.univ-amu.fr
Aix-Marseille University
Univ. Manouba, ENSI, Campus Universitaire Manouba
Marseille, France

Sebastien Fournier
Aix Marseille University
Marseille, France

Khaoula Mahmoudi
LTRSIRS
Tunis, Tunisia

Bernard Espinasse
Aix-marseille University
Marseille, France

Sami Faiz
LTSIRS
Tunis, Tunisia

ABSTRACT

Language model pretraining has yielded significant results in diverse natural language processing tasks. RoberTa, an efficient method for pretraining self-supervised NLP systems, is a good example. Our hypothesis in this paper is that the performance of Spatial Role Labeling (SpRL) can be improved by combining static word vectors and bags of features with RoberTa vectors. Furthermore, we show that our method is successful in several SpRL datasets.

CCS CONCEPTS

• **Computing methodologies**;

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Alaeddine Moussa, Sebastien Fournier, Khaoula Mahmoudi, Bernard Espinasse, and Sami Faiz. 2021. Mixing Word Embeddings and RoBERTa for Spatial Role Labeling. In *Proceedings of The 20th IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'21).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Spatial Role Labeling (SpRL) [13] introduces a language-independent annotation scheme to identify spatial scene objects and relationships within a text, including reasoning purposes. Spatial Role Labeling is used in many application areas such as robotics [18], maritime navigation [9], traffic management [1], in query response systems [3]. It involves processing sentences in a text and identifying the objects in spatial scenes and their relationships.

Linguistic constructs may communicate complex relationship structures and spatial relationships between them and movement patterns across space relative to a reference point. A spatial relationship between two objects is typically expressed in natural language through a preposition (e.g. in, on, at, ...) or prepositional expression (on top of, inside of, ...).

For instance, in the a sentence "Trees in the background" where the token "Trees" is a TRAJECTOR (TR), "background" a LANDMARK (LM), and "in" the SPATIAL INDICATOR (SI). The spatial indicator (often a preposition) establishes the type of spatial relationship. Thus, the task of SpRL is to process the sentences of a text automatically and identify the spatial scene objects and the relations between them.

A trajector (TR) describes a central object of a spatial scene. It can be static or dynamic; including persons, objects, or events. The landmark (LM) is a spatial role label that denotes the location or the motion of the TR. It indicates a secondary object of a spatial scene, to which a possible spatial relation can figure out. A spatial Indicator (SI) is a spatial role label allocated to a word or a phrase to flags a spatial relation between objects (TR and LM).

Most of the existing SpRL approaches imply Word Embedding (WE). WE is one of the most useful deep learning methods used for constructing vector representations of words. Although very effective, these methods have certain limits and need to be improved. Recently, contextualized word representations models like BERT [7] and RoBERTa [14] acheived state-of-art in many nlp tasks.

In this research, we propose to improve the accuracy of SpRL systems. We propose to combine static word vectors and bags of features with vectors from RoberTa. The results show that proposed method increases the accuracy of embeddings vectors for SpRL. The main contributions of our work are the following:

1) We present a RoberTa based approach for SpRL.
2) We combine RoberTa vectors with the WE vector ,and bag features, including Part of Speech and character-level representation.

The remainder of this paper is organized as follows : Section 2 describes the related works and literature review. Section 3 presents our proposed method and describes the proposed deep learning model. Section 4 reports the designed experimental setup. Results are discussed in Section 5. Finally, Section 6 summarizes the major findings.

WI-IAT'21, 14-17 December 2021, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

SpRL has been the subject of several evaluation compaigns in the context of SemEval. SemEval is a series of ongoing evaluations of computational semantic analysis systems to explore the nature of meaning in language. Although meaning is intuitive to humans, the transfer of these intuitions to computer analysis has proven complicated.

The SemEval campaigns in automatic information extraction of the years 2012, 2013, and 2015 involved challenges to be taken up in SpRL on a specific proposed corpus of reference texts. Thus, the SpRL task at SemEval 2012 [11] focused on the roles of TRs, LMs, SIs, and the links between these roles form spatial relations. The formal semantics of the relations were considered at a granular level, composed of three types of relations: directional, regional (topological), and distal (the part furthest from a reference center). The spatial roles are assigned both to phrases and their headwords, but only the headwords are evaluated for this task.

