# LIP6@CLEF2017: Multi-Modal Spatial Role Labeling using Word Embeddings Working notes

Éloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari

## ▶ To cite this version:

# LIP6@CLEF2017: Multi-Modal Spatial Role Labeling using Word Embeddings
## Working notes

Éloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and
Patrick Gallinari

Sorbonne Universités,
UPMC Univ Paris 06, UMR 7606, CNRS, LIP6, F-75005, Paris, France
{eloi.zablocki, patrick.bordes, laure.soulier, benjamin.piwowarski,
patrick.gallinari}@lip6.fr,

**Abstract.** We report our participation to the multi-modal Spatial Role
Labeling (mSpRL) lab at CLEF 2017. The task consists in extracting
and classifying spatial relationships from textual data and associated
images. Our approach focuses on the classification part as we use a base-
line system for the extraction of the relations: we train a linear Support
Vector Machine (SVM) model to classify hand-crafted vectors represent-
ing spatial relations. We present the obtained experiments and discuss
also the effect of model parameters. Finally, we conclude the paper and
introduce ideas for future developments.

**Keywords:** multi-modal spatial role labeling, linear SVM, multi-label
classification, spatial indicator, landmark, trajector, word embedding,
RCC8 regions

## 1 Introduction

In this paper, we report our participation to the multi-modal Spatial Role La-
beling lab [8] at CLEF 2017 [4]. The task consists in extracting and classifying
spatial relationships from textual data and associated images.

The mSpRL goal is composed of three successive sub-tasks. The first one
aims at extracting spatially-related entities and annotating the text using the
following labels: *Trajector*, *Spatial Indicator* and *Landmark*. The second one
consists in associating the previously found entities into spatial relation triplets
$r = $ (trajector, spatial indicator, landmark). The goal of the third sub-task is to
classify those relations. While the two first sub-tasks could be seen as a linguistic
conceptual representation (spatial role labeling), the third sub-task rather refers
to a formal semantic representation of relations (spatial qualitative labeling).
Possible labels for the relation classification are divided in three general types:

- *Region RCC8 [10]* (8 possible values): disconnected (DC), externally con-
  nected (EC), equal (EQ), partially overlapping (PO), tangential proper part
  (TPP), tangential proper part inverse (TPPi), non-tangential proper part
  (NTPP), non-tangential proper part inverse (NTPPi).

- *Direction* (6 possible values): left, right, above, below, behind, front
- *Distance* (5 possible values): middle, fast, close, far, near

The mSpRL task of CLEF 2017 is built upon SemEval 2012 task 3 [5] which was proposed several years ago. That task has been augmented with images paired with text as additional inputs and sub-task 3 (relation classification).

The training set consists in 275 images, 600 associated sentences (several sentences can be linked to a single image) and a total of 761 relations. Figure 1 shows an example of the task.

**Text:**
"A woman is sitting at table in the middle of a school yard."

**Image:**

**Sub-task 1:**
Trajector: "a woman", "table"
Spatial indicator: "at", "in the middle of"
Landmark: "table", "a schoolyard"

