CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview

Parisa Kordjamshidi[1], Taher Rahgooy[1], Marie-Francine Moens[2], James Pustejovsky[3], Umar Manzoor[1] and Kirk Roberts[4]

[1] Tulane University
[2] Katholieke Universiteit Leuven
[3] Brandies University
[4] The University of Texas Health Science Center at Houston

Abstract. The extraction of spatial semantics is important in many real-world applications such as geographical information systems, robotics and navigation, semantic search, etc. Moreover, spatial semantics are the most relevant semantics related to the visualization of language. The goal of multimodal spatial role labeling task is to extract spatial information from free text while exploiting accompanying images. This task is a multimodal extension of spatial role labeling task which has been previously introduced as a semantic evaluation task in the SemEval series. The multimodal aspect of the task makes it appropriate for the CLEF lab series. In this paper, we provide an overview of the task of multimodal spatial role labeling. We describe the task, sub-tasks, corpora, annotations, evaluation metrics, and the results of the baseline and the task participant.

1 Introduction

The multimodal spatial role labeling task (mSpRL) is a multimodal extension of the spatial role labeling shared task in SemEval-2012 [5]. Although there were proposed extensions of the data and the task in more extensive schemes in Kolomiyets et al. [4] and Pustejovsky et al. [13], the SemEval-2012 data was more appropriate for the goal of incorporating the multimodality aspect. SemEval-2012 annotates CLEF IAPRTC-12 Image Benchmark [1], which includes touristic pictures along with a textual description of the pictures. The descriptions are originally provided in multiple languages though we use the English annotations for the purpose of our research.

The goal of mSpRL is to develop natural language processing (NLP) methods for extraction of spatial information from both images and text. Extraction of spatial semantics is helpful for various domains such as semantic search, question answering, geographical information systems, and even in robotic settings when giving robots navigational instructions or instructions for grabbing and manipulating objects. It is also essential for some specific tasks such as text to scene conversion (or vice-versa), scene understanding as well as general information retrieval tasks when using a huge amount of available multimodal data from various resources. Moreover, we have noticed an increasing interest in the extraction of spatial information from medical images that are accompanied by natural language descriptions. The textual descriptions of a subset of images are annotated with spatial roles according to spatial role labeling annotation

scheme [7]. We should note that considering the vision and language modalities and combining the two media has become a very popular research challenge nowadays. We distinguish our work and our data from the existing research related to vision and language (inter alia, [11, 3]) in considering explicit formal spatial semantics representations and providing direct supervision for machine learning techniques by our annotated data. The formal meaning representation would help to exploit explicit spatial reasoning mechanisms in the future. In the rest of this overview paper, we introduce the task in Section 2; we describe the annotated corpus in Section 3; the baseline and the participant systems are described in Section 4; Section 5 reports the results and the evaluation metrics. Finally, we conclude in Section 6.

2 Task Description

The task of text-based spatial role labeling (SpRL) [8] aims at mapping natural language text to a formal spatial meaning representation. This formal representation includes specifying spatial entities based on cognitive linguistic concepts and the relationships between those entities, in addition to the type of relationships in terms of qualitative spatial calculi models. The ontology of the target concepts is drawn in Figure 1 and the concepts are described later in this section. The applied ontology includes a subset of concepts proposed in the scheme described in [7]. We divide this task to three sub-tasks. To clarify these sub-tasks, we use the example of Figure 2. This figure shows a photograph and a few English sentences that describe it. Given the first sentence *"About 20 kids in traditional clothing and hats waiting on stairs."*, we need to do the following tasks:

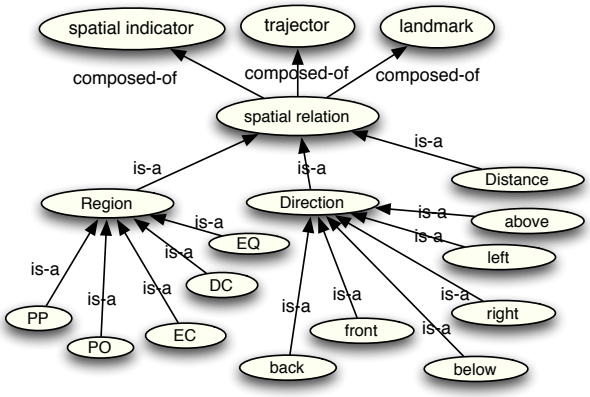

Fig. 1. Given spatial ontology [9]

– **Sub-task 1:** The first task is to identify the phrases that refer to spatial entities and classify their roles. The spatial roles include a) spatial indicators, b) trajectors, c)

landmarks. Spatial indicators indicate the existence of spatial information in a sentence. Trajector is an entity whose location is described and landmark is a reference object for describing the location of a trajector. In the above-mentioned sentence, the location of *about 20 kids* that is the *trajector* has been described with respect to the *the stairs* that is the *landmark* using the preposition *on* that is the *spatial indicator*. These are examples the spatial roles that we aim to extract form the sentence.