In the SemEval 2013 SpRL task, [10] introduced new roles to include motion indicators, paths, directions, and distances to capture the fine-grained spatial semantics of static spatial relationships and also to take into account dynamic spatial relationships. In this task, the entire span of spatial roles is evaluated, not only headwords.

The SpaceEval task, introduced in SemEval-2015, adopted a more advanced annotation specification with respect to ISOspace [21]. SpaceEval, first, enriches the semantics' granularity in both static and dynamic spatial configurations, and secondly, by extending the variety of annotated data and the domains considered. Indeed, the concept of place is distinguished from the concept of a spatial entity.

The KUL-SKIP-CHAIN-CRF [13] system adopted a pipeline approach. In the beginning, spatial indicators are found and labeled for each sentence. Then, given a spatial indicator, the second task consists of classifying parts of an input sentence with LM or TR labels. To identify the SI, authors use an external corpus The Preposition Project (TPP) labeled to learn the sens of preposition. As a result, the binary classification of a preposition's spatial is made. In the same way, TR and LM are identified to leverage many linguistically motivated features. Finally, SIs, TRs, and LMs are combined into spatial relation triplets.

The UTDSpRL [22] system used a join approach to tackle the SpRL task. Authors expose that the pipeline approach cannot perform well if a spatial roles arguments are considered in isolation. They propose an alternative system that jointly decides whether a given candidate triple expresses a spatial relation or not. The approach used a recall heuristics to find spatial relation candidate triple. A hand-crafted dictionary was used to detect SPATIAL_ INDICATOR candidates, allowing the decrease of negative relation candidates in comparison to the pipeline method, which considers every phrase as a spatial role. Then, noun phrase heads were treated as TRAJECTOR and LANDMARK candidates. Finally, a Support Vector Machine (SVM), was trained with some manually engineered features chosen by an automatic feature selector, to classify the relation candidate.

The UNITOR-HMM-TK [2] system uses a sequence-based classifier. First, classifies spatial and motion indicators, then built on these outcomes to identify spatial roles through a set of lexical and grammatical features. The generated candidate spatial relations are verified by a Support Vector Machine (SVM). This approach's significant contribution is to adopt smooth grammatical features instead of a full syntax of the sentence.

SpRL-CNN system [19] rely on an adapted version of the nlpnet[1] system. First, a spatial indicator is identified by a multilayer perceptron (MLP) witch it converts the tokens (spatial indicator) into feature vectors. Then, these vectors are fed to Conventional Neural Network (CNN) to classify the argument and spatial relation triples. Additionally, pre-trained Glove word embeddings were used to reduce the impact of words not seen in the training data.

The VIEW system [15] consists of producing word embedding vectors from multimodal. The goal is to be able to encode the visual information of images within the "Feature vector" produced by [12] by concatenating them with the "embedding" of VIEW.

Recently, [6] extended SpRL to encode data in radiology context. They presented a dataset with 2000 chest X-ray reports. The annotation involves a radiographic conclusion and its associated anatomical position.

Most of the existing systems depend heavily on manual features and do not benefit from the advances of NLP systems based on deep learning techniques.

3 PROPOSED METHOD

This section details the approach we advance to improve the SpRL task by mixing static word vectors and bags of features with RoberTa vectors. The prominent architecture of the proposed method is shown in figure 1.