**Sub-task 2:**
R1:= (a woman, at, table)
R2:= (table, in the middle of, a schoolyard)

**Sub-task 3:**
R1: REGION:EC
R2: REGION:EC & DISTANCE:MIDDLE

**INPUTS**          **OUTPUTS**
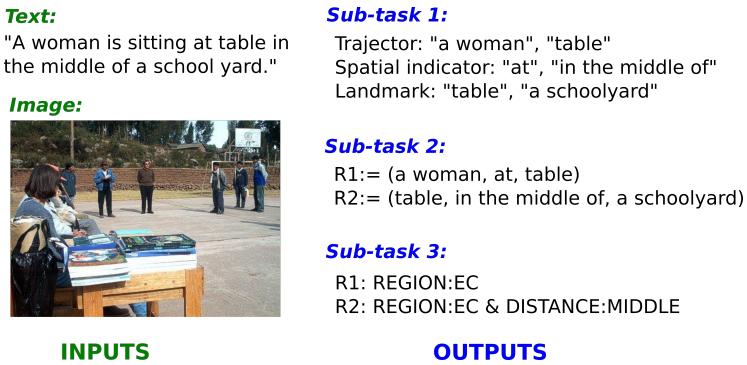
**Fig. 1.** Overview of the mSpRL task

In qualitative spatial representation and reasoning, spatial relations can be classified precisely (for example, RCC8 is a set of topological relations between regions). Identifying spatial relations using text only is a difficult task, due to the variety of meaning and interpretations words and sentences can have. Exploiting visual data could be paramount to recognize the spatial objects and their relations but gives rise to multi-modal alignment issues. In the dataset, images are segmented into annotated bounded boxes, and the spatial relations between these boxes is given. This spatial information enables us to enrich the textual data.

In our contribution, the extraction of spatial roles and relationships (sub-tasks 1 and 2) was done using the winning system of previous years [11], an implementation of this model being in a software called *Saul* [7]. It considers text solely as input (ignoring images) and returns a set of *trajector*, *spatial indicator* and *landmark* grouped into relations. Our contribution focuses on two aspects: using provided images as a complementary source of information and the relation classification sub-task. To do so, we handcraft a representation for spatial relations as a vector built using multi-modal inputs: the textual triplet and features from the associated image. We then train a SVM to classify the spatial relation. The SVM is trained to predict both general types (region, direction and distance labels) and specific values (EC, front, close, ...). Note that

the multiple labels can be associated with a single relation (as it is the case in the example of Figure 1).

The rest of this document is organized as follows: we describe our contribution including the classification pipeline and the design of relation embeddings. Finally, we present our experiments and the associated results.

## 2  Model

Our contribution mainly focuses on sub-task 3 since we used the previous state of the art model of [11] - re-implemented in *Saul* [7] - to perform sub-tasks 1 and 2. Sub-task 3 is a supervised classification problem in which the available input data is composed of 3 elements: the relation triplet, the original sentence from which the triplet was extracted, and an associated image. We convert that input data into a multi-modal embedding $e_{\text{relation}}$ that we describe in Section 2.1. We then use a linear SVM to classify the general types and specific values of the relations, the classification part being described in Section 2.2.

### 2.1  Relation Embedding

A relation is defined by visual (an image) and textual data (a triplet and the sentence from which it was extracted). We build our embedding by concatenating a textual embedding $e_{\text{text}}$ and an image embedding $e_{\text{image}}$.

$$e_{\text{relation}} = e_{\text{text}} \oplus e_{\text{image}}$$

*Textual embedding.* In our model, the text embedding contains information from the triplet only, as we drop the original sentence. Indeed, we assume that the useful information for the classification sub-task is contained in the extracted triplet, and that using the surrounding context of the sentence would lead to over-fitting of the model, given the small size of the training data.

We construct $e_{\text{text}}$ as follows:

$$e_{\text{text}} = u_{\text{trajector}} \oplus u_{\text{landmark}} \oplus \mathbb{1}_{\text{spatial indicator}}$$

where $u_{\star}$ is the average of the pre-trained embedding of the words that compose $\star$. In our experiments, we consider both Glove embeddings [9] and multi-modal word embedding described in [1]. $\mathbb{1}_{\star}$ is a one-hot encoding of the spatial-indicator $\star$ of the relation ; we use a fixed lexicon of 77 spatial-indicators as they are limited in number. $\oplus$ denotes the concatenation operator. Given the small amount of training data, ideally, $e_{\text{text}}$ has a small dimension to prevent over-fitting. With that objective in mind, we project word embeddings in a space of reduced dimensions ; we consider both random projections and Principal Component Analysis (PCA) techniques conducted on the training database.