– **Sub-task 2:** The second sub-task is to identify the relations/links between the spatial roles. Each spatial relation is represented as a triplet of (spatial-indicator, trajector, landmark). Each sentence can contain multiple relations and individual phrases can even take part in multiple relations. Furthermore, occasionally roles can be implicit in the sentence (i.e., a null item in the triplet). In the above example, we have the triplet *(*kids,on,stairs) that form a spatial relation/link between the three above mentioned roles. Recognizing the spatial relations is very challenging because there could be several spatial roles in the sentence and the model should be able to recognize the right connections. For example (waiting, on, stairs) is a wrong relation here because "kids" is the trajector in this sentence not "waiting".
– **Sub-task 3:** The third sub-task is to recognize the type of the spatial triplets. The types are expressed in terms of multiple formal qualitative spatial calculi models similar to Figure 1. At the most course-grained level the relations are classified into three categories of topological (regional), directional, or distal. Topological relations are classified according to the well-known RCC (regional connection calculus) qualitative representation. An RCC5 version that is shown in Figure 1 includes Externally connected (EC), Disconnected (DC), Partially overlapping (PO), Proper part (PP), and Equality (EQ). The data is originally annotated by RCC8 which distinguishes between Proper part (PP), Tangential proper part (TPP) and Inverse tangential proper part inverse (TPPI). For this lab the original RCC8 annotations are used. Directional relations include 6 relative directions: left, right, above, below, back, and front. In the above example, we can state the type of relation between the roles in the triplet (kids,on,stairs) is "above". In general, we can assign multiple types to each relation. This is due to the polysemy of spatial prepositions as well as the difference between the level of specificity of spatial relations expressed in the language compared to formal spatial representation models. However, multiple assignments are not frequently made in our dataset.

The task that we describe here is similar to the specifications that are provided in Kordjamshidi et al. [9], however, the main point of this CLEF lab was to provide an additional resource of information (the accompanying images) and investigate the ways that the images can be exploited to improve the accuracy of the text-based spatial extraction models. The way that the images can be used is left open to the participants. Previous research has shown that this task is very challenging [8], particularly given the small set of available training data and we aim to investigate if using the images that accompany textual data can improve the recognition of the spatial objects and their relations. Specifically, our hypothesis is that the images could improve the recognition of the type of relations given that the geometrical features of the boundaries of the objects in the images are closer to the formal qualitative representations of the relationships compared to the counterpart linguistic descriptions.

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

3 Annotated Corpora

The annotated data is a subset of the IAPR TC-12 image Benchmark [1]. It contains 613 text files with a total of 1,213 sentences. The original corpus was available without copyright restrictions. The corpus contains 20,000 images taken by tourists with textual descriptions in up to three languages (English, German, and Spanish). The texts describe objects and their absolute or relative positions in the image. This makes the corpus a rich resource for spatial information. However the descriptions are not always limited to spatial information which makes the task more challenging. The data has been annotated with the roles and relations that were described in Section 2, and the annotated data can be used to train machine leaning models to do this kind of extractions automatically. The text has been annotated in previous work (see [7, 6]). The role annotations are provided on phrases rather than single words. The statistics about the data is given in Table 1. For this lab, we augmented the textual spatial annotations with a reference to the aligned images in the xml annotations and fixed some of the annotation mistakes to provide a cleaner version of the data.

4 System Descriptions

We, as organizers of the lab, provided a baseline inspired by previous research for the sake of comparison. The shared task had one official participant who submitted two systems. In this section, we describe the submitted systems and the baseline.

– **Baseline:** For sub-task 1 and classifying each role (Spatial Indicator, Trajector, and Landmark), we created a sparse perceptron binary classifier that uses a set of

	Train	Test	All
Sentences	600	613	1213
Trajectors	716	874	1590
Landmarks	612	573	1185
Spatial Indicators	666	795	1461
Spatial Relations	761	939	1700
Region	560	483	1043
Direction	191	449	640
Distance	40	43	83

Table 1. The statistics of the annotated CLEF-Image Benchmark, some of the spatial relations are annotated with multiple types, e.g., having both region and direction labels.

