Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes

Simon Dobnik[1] and **John D. Kelleher**[2*]
[1]University of Gothenburg, Centre for Language Technology,
Dept. of Philosophy, Linguistics & Theory of Science, Gothenburg, Sweden
[2]Dublin Institute of Technology, Applied Intelligence Research Centre,
School of Computing, Dublin, Ireland
simon.dobnik@gu.se, john.d.kelleher@dit.ie

Abstract

We present a method of extracting functional semantic knowledge from corpora of descriptions of visual scenes. Such knowledge is required for interpretation and generation of spatial descriptions in tasks such as visual search. We identify semantic classes of target and landmark objects related by each preposition by abstracting over WordNet taxonomy. The inclusion of such knowledge in visual search should equip robots with a better, more human-like spatial cognition.

1 Introduction

Visual search is an area of growing research importance in mobile robotics; see (Sjöö, 2011; Kunze et al., 2014) among others. Visual search involves directing the visual sensors of a robot with the goal of locating a specific object. Several recent approaches have integrated non-visual (often linguistically motivated information) into the visual search process. The intuition behind this is that if the robot knows that object X is often located *near*/*on*/... object Y then in situations where Y is visually salient it may be easier for the system to search by first locating Y and then use relational information to direct the search for X. A key component of these approaches to visual search is the definition of spatial semantics of the relational information. Appropriately modelling these semantics is crucial because fundamentally it is these models that define the scope of the visual search in relation to Y.

In language spatial relations between objects are often expressed using *locative expressions* such as "the apple above a bowl", "the computer is on the shelf" and "the plane is over the house". In these expressions a *target* object is located relative to a *landmark* object using a *preposition* to describe the spatial relation. Crucially, there are differences between prepositions with respect to how their spatial relations are defined. The semantics of some prepositions can be modelled in terms of geometric primitives whereas the semantics of other prepositions are sensitive to the functional relations between the target and the landmark (Coventry and Garrod, 2004). Consider the example "Alex is at her desk". This description refers to a situation where Alex is not only geometrically proximate to her desk but also where she is sitting down and working. The extra constraints are coming from the functional relations that normally exist between an individual and a desk.

Returning to visual search, being able to identify whether a given preposition is primarily geometric of functional is important because this classification informs the design of the spatial semantics for the preposition and hence the appropriate definition of the search relative to the landmark object. In this paper we present some ongoing experiments which attempt to develop techniques that can classify prepositions as geometric or functional.

2 Spatial descriptions

(Coventry and Garrod, 2004) show in the experiments with human observers of images of a man in the rain holding an umbrella where the umbrella is providing a varying degree of protection from the rain that "above" is more sensitive to the geometrical component than "over" and that "over" is more

[*] Both authors are equal contributors.

Proceedings of the 25th International Conference on Computational Linguistics, pages 33–37,
Dublin, Ireland, August 23-29 2014.

sensitive to the object function component than "above". Descriptions of "the umbrella is over a man" were considered acceptable even in cases where the umbrella was held horizontally but was providing protection from the rain.

Modelling of functional knowledge in the computational models of meaning of spatial prepositions is not straightforward. Humans (and even expert linguists) do not seem to have clear intuitions what the functional component of each preposition sense may be and hence they must be confirmed experimentally (Garrod et al., 1999). This may take significant time for various combinations of prepositions and target and landmark objects. One needs to develop a complex ontology of object properties and then associate spatial prepositions with rules that pick out certain properties of objects for a particular preposition sense (for examples of such rules see (Garrod et al., 1999, p.170)).

In this paper we describe a method of extraction of these meaning components from a corpus of descriptions of visual scenes automatically. Unlike in the psycho-linguistic experiments described above we examine general corpora that obtain a wide and unrestricted set of images that humans described freely. The purpose of the experiment is to investigate whether functional knowledge can be extracted from contextual language use. For example, can we make generalisations about the semantics of the arguments that a particular prepositional sense takes automatically. Furthermore, we are also interested if the distinctions between geometric and functional sensitivity of prepositions reported experimentally could be determined this way. This information would allow us to weight the contributions of the geometric and functional knowledge when generating and interpreting spatial descriptions of scenes. We hypothesise, that if a preposition is sensitive to functional meaning then there will be functional relations between target and landmark objects that it is used with and consequently the preposition will will be much more restrictive or specific in the semantic type of targets and landmarks that it requires. Other prepositions may be more sensitive to the satisfaction of the geometric constraint and hence we expect that they will co-occur with objects of more general semantic types.