*Visual embedding.* Segmented images and pre-computed visual features are provided in the dataset. A label is provided for each region of the segmented image. The given visual feature of a region is a 27-dimensional vector containing low-level information such as region area, width and height of the region, mean and standard deviation of height and width in the $x$ and $y$ axis respectively, the boundary/area ratio, convexity, average, standard deviation and skewness in RGB and CIE-Lab color spaces. Segmenting the images was done manually and visual features of all the regions were computed in [3] and are included in the dataset. We construct $e_{\text{image}}$ as follows:

$$e_{\text{image}} = r_{\text{trajector}} \oplus r_{\text{landmark}} \oplus r_{\text{spatial}}$$

where $r_\star$ is the visual embedding of the region $\star$ ; we find the matching region of a landmark or a trajector by taking the region annotated with the most similar word (i.e. we compute cosine similarity scores on word embeddings). In $r_{\text{spatial}}$, we encode in a one-hot vector the connectivity relations between the landmark and trajector regions: adjacent/disjoint, beside/x-aligned, above/below/y-aligned. This information is also provided as input data. As with textual embedding, to avoid over-fitting of the SVM classifier, we project the $r_\star$ vectors in a space of smaller dimension.

Note that the relation embeddings are hand-crafted and they remain fixed during training. The main reason for that is to reduce over-fitting of the model given the small size of the training dataset.

## 2.2   Classification

Once the embeddings of spatial relations are built (as explained in section 2.1), we use linear Support Vector Machines for classification [2], according to three strategies as shown in Figure 2:

- **Mono-label**: predicts a single label corresponding to specific values. Multi-labels are considered as distinct classes. This gives a total of 28 classes (we remove multi-label classes that do not occur in the training set). The general type is simply deduced from the specific values.
- **Multi-label**: predicts multiple labels corresponding to the specific values. We use the One-vs-Rest (OvR) strategy to do so, which gives a total of 19 classes. The general type is simply deduced from the specific values.
- **Hierarchical Multi-label**: first predicts multiple labels corresponding to the general types, then uses appropriate classifiers (each one trained on a particular general type) to predict multiple labels corresponding to the specific values for each of the predicted general type. That gives us a total of 4 classifiers (one to determine the general types and one for each possible general type).

**Mono-label**

Relation —SVM→ ( DC, EC, ..., middle, close, ..., EC/middle, DC/close)

**Multi-label**

Relation —OvR SVM→ ( DC, EC, ..., middle, close)

**Hierarchical Multi-label**

Relation —OvR SVM→ (region, direction, distance)
                        OvR|SVM          OvR|SVM
                   (DC, EC, ...)    (middle, close, ...)
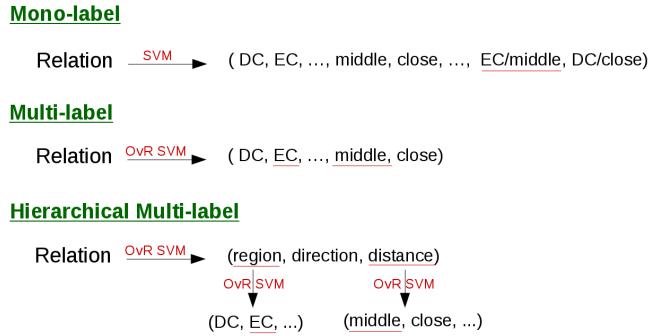
**Fig. 2.** Overview of the considered classification strategies

# 3 Experiments

## 3.1 Sub-tasks 1 and 2: Spatial Relation Extraction

*Sub-Task 1. Identification of spatial entities.* A sparse perceptron classifier is trained for each role: Trajector, Spatial Indicator, and Landmark. The features are designed using a set of lexical, syntactical, and contextual features (lexical surface of the phrases, headwords phrases, POS-tags, dependency relations, subcategorization, etc.). Results are presented in Table 1.