3.1 Generate RoBERTa contextualized word embeddings

BERT is meant to pre-train deep bidirectional representations. Then, the pre-trained BERT model can be fine-tuned to create state-of-the-art models for a wide range of tasks. However, RoBERTa authors find that BERT was under-trained and present a replication study of BERT that outperformed the BERT model. In this work, we attend to use RoBERTa vector in SpRL tasks. To fine-tune RoBERTa for SpRL we need first the input data to be tokenized. Words will be splitted into smaller word-pieces and characters. Some subwords have two hash characters '##' preceding them, indicating that they are part of a bigger word and are preceded by another subword. For instance, word “pylons” is divided into three tokens as 'p', '##yl', and '##ons'. To generate the embedding vector for the word "pylons," we average the three tokens vectors to get a unique vector of the word "pylons." In order to retrieve the vectors, the authors of original BERT propose several choices [7]. We opted for the concatenation of the last four layers in this work because this method gave us the best results.

3.2 Part-of-speech to vector

Part-of-speech (POS) tagging is an essential and fundamental step in Natural Language Processing that specifies the appropriate POS tag for each word in a text. The POS contains a wealth of information about a word and its neighbors and grammatical groups of

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

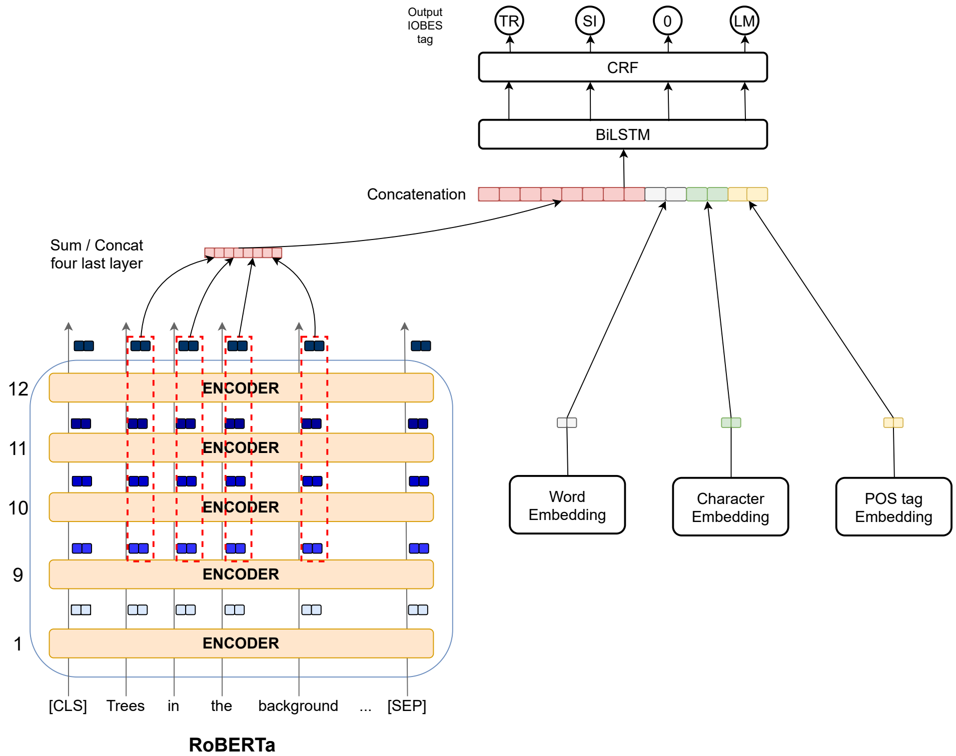

Figure 1: The global architecture of our proposed method.

words (nouns, verbs, adjectives, and adverbs) and similarities and differences. Each produced POS tag is transformed into a vector representation.

3.3 Character to vector

The character-level offers much information about the word's meaning. Using this knowledge, the model's quality can be considerably improved [4]. We utilized a Convolution Neural Network (CNN) to encode the structure of words. We used one-hot encoding to encode each character into a vector, which we will be feed into a CNN.

