Combining Geometric, Textual and Visual Features for Predicting Prepositions in Image Descriptions

Arnau Ramisa*[1] **Josiah Wang*[2]** **Ying Lu[3]** **Emmanuel Dellandrea[3]**
Francesc Moreno-Noguer[1] **Robert Gaizauskas[2]**

[1] Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Barcelona, Spain
[2] Department of Computer Science, University of Sheffield, UK
[3] LIRIS, École Centrale de Lyon, France

{aramisa, fmoreno}@iri.upc.edu {j.k.wang, r.gaizauskas}@sheffield.ac.uk
{ying.lu, emmanuel.dellandrea}@ec-lyon.fr

Abstract

We investigate the role that geometric, textual and visual features play in the task of predicting a preposition that links two visual entities depicted in an image. The task is an important part of the subsequent process of generating image descriptions. We explore the prediction of prepositions for a pair of entities, both in the case when the labels of such entities are known and unknown. In all situations we found clear evidence that all three features contribute to the prediction task.

1 Introduction

In recent years, there has been an increased interest in the task of automatic generation of natural language image descriptions at sentence level, compared to earlier work that annotates images with a laundry list of terms (Duygulu et al., 2002). The task is important in that such detailed annotations are more informative and discriminative compared to isolated textual labels, and are essential for improved text and image retrieval.

The most standard approach to generating such descriptions involves first detecting instances of pre-defined concepts in the image, and then reasoning about these concepts to generate image descriptions e.g. (Kulkarni et al., 2011; Yang et al., 2011). Our work is also based on this paradigm. However, we assume that object instances have already been pre-detected by visual recognisers, and concentrate on a specific subtask of description generation. More specifically, given two visual entity instances where one could potentially act as a modifier to the other, we address the problem of identifying the appropriate preposition to connect these two entities (Figure 1). The inferred prepositional relations will subsequently act as an important intermediate representation towards the eventual goal of generating image descriptions.

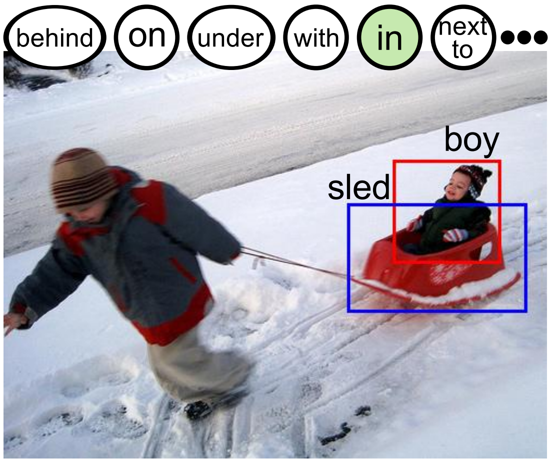

Figure 1: Given a subject *boy* and an object *sled* and their location in the image, what would the best preposition be to connect the two entities?

The main contribution of this paper is therefore to learn to predict the most suitable preposition given its context, and to learn this jointly from images and their descriptions. In particular, we concentrate on learning from (i) geometric relations between two visual entities from image annotations; (ii) textual features from textual descriptions; (iii) visual features from images. Previous work exists (Yang et al., 2011) that uses text corpora to 'guess' the prepositions given the context without considering the appropriate spatial relations between the entities in the image, signifying a gap between visual content and its corresponding description. For example, although *person on horse* might commonly occur in text corpora, a particular image might actually depict a person standing beside a horse. On the other hand, work that does consider the image content for generating prepositions (Kulkarni et al., 2011; Elliott and Keller, 2013) map geometric relations to a limited set of prepositions using manually defined rules,

*A. Ramisa and J. Wang contributed equally to this work.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 214–220,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

not as humans would naturally use them with a richer vocabulary. We would like to have the best of both worlds, by considering image content as well as textual information to select the preposition best used to express the relation between two entities. Our hypothesis is that the combination of geometric, textual and visual features can help with the task of predicting the most appropriate preposition, since incorporating geometric and visual information should help generate a relation that is consistent with the image content, whilst incorporating textual information should help generate a description that is consistent with natural language.