3 Datasets and extraction of spatial descriptions

The goal of this work is to analyse the semantics of linguistic expressions that are used to describe the relative location of objects in visual contexts. We base our analysis on two corpora of image descriptions: specifically, the IAPR TC-12 Benchmark corpus (Grubinger et al., 2006)[1] which contains 20,000 images and multi-sentence descriptions and the 8K ImageFlickr dataset (Rashtchian et al., 2010)[2] which contains 8108 images. In both corpora the situations and events represented by images are described by several sentences which contain spatial descriptions with prepositions: in the first case all sentences are by a single annotator and in the second case each sentence is by a different annotator. The descriptions are geometrically constrained by the visual context. On the other hand, the describers' choice of the target and the landmarks objects and the preposition in these descriptions will tell us about their functional semantics. The main pre-processing step was to extract parts of spatial expressions used in the image descriptions. Once extracted each spatial expression was stored in a type with the following structure: ⟨preposition, target, landmark⟩. To do this extraction we first parsed both corpora of linguistic descriptions for dependencies (Marneffe et al., 2006) using Stanford CoreNLP tools[3]. Then we wrote several extraction rules that matched dependency parses and extracted all three parts of spatial expressions that we are looking for. All words were lemmatized and converted to lower case, compound prepositions such as "on the left side of" were rewritten as single words and names, etc. were converted to their named entity categories such as "person". The extracted patterns from both corpora were combined into a single dataset from which we can determine their frequency counts.

4 Determining conceptual categories of objects

The intuition behind our experiment is that functionally defined prepositions can be distinguished from geometrically defined prepositions by virtue of the fact that the functional relations between the target

[1] http://imageclef.org/photodata

[2] http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html

[3] http://nlp.stanford.edu/software/corenlp.shtml

and landmark objects that are encoded in the semantics of functional prepositions result in less variance across the object types that occur with geometric prepositions. In other words, the target objects that occur with a functional preposition will be more semantically similar to each other than the target objects that occur with a geometric preposition. Likewise the landmark objects that occur with a functional preposition will be more semantically similar than the landmark objects that occur with a geometric preposition. In order to test this intuition we need to be able to cluster the target and landmark objects that occur with a given preposition into conceptual categories. We can then define patterns that describe the pairs of conceptual categories that a preposition occurs with. Based on the intuition that functional prepositions occur with more specific object classes we would expect that functional prepositions generate more patterns of use and that the patterns include more specific classes of objects.

To determine conceptual categories that objects of prepositions belong to, WordNet (Fellbaum, 1998) appears to be an ideal tool. It contains taxonomies of words that were constructed by humans using their intuitions. While certain classification of words in the ontology are not entirely unproblematic it is nonetheless considered a gold-standard resource for lexical semantics. In particular, we are interested in finding out given a certain preposition what are the possible semantic classes of its target and landmark objects. To determine the class synset (a sense in WordNet terminology) that covers a bag of words best we use the *class-labelling algorithm* of Widdows (2003). Given a list of words, this algorithm finds hypernyms which subsume as many as possible words in the list, as closely as possible. The algorithm works by first defining the set of all possible hypernyms for the words in the list. In then computes a score for each of the hypernyms: a hypernym score is incremented by a small positive value for each word it subsumes (this positive value is defined as 1 divided by the square of the vertical distance in the hierarchy between the hypernym and the word) and decremented by a small negative value g for each word it does not subsume. The algorithm returns the hypernym with the highest score. Widdows (2003) sets g to 0.25 for lists that contain 5 words on average. The lists in our experiments are much longer so we scale g to the length of the list input: $g = 0.25 \times \frac{5}{list\ length}$. Given a bag of nouns we use the class-labelling algorithm to determine the bets matching hypernym. Words that are subsumed by this hypernym are labelled as belonging to this class and are removed from the bag of words. The algorithm is repeated on the remaining words recursively until all words from the bag of words are exhausted. The procedure allows us to greedily create classes of words that are most general categories representing these words, the level of generality can be tweaked by the parameter g. The bag of words is allowed to contain duplicates as these are indicators of coherent classes. Duplicate words are all covered by a common hypernym and hence this hypernym will be given more weight in the overall scoring. The hypernyms introduced by infrequent and non-similar words are given less weight. This filters words that may be included due to an error.