3.4 BiLSTM-CRF Architecture

After concatenating RoBERTa vector along with static word embedding, POS tag vector, and character level vector into a single vector. We feed the resulting vector to a BiLSTM-CRF model [16]. When processing sequence labeling tasks, each labeling decision for each word is produced independently of the others, but in many cases neighboring decisions leverage each other. The Bi-directional LSTM (BiLSTM) encodes each sequence forwards and backward to capture past and future information. Given an input sequence $x = (x_1, ..., x_n)$ and a sequence of SpRL labels $y = (y_1, ..., y_n)$ for x. The probabilistic model for overall possible tag sequences will be defined as:

$$p(s|x; w) = \frac{e^{(w.\Phi(x,y))}}{\sum_{x'} e^{(w.\Phi(x',y))'}} \tag{1}$$

where s extends overall possible output sequence and w is the parameter vector. Indeed, the expression $w.\Phi(x, y) = score_{crf}(x, y)$ represents the scoring function that indicates how y fits x. In the CRF layer, we aim to replace the linear scoring function with a non-linear neural network. We define the score by:

$$score_{BiLSTM-crf}(x, y) = \sum_{i=0}^{n} W_{y_{i-1},y_i}.BiLSTM(x)_i + b_{y_{i-1},y_i} \tag{2}$$

With the score function constructed, we can efficiently use past and future tags to predict the current tag and optimize the conditional probability $p(y|x; w, b)$ and propagating back through the network. We use an interaction function $I(\vec{wc_i}^{\,s}, \vec{p_i})$ to capture the interaction between the child and its parent in a dependency tree. The function I is based on *average*. This meta-embedding technique gives results same as concatenation, with the additional benefit of reduced dimensionality [5].

The SpRL task is divided mainly into two subtasks: spatial role extraction and spatial relation extraction. For the relation extraction task, we aim to figure out which spatial roles discovered in the previous classification phase can be composed as valid spatial relations. All possible spatial roles are first generated by heuristics and then combinatorially combined to acquire candidate relationships. The network was feed with three input representation vectors v, each referring to the trajector, spatial indicator, and landmark.

4 EXPERIMENTS

We conduct experiments to evaluate the performance of the proposed approach for SpRL. In this section, we describe datasets used for testing followed by the experimental setup.

4.1 Datasets and preprocessing

We use the SemEval-2012 and SemEval-2013 shared tasks. We have chosen these two corpora because they are the reference corpora for SpRL and most systems have been tested in.

4.1.1 SemEval-2012 Data. The first corpus is a subset of the IAPR TC-12 image Benchmark [8]. It contains 613 text files that include 1213 sentences in total. The original corpus was available without copyright restrictions [2]. Tourists took 20,000 images with textual descriptions in different languages. The texts describe objects in a scene together with their absolute and relative positions in the image. The data is released in XML format.

4.1.2 SemEval-2013 Data. The dataset for SemEval2013 includes two distinctive corpora. The first one is the SemEval2012 Data presented in the previous sub-section but with some modifications. The Data was transformed into a span-based annotation, and some annotation error was corrected. The second corpus originates from the Confluence Project [3]. This project tries to outline all possible latitude-longitude intersections on earth, and people who visit these intersections provide written narratives of the visit.

4.1.3 Preprocessing. The original XML encoding for the previous datasets was converted into CoNLL-U format with Stanford CoreNLP toolkit [17] to get POS tag. We added a new column that holds the tag for each word in the IOBES labeling scheme.

4.2 Experimental Setup

We use *PyTorch* toolkit to test our approach for SpRL. We choose the RoBERTa base model provided by the Transformers package from the *HuggingFace* library. We limit our sentence length to 63 tokens, equal to the longest sentence of the corpus, instead of 512 initially supported by RoBERTa. We have kept the default suggested size of the batch, which is 32. In order to prepare the data for RoBERTa we need to tokenize all sentences. To prepare the data for RoBERTa, we need to tokenize all sentences. For that, we have used the *RobertaTokenizer* class from *roberta-base* model. Then, to fine-tuning RoBERTa for SpRL tasks, we used the RobertaForTokenClassification class for token classification. Hence, we encased RoBERTaModel and adds a token-level classifier on top of the RoBERTaModel. We utilized the AdamW optimizer. The hyper-parameters search for the fine-tuning process involves the number of epochs (3 to 6), the learning rates 3e-5, and eps = 1e-8. We also include a scheduler that reduces the learning rate linearly over the epochs. After the fine-tuning process is complete, for each token we have 13 separate vectors (1 input vector + 12 from RoBERTa hidden layers) of length 768. We then concatenate the last four layers, providing a single word vector per token with length 3,072.