lexical, syntactical, and contextual features, such as lexical surface patterns, headwords phrases, part-of-speech tags, dependency relations, subcategorization, etc. For classifying the spatial relations, we first trained two binary classifiers on pairs of phrases. One classifier detects Trajector-SpatialIndicator pairs and another detects Landmark-SpatialIndicator pairs. We used the spatial indicator classifier from sub-task 1 to find the indicator candidates and considered all noun phrases as role candidates. Each combination of SpatialRole-SpatialIndicator candidates considered as a pair candidate and the pair classifiers are trained on. We used a number of relational features between the pairs of phrases such as distance, before, etc to classify them. In the final phase, we combined the predicted phrase pairs that have a common spatial indicator in order to create the final relation/triplet for subtask 2. for example if (kids,on) pair is classified as Trajector-SpatialIndicator and (stairs,on) is predicted as Landmark-SpatialIndicator then we generate the triplet, (on,kids,stairs) as a spatial triplet since both trajector and landmark relate to the same preposition 'on'. The features of this baseline model are inspired by the work in [9]. For sub-task 3 and training general type and specific value classifiers, we used a very naive pipeline model as the baseline. In this pipeline, the predicted triplets from the last stage are used for training the relations types. For these type classifiers, simply, the phrase features of each argument of the triplets are concatenated and used as features. Obviously, we miss a large number of relations at the stage of spatial relation extraction in sub-task 2 since we depend on its recall.

– **LIP6:** The LIP6 group built a system for sub-task 3 that classifies relation types. For the sub-task 1 and 2, the proposed model in Roberts and Harabagiu [14] was used. Particularly, an implementation of that model in the Saul [10] language/library was applied. For every relation, an embedding is built with available data: the textual relation triplet and visual features from the associated image. Pre-trained word embeddings are used [12] to represent the trajector and landmark and a one-hot vector indicates which spatial indicator is used; the visual features and embeddings from the segmented regions of the trajectors and landmarks are extracted and projected into a low dimensional space. Given those generated embeddings, a linear SVM model is trained to classify the spatial relations and the embeddings remain fixed. Several experiments were made to try various classification modes and dis-

cuss the effect of the model parameters, and more particularly the impact of the visual modality. As the best performing model ignores the visual modality, these results highlight that considering multimodal data for enhancing natural language processing is a difficult task and requires more efforts in terms of model design.

5 Evaluation Metrics and Results

About 50% of the data was used as the test set for the evaluation of the systems. The evaluation metrics were precision, recall, and F1-measure, defined as:

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}, \quad F1 = \frac{2 * recall * precision}{(recall + precision)}$$

where, TP (true positives) is the number of predicted components that match the ground truth, FP (false positives) is the number of predicted components that do not match the ground truth, and FN (false negatives) is the number of ground truth components that do not match the predicted components. These metrics are used to evaluate the performance on recognizing each type of role, the relations and each type of relation separately. Since the annotations are provided based on phrases, the overlapping phrases are counted as correct predictions. The evaluation with exact matching between phrases would provide lower performance than the reported ones. The relation type evaluation for sub-task 3 includes course- and fine-grained metrics. The coarse-grained metric (overall-CG) averages over the labels of region, direction, and distance. The fine-grained metric (overall-FG) shows the performance over all lower-level nodes in the ontology including the RCC8 types (e.g., EC) and directional relative types (e.g., above, below).

Table 2 shows the results of our baseline system that was described in the previous section. Though the results of the roles and relation extraction are fairly comparable to the state of the art [14, 9], the results of the relations type classifiers are less matured because a simple pipeline, described in Section 4, was used. Table 3 shows the results of the participant systems.

As mentioned before, LIP6 uses the model suggested in [14] and its implementation in Saul [10] for sub-task 1 and sub-task 2. It has a focus in designing a model for sub-task 3. The experimental results using textual embeddings alone are shown under *text only* in the table, and a set of results are reported by exploiting the accompanying images and training the visual embeddings from the corpora. The LIP6's system significantly outperforms the provided baseline for relation type classifiers. Despite our expectations, the results that use the visual embeddings perform worse than the one that ignores images. In addition to the submitted systems, the LIP6 team improved their results slightly by using a larger feature size in their dimensionality reduction procedure with their text-only features. This model outperforms their submitted systems and is listed in Table 3 as *Best model*.

Discussion. The previous research and the results of LIP6 team show this task is challenging, particularly, using this small set of training data. LIP6 was able to outperform the provided baseline using the textual embeddings for relation types but the

Label	P	R	F1
SP	94.76	97.74	96.22
TR	56.72	69.56	62.49
LM	72.97	86.21	79.04
Overall	74.36	83.81	78.68
Triplets	75.18	45.47	56.67
Overall-CG	64.72	37.91	46.97
Overall-FG	47.768	23.490	26.995

Table 2. Baseline: classic classifiers and linguistically motivated features based on [9]