5 Patterns of prepositional use

The algorithm for class-labelling of words backed by the WordNet ontology allows us to predict the typical classes or semantic types of the landmark and the target objects related by a preposition. From these several patterns can be generated. Such patterns can be used both in the interpretation of visual scenes (visual search) or generation of spatial referring expressions that optimally constrain the set of intended objects. We create the patterns by collecting all targets and all landmarks that occur with a particular preposition. We apply the class labelling on the bags of words representing the targets and the landmarks to obtain a set of classes representing targets and landmarks. Finally, for every tuple $\langle$target, preposition, landmark$\rangle$ we replace target and landmark with target class and landmark class and collect a set of $\langle$target class, preposition, landmark class$\rangle$ patterns. This method assumes that targets and landmarks are semantically independent of each other. Below are some examples of patterns that our algorithm has found (the notation for the names of the objects corresponds to the names of synsets in the WordNet taxonomy, the numbers in brackets indicate the number of examples out of total examples covered by this pattern): (i) travel.v.01 over object.n.01 (9/713), sunlight.n.01 over object.n.01 (13/713), bridge.n.01 over object.n.01 (23/713), bridge.n.01 over body_of_water.n.01 (42/713), air.v.03 over object.n.01 (36/713), artifact.n.01 over body_of_water.n.01 (42/713), artifact.n.01 over object.n.01

(175/713),... (ii) breeze.n.01 above body_of_water.n.01 (8/183), person.n.01 above artifact.n.01 (9/183), artifact.n.01 above steer.v.01 (14/183), artifact.n.01 above entrance.n.01 (16/183) artifact.n.01 above artifact.n.01 (27/183),... (iii) person.n.01 under tree.n.01 (7/213), shirt.n.01 under sweater.n.01 (8/213), person.n.01 under body_of_water.n.01 (11/213), person.n.01 under artifact.n.01 (13/213) artifact.n.01 under artifact.n.01 (16/213), person.n.01 under structure.n.01 (17/213), artifact.n.01 under structure.n.01 (21/213),... (iv) box.n.05 below window.n.08 (1/14), crown.n.04 below script.n.01 (2/14),...[4] The patterns show that different prepositions which are seemingly synonyms when considering geometry ("over"/"above" and "below"/"under") do relate different types of objects and from the labels of the semantic classes we may speculate what kind of semantic knowledge the relations are sensitive to. Importantly, the labels of the classes and different patterns extracted show that there may be several distinct and even unrelated situations that a preposition is referring to. Consider for example, person.n.01 under tree.n.01, shirt.n.01 under sweater.n.01 (8/213), person.n.01 under body of water.n.01 are denoting three different kinds of situations which require distinct and unrelated geometric arrangements. Overall, the results indicate that the method is able to extract functional knowledge which is a reflection of the way humans conceptualise objects and relations between them and which may be useful for improving visual processing of scenes.

Another question we set off to answer is whether from the patterns one can determine the functional and geometric bias of a preposition. As noted previously, our application of the class labelling algorithm is greedy and attempts to cover a large number of words. The more words are generalised over the more generic classes are created. To counter this confounding factor we down-sampled the dataset by creating, for the results we report here, 50 samples of 20 randomly chosen words. The same procedure for creating patterns of prepositional use was applied as before. On each sampling iteration we estimate for each preposition **(i) the average depth of the target and landmark hypernyms** in the WordNet taxonomy, **(ii) the number of patterns created**, and **(iii) the entropy of the patterns** over the examples in the dataset. Finally, we average all values obtained from iterations and we rank the prepositions by the ascending values of the these measures as shown below:

(i) on (3.17), near (3.55), with (3.66), next to (3.83), of (3.95), between (4.17), in front of (4.26), above (4.27), over (4.48), around (4.52), behind (4.65), from (4.74), at (4.89), under (4.93), for (4.97), through (5.27), in (5.45)