To generate word embeddings, we use torch.nn.Embedding module from PyTorch along with Global Vectors (GloVe) proposed by [20]. We have also tested word2vec embedding, but Glove gives us a better result. We choose 1D CNN to construct character-level embeddings followed by a 1d maxpool operation. Hyper-parameters of the CNN was set after tuning to an embedding size of 30, a window size of 3, and 30 filters. The size of the POS tag came empirically with 30 dimensions each.

[2]http://www.cs.tulane.edu/ pkordjam/SpRL.htm# data
[3]http://confluence.org

5 RESULTS AND DISCUSSION

In this section, the experimental results are shown and discussed. First of all we present an ablation analysis of results, then an error analysis of experimental results. We present our main results for SpRL individual spatial role identification and relation extraction in Table 1 and Table 2.

5.1 Ablation study

An ablation analysis was used to investigate the impacts of mixing static word embedding, POS tag, character embedding ,and RoBERTa vector. Table 1 and Table 2 shows that the proposed model improves the F1-score across almost all datasets. Based on those results, we can deduce the influence of supplementing a static word embedding with RoBERTa embedding. We also observe that adding POS tags and character representation improves the performance compared to only using the word embedding.

5.2 Error analysis

We manually examined the errors to determine their nature. For example, the model wrongly assigns a spacial role to a word that is not one. For instance, consider the following sentence:

A mountain range with three outstanding rock summits at sunrise.

The system classifies “at” as an SI and “sunrise” as LM while “in” has no spatial sense.

In addition, the model assigns none tag to a token that plays spatial roles. In this example,

a lake with a couple of mountains behind it .

"lake" is wrongly classified as none and "it" labeled with none tag, whereas it should be tagged as LM.

6 CONCLUSION

In this paper, we proposed a new method to improve the accuracy of SpRL tasks. First of all, fine-tune RoBERTa for SpRL and generate relative contextualized word embeddings. Then, we mix the RoBERTa vectors with static word vectors enhanced with both POS tags and CNN-Based character representation. In the next step, we feed the resulting vector to a BiLSTM network. Finally, we add a CRF layer to exploit the neighbor tags while predicting current tags at the last stage.

Experimentation results are conducted on SemEval-2012 and SemEval-2013 Task 3 datasets and show that our proposed method increases the accuracy of Spatial information classification tasks in all datasets.

For future work, we ought to improve the performance of our work by introducing syntactic dependencies and mixing them with RoBERTa contextualized vectors.

Table 1: F1 scores comparison on the SemEval 2012 dataset with TR=Trajector, LM=Landmark, SI=Spatial Indicator, and relation=Spatial relation

System	TR	LM	SI	Relation
Glove	0.633	0.685	0.887	0.601
BERT	0.708	0.773	0.916	0.699
RoBERTa	0.720	0.799	0.923	0.719
Glove+RoBERTa	0.723	0.801	0.927	0.722
Glove+RoBERTa+POS+Char	**0.734**	**0.803**	**0.931**	**0.729**
SpRL systems				
EtoE-IBT-CLCP [12]	0.673	0.797	0.869	0.617
KUL-SKIP-CHAIN-CRF [13]	0.646	0.756	0.900	0.500
UTDSpRL [22]	0.707	0.772	0.823	0.573

Table 2: F1 scores comparison on the SemEval 2013 dataset with TR=Trajector, LM=Landmark, SI=Spatial Indicator, and relation=Spatial relation

System	TR	LM	SI	Relation
Glove	0.613	0.652	0.831	0.427
BERT	0.669	0.780	0.901	0.600
RoBERTa	0.701	0.789	0.901	0.610
Glove+RoBERTa	0.726	0.790	0.915	0.611
Glove+RoBERTa+POS+Char	**0.733**	**0.792**	0.919	**0.613**
SpRL systems				
SpRL-CNN [19]	NA	NA	NA	0.460
UNITOR-HMM-TK [2]	0.682	0.785	**0.926**	0.458
VIEW [15]	0.732	0.678	0.749	0.235

REFERENCES

[1] Fatma Başak Aydemir and Fabiano Dalpiaz. 2017. Towards aligning multi-concern models via nlp. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*. IEEE, 46–50.