Label		P	R	F1
SP		97.59	61.13	75.17
TR		79.29	53.43	63.84
LM		94.05	60.73	73.81
Overall		89.55	58.03	70.41
Triplets		68.33	48.03	56.41
Text only	Overall-CG	63.829	44.835	52.419
	Overall-FG	56.488	39.038	43.536
Text+Image	Overall-CG	66.366	46.539	54.635
	Overall-FG	58.744	40.716	45.644
Best model	Overall-CG	66.76	46.96	55.02
	Overall-FG	58.20	41.05	45.93

Table 3. LIP6 performance with various models for Sub-task 3; LIP6 uses Roberts and Harabagiu [14] for Sub-tasks 1 and 2.

results of combining the images, in the contrary, dropped the performance. This result indicates that integrating the visual information needs more investigation otherwise it can only add noise to the learning system. One very basic question to be answered is whether the images of this specific dataset can potentially provide complementary information or help resolving ambiguities in the text at all; this investigation might need a human analysis. Although the visual embeddings did not help the best participant system with the current experiments, using other alternative embeddings trained from large corpora might help improving this task. Given the current interest of the vision and language communities in combining the two modalities and the benefits that this trend will have for the information retrieval, there are many new corpora becoming available(e.g. [11]) which can be valuable sources of information for obtaining appropriate join features. There is a separate annotation on the same benchmark that includes the ground-truth of the co-references in the text and image [2]. This annotation has been generated for co-reference resolution task but it seems to be very useful to be used on top of our spatial annotations for finding better alignment between spatial roles and image segments. In general, current related language and vision resources do not consider formal spatial meaning representation but can be used indirectly to train informative representations or be used as source for indirect supervision for extraction of formal spatial meaning.

6 Conclusion

The goal of the multimodal spatial role labeling lab was to provide a benchmark to investigate how adding grounded visual information can help understanding the spatial semantics of natural language text and mapping language to a formal spatial meaning representation. The prior hypothesis has been that the visual information should help the extraction of such semantics because spatial semantics are the most relevant semantics for visualization and the geometrical information conveyed in the vision media should be able to easily help in disambiguation of spatial meaning. Although, there are many recent research works on combining vision and language, none of them consider obtaining a formal spatial meaning representation as a target nor provide supervision for training such representations. However, the experimental results of our mSpRL lab participant show that even given ground truth segmented objects in the images and having the exact geometrical information about their relative positions, adding useful information for understanding the spatial meaning of the text is very challenging. The experimental results indicate that using the visual embeddings and using the similarity between the objects in the image and spatial entities in the text can turn to adding noise to the learning system reducing the performance. However, we believe our prior hypothesis is still valid, but finding an effective way to exploit vision for spatial language understanding, particularly obtaining a formal spatial representation appropriate for explicit reasoning, remains as an important research question.

References

1. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR benchmark: a new evaluation resource for visual information systems. In: Proceedings of the International Conference on

Language Resources and Evaluation (LREC). pp. 13–23 (2006)
2. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referit game: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
3. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. CoRR abs/1411.2539 (2014), http://arxiv.org/abs/1411.2539
4. Kolomiyets, O., Kordjamshidi, P., Moens, M., Bethard, S.: Semeval-2013 task 3: Spatial role labeling. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 255–262. Atlanta, Georgia, USA (June 2013)
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: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval). vol. 2, pp. 365–373 (2012)
6. Kordjamshidi, P., van Otterlo, M., Moens, M.: Spatial role labeling annotation scheme. In: Pustejovsky J., I.N. (ed.) Handbook of Linguistic Annotation. Springer Verlag (2015)
7. Kordjamshidi, P., van Otterlo, M., Moens, M.F.: Spatial role labeling: task definition and annotation scheme. In: Calzolari, N., Khalid, C., Bente, M. (eds.) Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). pp. 413–420 (2010)
8. Kordjamshidi, P., van Otterlo, M., Moens, M.F.: Spatial role labeling: towards extraction of spatial relations from natural language. ACM - Transactions on Speech and Language Processing 8, 1–36 (2011)
9. Kordjamshidi, P., Moens, M.F.: Global machine learning for spatial ontology population. Web Semant. 30(C), 3–21 (Jan 2015)
10. Kordjamshidi, P., Wu, H., Roth, D.: Saul: Towards declarative learning based programming. In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI) (7 2015)
11. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Li, F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (2017)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
13. Pustejovsky, J., Kordjamshidi, P., Moens, M.F., Levine, A., Dworman, S., Yocum, Z.: SemEval-2015 task 8: SpaceEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 9th international workshop on semantic evaluation (SemEval 2015), Denver, Colorado, 4-5 June 2015. pp. 884–894. ACL (2015), https://lirias.kuleuven.be/handle/123456789/500427
14. Roberts, K., Harabagiu, S.: UTD-SpRL: A joint approach to spatial role labeling. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval'12). pp. 419–424 (2012)