2 Related Work

The Natural Language Processing Community has significant interest in different aspects of prepositions. The Prepositions Project (Litkowski and Hargraves, 2005) analysed and produced a lexicon of English prepositions and their senses, and subsequently used them in the Word Sense Disambiguation of Prepositions task in SemEval-2007 (Litkowski and Hargraves, 2007). In SemEval-2012, Kordjamshidi et al. (2012) introduce the more fine-grained task of spatial role labelling to detect and classify spatial relations expressed by triples *(trajector, landmark, spatial_indicator)*. In the latest edition of SemEval-2015, the SpaceEval task (Pustejovsky et al., 2015) introduce further tasks of identifying spatial and motion signals, as well as spatial configurations/orientation and motion relation.

In work that links prepositions more strongly to image content, Gupta and Davis (2008) model prepositions implicitly to disambiguate image regions, rather than for predicting prepositions. Their work also require manual annotation of prepositional relations. In image description generation work, Kulkarni et al. (2011) manually map spatial relations to pre-defined prepositions, whilst Yang et al. (2011) predict prepositions from large-scale text corpora solely based on the complement term, with the prepositions constrained to describing scenes (*on the street*). Elliott and Keller (2013) define a list of eight spatial relations and their corresponding prepositional term for sentence generation. Although they also present alternative models that use text corpora for descriptions that are more human-like, they are limited to verbs and do not cover prepositions. Le et al. (2014) examine prepositions modifying human actions (verbs), and conclude that these relate to positional information to a certain extent. Other related work include training classifiers for prepositions with spatial relation features to improve image segmentation and detection (Fidler et al., 2013); this work is however limited to four prepositions.

3 Task Definition

We formally define the task of predicting prepositions as follows: Let P be the set of possible prepositions. Let L be the set of possible *landmark* entities acting as the complement of a preposition, and let T be the set of possible *trajector* entities modified by the prepositional phrase comprising a preposition and its landmark[1]. For example, for the phrase *person on bicycle*, *on* would be the preposition, *bicycle* the landmark, and *person* the trajector. For this paper, we constrain *trajector* and *landmark* to be entities that are visually identifiable in an image since we are interested in discovering the role of visual features and geometric configurations between two entities in the preposition prediction task.

Let $D = \{d_1, d_2, ..., d_N\}$ be the set of N observations, where each d_i for $i = 1, 2..., N$ is represented by $d_i = (x_i, y_i, r_i)$, where x_i and y_i are the feature representations for the trajector and the landmark entities respectively, and r_i the relative geometric feature between the two visual entities.

Given d_i, the objective of the preposition prediction task is to produce a ranked list of prepositions $(p_1, p_2, ...p_{|P|})$ according to how likely they are to express the appropriate spatial relation between the given trajector and landmark entities that are either known (Section 6.1) or only represented by visual features (Section 6.2).

4 Dataset

We base the preposition prediction task on two large-scale image datasets with human authored descriptions, namely MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014; Plummer et al., 2015). To extract instances of triples *(trajector, preposition, landmark)* from image descriptions, we used the Neural Network, transition-based dependency parser of Chen and Manning (2014) as implemented in Stanford CoreNLP (Manning et al., 2014). Dependencies signifying prepositional

[1] The terminologies *trajector* and *landmark* are adopted from spatial role labelling (Kordjamshidi et al., 2011)

Bounding Box feature (number of dimensions)
• Vector (x, y) from centroid of trajector to centroid of landmark, normalised by the size of the bounding box enclosing both objects (2)
• Area of trajector bounding box relative to landmark (1)
• Aspect ratio of each bounding box (2)
• Area of each bounding box w.r.t. enclosing box (2)
• Intersection over union of the bounding boxes (1)
• Euclidean distance between the trajector and landmark bounding boxes, normalised by the image size (1)
• Area of each bounding box w.r.t. the whole image (2)

Table 1: Geometric features derived from bounding boxes.

relations are retained where both the governor and its dependent overlap with the entity mentions in the descriptions, and where both mentions have corresponding bounding boxes. The MSCOCO validation set is further annotated to remove errors arising from dependency parsing (notably PP attachment errors), and is used as our clean test set. Our final dataset comprises 8,029 training and 3,431 test instances for MSCOCO, and 46,847 training and 20,010 test instances for Flickr30k. Details on how the triples were extracted from captions and matched to instances in images are available in the supplementary material.