| Role | Precision | Recall | F1 | LCount | PCount |
|---|---|---|---|---|---|
| Trajector | 79.29 | 53.43 | 63.84 | 874 | 589 |
| Spatial Indicator | 97.59 | 61.13 | 75.17 | 95 | 498 |
| Landmark | 94.05 | 60.73 | 73.81 | 573 | 370 |
| Overall | 89.55 | 58.30 | 70.41 | 2242 | 1457 |

**Table 1.** Results of Saul's baseline on test data for spatial roles prediction. LCount stands for the count of labels in the gold data and PCount stands for the number of predicted labels

*Sub-Task 2. Identification of spatial relations.* In order to classify spatial relations, two binary classifiers are trained on pairs of phrases: one takes as input Trajector-Spatial Indicator pairs, the other on considers Spatial Indicator-Landmark pairs. With the perceptron assigned to Spatial Indicators, the indicator candidates are found, and all possible role-indicator pairs are possible candidates for the binary classifiers trained earlier. At the end, pairs with common indicators are the final triplets. Results are presented in Table 2.

| | Precision | Recall | F1 | LCount | PCount |
|---|---|---|---|---|---|
| Overall | 68.33 | 48.03 | 56.41 | 939 | 660 |

**Table 2.** Results of Saul's baseline on test data for spatial relations extraction

### 3.2 Sub-task 3: Spatial Relation Classification

This sub-task is the main focus of our contribution since we were interested in using multi-modal embedding for classifying spatial relations. For this purpose, we run two different scenarios:

- Our submitted best model (no image) which uses the mono-label classification strategy. We have $e_{\text{relation}} = e_{\text{text}}$ as $e_{\text{image}}$ is ignored. Word embeddings are projected in a space of dimension 25 with PCA (outperforming random projection). Glove embeddings are used as they outperformed multi-modal embeddings of [1].
- Our submitted model with image is the same than the model without image but the relation embedding includes $r_{\text{spatial}}$ in addition: $e_{\text{relation}} = e_{\text{text}} \oplus r_{\text{spatial}}$

*Overall results.* Table 3 presents the obtained results for sub-task 3 for our scenarios with respect to two baselines:

- Organizer's baseline: The features for the type classifiers are the concatenation of the features from sub-task 1 for each argument of the triplet.
- Best model which is the same model as the one submitted without image but the word embeddings are not projected in a lower-dimensional space and stay unchanged.

| | Precision | Recall | F1 |
|---|---|---|---|
| Organizer's baseline | 47.77 | 23.49 | 27.00 |
| Submitted model with image | 56.49 | 39.04 | 43.54 |
| Submitted model (no image) | **58.74** | 40.72 | 45.64 |
| Best model | 58.66 | **41.39** | **46.29** |

**Table 3.** Global scores on test data for the specific value prediction task

Please, note that all hyper-parameters are chosen with a 5-fold cross-validation computation on the training set. Our model, with its different settings reaches better scores (precision, recall and F1) than the baseline by a large margin. [6] reports comparable and slightly better results with a 10-fold cross-validation. For more detailed results, we refer the reader to Appendix A.

*Classification strategies.* To refine these results, we also compare the different classification strategies, as shown in Table 4. Mono-label classification appears to work better than other strategies. Interestingly, hierarchical strategy gives a higher precision but worse recall (and overall worse F1 score). A joint model, coupling mono-label and hierarchical models, might lead to even better performance by taking advantage of both models (best precision for Hierarchical, best recall for Mono-label).

| Classif. type | Precision | Recall | F1 |
|---|---|---|---|
| Mono-label | 58.66 | **41.39** | **46.29** |
| Multi-label | 62.14 | 39.49 | 45.32 |
| Hierarchical | **62.17** | 40.05 | 45.88 |