[2] Emanuele Bastianelli, Danilo Croce, Daniele Nardi, and Roberto Basili. 2013. UNITOR-HMM-TK : Structured Kernel-based Learning for Spatial Role Labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Vol. 2. 573–579. https://www.cs.york.ac.uk/semeval-2013/accepted/52{_}Paper.pdf

[3] Konstantina Bereta, Guohui Xiao, and Manolis Koubarakis. 2019. Ontop-spatial: Ontop of geospatial databases. *Journal of Web Semantics* 58 (2019), 100514.

[4] Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint Learning of Character and Word Embeddings. Ijcai (2015), 1236–1242.

[5] Joshua N Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding-computing meta-embeddings by averaging source word embeddings. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 2. 194–198.

[6] Surabhi Datta and Kirk Roberts. 2020. A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations. *Data in Brief* 32 (oct 2020), 103473. https://doi.org/10.1016/j.dib.2020.106056

[7] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. , 4171–4186 pages. https://doi.org/10.18653/v1/N19-1423

[8] M Grübinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *LREC Workshop OntoImage Language Resources for Content-Based Image Retrieval*. 13–23. http://www-i6.informatik.rwth-aachen.de/publications/download/34/TheIAPRBenchmark:ANewEvaluationResourceforVisualInformationSystems.pdf

[9] Melita Hadzagic, Maxime Isabelle, and Nathan Kashyap. 2020. Hard and Soft Data Fusion for Maritime Traffic Monitoring Using the Integrated Ornstein-Uhlenbeck Process. In *2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 98–105.

[10] Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie-Francine Moens, and Steven Bethard. 2013. SemEval-2013 Task 3: Spatial Role Labeling. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* 2, SemEval (2013), 255–262. http://www.aclweb.org/anthology/S13-2044

[11] Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 Task 3: Spatial Role Labeling. *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2012)* (2012), 365–373. https://dl.acm.org/citation.cfm?id=2387694

[12] Parisa Kordjamshidi and Marie Francine Moens. 2015. Global machine learning for spatial ontology population. *Journal of Web Semantics* 30 (2015), 3–21. https://doi.org/10.1016/j.websem.2014.06.001

[13] Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling. *ACM Transactions on Speech and Language Processing* 8, 3 (2011), 1–36. https://doi.org/10.1145/2050104.2050105

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[15] Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. 2016. Deep Embedding for Spatial Role Labeling. 2015 (2016), 1–19. https://doi.org/10.13140/RG.2.1.5080.4884 arXiv:1603.08474

[16] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 2. 1064–1074. https://doi.org/10.18653/v1/p16-1101 arXiv:1603.01354

[17] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Stroudsburg, PA, USA, 55–60. https://doi.org/10.3115/v1/P14-5010

[18] Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet (invited talk). In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[19] Alexey Mazalov. 2016. Understanding Spatial Semantics in Natural Language. October (2016). https://fenix.tecnico.ulisboa.pt/downloadFile/1689244997257025/Extended{_}Abstract{_}Mazalov{_}80462.pdf

[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 1532–1543. https://doi.org/10.

3115/v1/d14-1162
[21] James Pustejovsky, Parisa Kordjamshidi, Marie-francine Moens, Aaron Levine, Seth Dworman, Zachary Yocum, and Katholieke Universiteit Leuven. 2015. SemEval-2015 Task 8: SpaceEval. 1 (2015), 884–894.
[22] Kirk Roberts and Sanda M Harabagiu. 2012. UTD-SpRL: A Joint Approach to Spatial Role Labeling. (2012), 419–424. http://www.aclweb.org/anthology/S12-1056