We consider two variants of trajector and landmark terms in our experiments: (i) using the provided high level categories as terms (80 for MSCOCO and 8 for Flickr30k); (ii) using the terms occurring in the sentence directly, which constitute a bigger and more realistic challenge. For Flickr30k, the descriptive phrases may cause data sparseness (*the furry, black and white dog*). Thus, we extracted the lemmatised head word of each phrase, using a 'semantic head' variant of the head finding rules of Collins (2003) in Stanford CoreNLP. Entities from the same coreference chain are denoted with a common head noun chosen by majority vote among the group, with ties broken by the most frequent head noun in the corpus, and further ties broken at random.

5 Features

Geometric Features: Geometric features between a trajector and a landmark entity are derived from bounding box annotations. We defined an 11-dimensional vector of bounding box features, covering geometric relations such as distance, orientation, relative bounding box sizes and overlaps between bounding boxes (Table 1). We chose to use continuous features as we felt these may be more powerful and expressive compared to discrete, binned features. Despite some of these features being correlated, we left it to the classifier to determine the most useful features for discrimination without having to withhold any unnecessarily.

Textual features: We consider two textual features to encode the trajector and landmark terms w_i^t and w_i^l. The first feature is a one-hot indicator vector x_i^I and y_i^I for the trajector and landmark respectively, where $x_{i,t}^I = 1$ if index t corresponds to the trajector term w_i^t and 0 elsewhere (and similarly for landmark). As data sparseness may be an issue, we also explore an alternative textual feature which encodes the terms as word2vec embeddings (Mikolov et al., 2013). This encodes each term as a vector such that semantically related terms are close in the vector space. This allows information to be transferred across semantically related terms during training (e.g. information from *person on boat* can help predict the preposition that mediates *man* and *boat*).

Image Features: While it is ideal to have vision systems produce a firm decision about the visual entity instance detected in an image, in reality it may be beneficial to defer the decision by allowing several possible interpretations of the instance being detected. In such cases, we will not have a single concept label for the entity, but instead a high-level visual representation. For this scenario, we extracted visual representations from the final layer of a Convolutional Neural Network trained on ImageNet (Krizhevsky et al., 2012), and used them as representations for entity instances in place of textual features.

6 Preposition Prediction

Here we highlight interesting findings from experiments performed for the task of predicting prepositions for two different scenarios (Sections 6.1 and 6.2). Detailed results can be found in the supplementary material.

Evaluation metrics. As there may be more than one 'correct' preposition for a given context (*person on horse* and *person atop horse*), we propose the mean rank of the correct preposition as the main evaluation metric, as it accommodates

		IND	W2V	GF	IND+GF	W2V+GF	Baseline
Mean rank	MSCOCO (max rank 17)	1.45	1.43	1.72	1.44	**1.42**	2.14
	MSCOCO (balanced)	3.20	3.10	4.60	3.00	**2.90**	5.40
	Flickr30k (max rank 52)	1.91	1.87	2.35	1.88	**1.85**	2.54
	Flickr30k (balanced)	11.10	9.04	15.55	10.23	**8.90**	15.13
Accuracy	MSCOCO	79.7%	80.3%	68.4%	79.8%	**80.4%**	40.2%
	MSCOCO (balanced)	52.5%	**54.2%**	31.5%	52.7%	53.9%	11.9%
	Flickr30k	75.4%	75.2%	58.5%	**75.8%**	75.4%	53.7%
	Flickr30k (balanced)	24.6%	25.9%	9.0%	25.2%	**26.9%**	4.0%

Table 2: Top: Mean rank of the correct preposition (lower is better). Bottom: Accuracy with different feature configurations. All results are with the original trajector/landmark terms from descriptions. IND stands for Indicator Vectors, W2V for Word2Vec, and GF for Geometric Features. As baseline we rank the prepositions by their relative frequencies in the training dataset.