**Table 4.** Global score on test data for the specific value prediction task, using several classification techniques

*Influence of the components of the embeddings.* At a finer level, we also measure the influence of each component of the global relation embedding vector $e_{\text{relation}}$ with an ablation study. Instead of using full relation embeddings with all of their components, we remove one or several parts of them and we report in Table 5 obtained results of classification models trained on those partial relation embeddings. Each line of the table contains the results of a 5-fold cross-validation training on $e_{\text{relation}}$ without the ablated part, namely image, text, spatial indicator, visual region embeddings. This experimentally highlights the importance of which spatial indicator is being used and the textual embeddings for the trajector and then landmark. Visual parts of the relation embedding are useless or even harmful for the overall performance.

| Ablated part | Precision | Recall | F1 |
|---|---|---|---|
| None | 87.26 | 87.32 | 85.57 |
| $e_{\text{image}}$ | **87.60** | **88.49** | **86.50** |
| $e_{\text{text}}$ | 71.67 | 45.95 | 28.85 |
| $r_{\text{spatial}}$ | 87.04 | 87.62 | 85.95 |
| $r_{\star}$ | 85.76 | 86.51 | 84.70 |
| $\mathbb{1}_{\text{indicator}}$ | 75.80 | 75.69 | 72.16 |
| $u_{\star}$ | 74.72 | 73.03 | 70.34 |

**Table 5.** Ablation study of the influence of the different parts of the relation embedding

While it is unclear why using the visual embeddings slightly degrades the overall performance, we note that many images do not contain areas for the entities found by sub-task 1 in the sentence (for example "bench" is found as trajector in the sentence "a bench in a park" but the associated image might

not contain any region labeled "bench"). Moreover, sometimes our algorithm misses the good region: for example sub-task 1 gives the entity "head" in the text but there is no "head" region in the image but rather "face-of-person". Despite some handcrafted rules that we have added to account for this problem, lots of regions are not considered. Also, even though high-level features from the images are provided, we assume that there are not enough images in the training set to learn something complementary from the text. Eventually, since the classification labels of sub-task 3 are only gold labels prone to annotation subjectivity, a human annotator would not get 100% f1-score. It would then be interesting to know about human performance on that sub-task for a comparison with our results.

*Word embedding dimension influence.* Since textual embeddings are proved to be major components of the spatial relation representation, we evaluate the impact of the choice of the dimension of the space in which word embeddings are projected with PCA. We make that parameter vary while others are kept fixed, in the experiment reported in Table 3. We can see that increasing the word embedding dimension improves the effectiveness of our approach and our best performing model does not project word embeddings and keep them unchanged. Intuitively, having too high-dimension embeddings leads to more parameters and a higher risk in over-fitting the model. For a good trade-off between performance and limited size for relation embeddings, 50 is also a suitable choice.
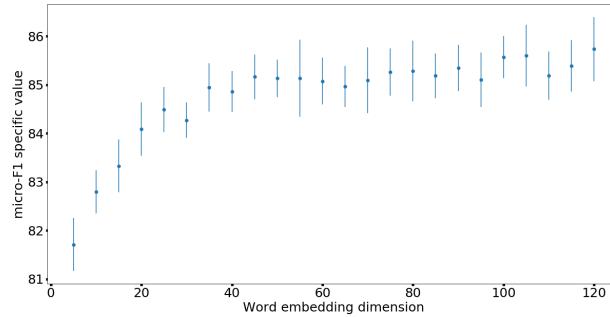


**Fig. 3.** Influence of the word embedding dimension on the F1 score. Vertical bars are the standard deviation of the results on 10 experiments, where each experiment is a 5-fold cross-validation

## 4  Conclusion and Future Work

In this work, we focused on sub-task 3 of the mSpRL lab of CLEF 2017: predicting general types and specific values for relations. Our system relies on a baseline

to extract spatial roles and relations from raw textual data. We build fixed embeddings for spatial triplets, and a linear SVM classifies relations. Unfortunately, we were not able to use provided visual inputs in a profitable way as our best model ignores images. These results highlight that considering multi-modal data for enhancing natural language processing is a difficult task and requires more efforts in terms of model design.