(ii) on (10.5), with (11.5), near (12), next to (12.1), between (12.6), of (12.6), above (12.7), around (13), in front of (13.1), over (13.7), from (13.8), behind (13.9), for (14.2), under (14.3), in (14.5), through (15.1), at (15.2)

(iii) on (2.74), next to (3.05), with (3.07), near (3.1), between (3.2), of (3.29), above (3.33), around (3.36), in front of (3.39), over (3.48), from (3.5), behind (3.51), for (3.59), under (3.62), in (3.62), through (3.75), at (3.76)

With small variations all measures (i), (ii) and (iii) rank the prepositions very similarly. Items at the beginning of the list are used with target and landmark objects that belong to more general classes (i), they are covered by a lower number of preposition patterns (ii), and the entropy of these patterns is low (iii)[5]. Hence, we expect that items at the top of the lists are less affected by the type of the target and landmark objects than items at the bottom of the lists. They are the prepositions where the geometric component is stronger to determine the spatial relation. On the other hand, prepositions at the bottom of the list rely more on the functional component of meaning. The results predict the observations from the literature. Although, being sometimes quite close "above" precedes "over" in respect to to all three measures. "Below" was not processed in this study as there were too few examples but "under" is found at the tail of the list. The ranking of other prepositions also aligns with our intuitions. For example, "on", appearing as the head of the list, requires a contact between objects (a geometric notion), whereas "in", appearing in the tail of the list requires that the target is constrained by the landmark (a functional notion).

[4]There are only 14 examples of "below" in the dataset for which nearly always unique patterns were created.

[5]Entropy balances between the number of classes and the frequency of items in them. Low entropy indicates that there is a tendency of items clustering in small number of specific classes rather rather than being equally dispersed over classes.

6 Conclusions and further work

In the preceding discussion we have demonstrated a method of extracting (i) functional semantic knowledge – required for the generation and interpretation of spatial descriptions – from the corpora of descriptions referring to visual scenes and (ii) made predictions about the bias of spatial descriptions to functional knowledge. We hope that this information can facilitate visual search as it further restricts the set of possible situations and objects involved. We have constructed several patterns of preposition use and have shown that a preposition such as "under" may refer to several distinct situations constrained by the knowledge of object function and that these situations would require geometric representations that are likely to be quite different. This knowledge should allow us to create different routines for visual search that could be applied over a wide set of domains and would better approximate the way humans perceive and reason about space. The applicability of this information for visual search must be properly evaluated in a visual application. Estimating the functional or geometric bias of prepositions informs us for their modelling but more importantly confirms that the patterns extracted here follow the experimental observations reported in the literature.

In the procedure we have made several design choices: the choice of the corpora and the way the corpora is processed and information extracted, the algorithm with which words are labelled for semantic classes and finally the method with which patterns are created. For example, before class labelling we could use an algorithm that clusters words of similar hypernym depth into discrete classes over which hypernyms are generalised. This would allow us to distinguish better between different situations that a preposition is referring to. When creating patterns we assume that target and landmark objects are independent of each other. However, this may not necessarily be the case. For example, the category of the target object may constrain the category of the landmark which means that the latter category should only be generalised over landmark words that occur with some target category.

References

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Mass.

Simon Garrod, Gillian Ferrier, and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *Proceedings of OntoImage 2006: Workshop on language resources for content-based mage retrieval during LREC 2006*, Genoa, Italy, 22 May. European Language Resources Association.

Lars Kunze, Chris Burbridge, and Nick Hawes. 2014. Bootstrapping probabilistic models of qualitative spatial relations for active visual object search. In *AAAI Spring Symposium 2014 on Qualitative Representations for Robots*, Stanford University in Palo Alto, California, US, March, 24–26.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of Int'l Conf. on Language Resources and Evaluation (LREC)*, pages 449–454, Genoa, Italy. European Language Resources Association.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA, 6 June. North American Chapter of the Association for Computational Linguistics (NAACL).

Kristoffer Sjöö. 2011. *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*. Ph.D. thesis, KTH, Computer Vision and Active Perception (CVAP), Centre for Autonomous Systems (CAS), Stockholm, Sweden.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 197–204. Association for Computational Linguistics.