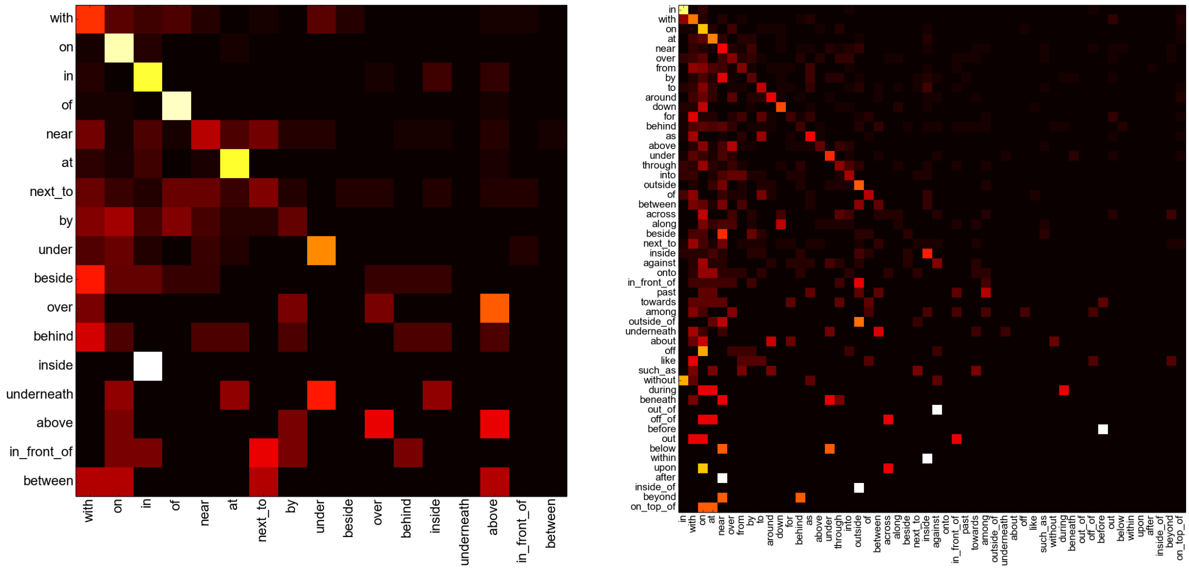

Figure 2: Normalised confusion matrices on the balanced test subsets for the two datasets (left: MSCOCO, right: Flickr30k), using geometric features and word2vec with the original terms.

multiple possible prepositions that may be equally valid. For completeness we also report classification accuracy results.

Baseline. As baseline, we rank the prepositions by their relative frequencies in the training dataset. We found this to be a sufficiently strong baseline, as ubiquitous prepositions such as *with* and *in* tend to occur frequently in the dataset.

6.1 Ranking with known entity labels

In this section, we focus on predicting the best preposition given the geometric and textual features of the trajector and landmark entities. This simulates the scenario of a vision detector providing a firm decision on the concept label for the detected entities. We use a multi-class logistic regression classifier (Fan et al., 2008), and concatenate multiple features into a single vector. We compare high-level categories and terms from descriptions as trajector/landmark labels. Prepositions are ranked in descending order of the classifier output scores.

We found a few prepositions (e.g. *with*) dominating the datasets. Thus, we also evaluated our models on a balanced subset where each preposition is limited to a maximum of 50 random test samples. The training samples are weighted according to their class frequency in order to train non-biased classifiers to predict this balanced test set. The results on both the original and balanced

Dataset	Prep (known labels)		Preposition		Trajector		Landmark	
	acc	rank	acc	rank	acc	rank	acc	rank
MSCOCO	79.8%	1.46 (17)	62.9%	1.92 (17)	65.6%	4.64 (74)	44.5%	7.30 (77)
Flickr30k	67.1%	2.16 (52)	61.7%	2.28 (52)	77.3%	1.43 (8)	66.4%	1.64 (8)

Table 3: Accuracy (**acc**) and mean rank (**rank**, with max rank in parenthesis) for each variable of the CRF model, trained using the high-level concept labels. Columns under **Prep (known labels)** refer to the results of predicting prepositions with the trajector and landmark labels fixed to the correct values.

test sets are compared.