As future work, we have two objectives. First, we want to use the image data for sub-tasks 1 and 2 in an end-to-end fashion, as visual information might be useful to disambiguate between several candidate relations. Our other goal aims at addressing the problem of the limited quantity of training data: we wish to explore transfer learning techniques to train spatial word embeddings on auxiliary tasks.

## Acknowledgements

## A  Detailed results of the best performing model

| General type | Precision | Recall | F1 | LCount | PCount |
|---|---|---|---|---|---|
| Direction | 65.10 | 43.60 | 52.22 | 445 | 298 |
| Region | 68.16 | 51.00 | 58.34 | 449 | 336 |
| Distance | 66.67 | 22.22 | 33.33 | 9 | 3 |
| Region/Distance | 73.91 | 50.00 | 59.65 | 34 | 23 |
| Direction/Direction | 100,00 | 0.00 | 0.00 | 1 | 0 |
| Region/Direction | 100,00 | 0.00 | 0.00 | 1 | 0 |
| Overall | 66.97 | 47.07 | 5.13 | 939 | 660 |

**Table 6.** Coarse-grained results on general type and specific value predictions for the best model.

| Specific Value | Precision | Recall | F1 | LCount | PCount |
|---|---|---|---|---|---|
| left | 59.83 | 54.26 | 56.91 | 129 | 117 |
| front | 78.08 | 56.44 | 65.52 | 101 | 73 |
| right | 53.06 | 29.55 | 37.96 | 88 | 49 |
| behind | 68.57 | 40.00 | 50.53 | 60 | 35 |
| above | 63.64 | 23.73 | 34.57 | 59 | 22 |
| below | 100.00 | 22.22 | 36.36 | 9 | 2 |
| TPP | 60.35 | 48.39 | 53.71 | 217 | 174 |
| EC | 45.75 | 50.36 | 47.95 | 139 | 153 |
| DC | 28.57 | 2.94 | 5.33 | 68 | 7 |
| PO | 100.00 | 0.00 | 0.00 | 10 | 0 |
| NTPP | 0.00 | 0.00 | 0.00 | 4 | 2 |
| EQ | 100.00 | 0.00 | 0.00 | 1 | 0 |
| TPPI | 100.00 | 0.00 | 0.00 | 1 | 0 |
| Overall | 58.66 | 41.39 | 46.29 | 894 | 634 |

**Table 7.** Fine-grained results on general type and specific value predictions for the best model.

# References

1. Collell Talleda, G., Zhang, T., Moens, M.F.: Imagined visual representations as multimodal embeddings. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). AAAI (2017)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
3. Escalante, H.J., Hernández, C.A., Gonzalez, J.A., López-López, A., Montes, M., Morales, E.F., Sucar, L.E., Villaseñor, L., Grubinger, M.: The segmented and annotated iapr tc-12 benchmark. Computer Vision and Image Understanding 114(4), 419–428 (2010)
4. Gareth J. F. Jones, Séamus Lawless, J.G.L.K.L.G.T.M.L.C.N.F. (ed.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings, vol. 10456
5. Kordjamshidi, P., Bethard, S., Moens, M.F.: Semeval-2012 task 3: Spatial role labeling. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. pp. 365–373. Association for Computational Linguistics (2012)
6. Kordjamshidi, P., Moens, M.F.: Global machine learning for spatial ontology population. Web Semantics: Science, Services and Agents on the World Wide Web 30, 3–21 (2015)
7. Kordjamshidi, P., Roth, D., Wu, H.: Saul: Towards declarative learning based programming. In: IJCAI: proceedings of the conference/sponsored by the International Joint Conferences on Artificial Intelligence. vol. 2015, p. 1844. NIH Public Access (2015)
8. Linda Cappellato, Nicola Ferro, L.G., Mandl, T. (eds.): CLEF 2017 Labs Working Notes

9. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)

10. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. KR 92, 165–176 (1992)

11. Roberts, K., Harabagiu, S.M.: Utd-sprl: A joint approach to spatial role labeling. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. pp. 419–424. Association for Computational Linguistics (2012)