As shown in Table 2, the system performed significantly better than the baseline in most cases. In general, geometric features perform better than the baseline, and when combined with text features further improve the results. In a per-preposition analysis, the geometric features show up to 14% improvement in the mean rank for Flickr30k.

In feature ablation tests on MSCOCO (balanced), we found the y component of the trajector to landmark vector to be important to most prepositions, especially for *under*, *above* and *on*. Other important geometric features include the final two features in Table 1 (Euclidean distance and area).

The benefit of the word2vec text feature is clear when moving from high-level categories to original terms from descriptions, where it consistently improves the mean rank (up to 25%). In contrast, the indicator vectors resulted in a less significant improvement, if not worse performance, when using the sparse original terms.

We also evaluated the relative importance of the trajector and the landmark, by withholding either from the textual feature vector. We found that the landmark plays a larger role in preposition prediction as omitting the trajector produces 10%-30% better results than omitting the landmark.

Figure 2 shows the confusion matrices of the best-performing systems. Note that many mistakes arise from prepositions that are often equally valid (e.g. predicting *near* instead of *next_to*).

6.2 Ranking with unknown entity labels

Here, we investigate the task of jointly predicting prepositions with the entity labels given geometric and visual features (*without* the trajector and landmark labels). This simulates the scenario of a vision detector output. For this structured prediction task, we use a 3-node chain CRF model[2], with the centre node representing the preposition and the two end nodes representing the trajector and landmark. We use *image features* for the entity nodes, and *geometric features* for the preposition node (Section 5). Due to computational constraints only high-level category labels are used, but as seen in Section 6.1, this may actually be *hurting* the performance.

Table 3 shows the results of the structured model used to predict the most likely *(trajector, preposition, landmark)* combination. To facilitate comparison with Section 6.1, column **Prep (known labels)** shows the results with the trajector and landmark labels as known conditions and fixed to the correct values, thus only needing to predict the preposition. The model achieved excellent performance considering the added difficulty of the task.

7 Conclusions and Future Work

We explored the role of geometric, textual and visual features in learning to predict a preposition given two bounding box instances in an image, and found clear evidence that all three features play a part in the task. Our system performs well even with uncertainties surrounding the entity labels. Future work could include non-prepositional terms like verbs, having prepositions modify verbs, adding word2vec embeddings to the structured prediction model, and providing stronger features – whether textual, visual or geometric.

Acknowledgments

This work was funded by the ERA-Net CHIST-ERA D2K VisualSense project (Spanish MINECO PCIN-2013-047, UK EPSRC EP/K019082/1 and French ANR Grant ANR-12-CHRI-0002-04) and the Spanish MINECO RobInstruct project TIN2014-58178-R. Ying Lu was also supported by the China Scholarship Council.

[2] We used the toolbox by Mark Schmidt: `http://www.cs.ubc.ca/~schmidtm/Software/UGM.html`

References

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, December.

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA, October. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. 2013. A sentence is worth a thousand pixels. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Abhinav Gupta and Larry S. Davis. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of the European Conference on Computer Vision*, pages 16–29.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):article 4, 36 p.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. In **SEM 2012: 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 (SemEval 2012)*, pages 365–373, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. 2014. TUHOI: Trento Universal Human Object Interaction dataset. In *Proceedings of the Third Workshop on Vision and Language*, pages 17–24, Dublin, Ireland, August. Dublin City University and the Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Kenneth C. Litkowski and Orin Hargraves. 2005. The preposition project. In *ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications*, pages 171–179.

Kenneth C. Litkowski and Orin Hargraves. 2007. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic, June. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. 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*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June. Association for Computational Linguistics.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 444–454. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, February.