Commonsense Reasoning Based on Betweenness and Direction in Distributional Models*

Steven Schockaert
School of Computer Science and Informatics
Cardiff University, Cardiff, UK
SchockaertS1@cardiff.ac.uk

Joaquín Derrac
School of Computer Science and Informatics
Cardiff University, Cardiff, UK
DerracRusJ@cardiff.ac.uk

Abstract

Several recent approaches use distributional similarity for making symbolic reasoning more flexible. While an important step in the right direction, the use of similarity has a number of inherent limitations. We argue that similarity-based reasoning should be complemented with commonsense reasoning patterns such as interpolation and a fortiori inference. We show how the required background knowledge for these inference patterns can be obtained from distributional models.

Introduction

Structured knowledge bases such as Freebase, YAGO, CYC and ConceptNet are becoming increasingly important in applications (e.g. semantic search). As a result, several authors have recently looked at techniques for automatically extending such knowledge bases. One possibility is to use external knowledge (West et al. 2014), and for example rely on information extraction techniques to fill in missing values for prominent attributes (e.g. missing birth dates). Other approaches rely on exploiting regularities within the knowledge base, e.g. learning probabilistic dependencies (Lao, Mitchell, and Cohen) or using matrix factorisation (Speer, Havasi, and Lieberman 2008). A third class of approaches relies on commonsense reasoning, and in particular on similarity-based reasoning (Beltagy et al. 2013; Freitas et al. 2014; d'Amato et al. 2010). These approaches are based on the assumption that similar concepts tend to have similar properties, where similarity degrees are often obtained from distributional models. For example, knowing that Cabernet Sauvignon pairs well with a grilled steak, we can plausibly derive that this wine will also pair well with a barbecued steak, given the similarity between both types of steaks.

Similarity-based reasoning has two main limitations. First, it can only be used when there are sufficiently similar concepts for which the required type of knowledge is available. Second, similarity degrees are highly context-dependent (e.g. red and white Burgundy wine are similar in some sense, but they should be paired with very different types of food). To alleviate these issues, we propose to complement similarity-based reasoning with additional commonsense reasoning patterns:

Interpolative reasoning Suppose we know that undergraduate students and PhD students are both exempt from paying council tax in the UK, then we can plausibly conclude that master's students are also exempt from paying this tax, given that master's students are conceptually between undergraduate students and PhD students.

A fortiori reasoning Suppose we know that buying beer is illegal under the age of 18 in the UK, then we can plausibly derive that buying whiskey is also illegal under the age of 18, since whiskey is stronger than beer.

These reasoning patterns require background knowledge: conceptual betweenness for interpolation and relative properties for a fortiori reasoning. The aim of this paper is to show how this type of background knowledge can be obtained from distributional models. In the next section, we explain how we use multi-dimensional scaling (MDS) to induce from a text corpus a Euclidean space (called a semantic space or conceptual spaces) in which entities of a given type (e.g. wines) are represented as points, categories correspond to (usually convex) regions and relative properties correspond to directions. Subsequently, we explain in more detail how conceptual betweenness and relative properties correspond to spatial relations in a semantic space.

Inducing a semantic space

The examples in this paper are mostly based on a semantic space of place types, containing place types from three taxonomies: GeoNames[1] (667 place types organised in 9 categories), Foursquare[2] (435 place types organised 9 top-level categories) and OpenCYC[3] (3388 place types, being the refinements of the category *site*). We have used the Flickr API to collect meta-data of photos that have been tagged with each of these place types. Our assumption is that photos which are tagged with a given place type (e.g. *restaurant*) will often contain other tags that relate to that place type (e.g.

*This work has been supported by EPSRC (EP/K021788/1).

[1] http://www.geonames.org/export/codes.html, accessed September 2013.

[2] http://aboutfoursquare.com/foursquare-categories/, accessed September 2013.

[3] http://www.cyc.com/platform/opencyc, accessed April 2014.

food, *waiter*, *dessert*), and will thus enable us to model the meaning of that place type. In total we collected 22 816 139 photos in April 2014. Place types with fewer than 1000 associated photos on Flickr have been omitted from our analysis, resulting in a total of 1383 remaining place types.

Following (Turney and Pantel 2010), we use Positive Pointwise Mutual Information (PPMI) to represent the bag of words corresponding to each place type as a real-valued vector. Subsequently, following (Gärdenfors and Williams 2001), we apply MDS to represent place types as points in a lower-dimensional space; unless otherwise stated, we will assume a 100-dimensional space throughout this paper. As input, MDS requires a dissimilarity matrix. To measure the dissimilarity between place types, we used the angular difference between the corresponding PPMI weighted vectors.

Several authors have already proposed the use of dimensionality reduction methods for commonsense reasoning. For example, (Speer, Havasi, and Lieberman 2008) uses Singular Value Decomposition (SVD) to find missing properties in ConceptNet. However, SVD produces a representation in which entities correspond to vectors, which should be compared in terms of cosine similarity rather than Euclidean distance. While this makes no difference in applications which only rely on similarity, spatial relations such as betweenness are only meaningful if Euclidean distance is meaningful. This has been confirmed by experiments in (Derrac and Schockaert 2014b), where MDS was found to substantially outperform SVD for obtaining betweenness relations.

We also consider a semantic space of wine varieties, which was induced from a corpus of wine reviews[4]. The entities we will consider in this paper are the wine variants (e.g. Syrah), rather than the specific wines (e.g. 2001 Thierry Allemand Cornas Reynard). In particular, we consider the 330 wine varieties for which the available reviews together contained at least 1000 words. In the bag-of-words representation of a given wine variety, we consider all terms occurring in the corresponding reviews, after removing stop words and diacritics and converting everything to lowercase. As for the place types, we then use MDS to obtain a lower-dimensional representation.

Betweenness and interpolative reasoning

We say that an entity b is conceptually between entities a and c if b has all the *natural* properties that a and c have in common. Geometrically, we say that the point b is between points a and c if $\cos(\overrightarrow{ab}, \overrightarrow{bc}) = 1$ (assuming that a, b and c are disjoint), where we identify an entity a with the point representing it in the semantic space for the ease of presentation. Conceptual and geometric betweenness can be linked to each other, by taking into consideration that natural properties tend to correspond to convex regions in a suitable semantic space (Gärdenfors and Williams 2001). Indeed, b is geometrically between a and c iff all convex regions that contain a and c also contain b. This suggests that we identify conceptual betweenness by checking for geometric betweenness in a semantic space.

Table 1: Examples of place types b which were classified correctly using the betweenness classifier, because b was found to be between place types a and c.

Place b	Places (a, c)	Category found
marina	(harbor, plaza)	Parks & Outdoor
music school	(auditorium, elementary school)	Professional & Other
campground	(playground, scenic lookout)	Parks & Outdoor
bike shop	(bookstore, motorcycle shop)	Shops & Services
medical center	(fire station, hospital)	Professional & Other
legal services	(dojo, financial services)	Shops & Services
candy store	(grocery store, toy store)	Shops & Services
art gallery	(comedy club, museum)	Arts & Entertainment
skate park	(playground, plaza)	Parks & Outdoor
veterinarian	(animal shelter, emergency room)	Professional & Other

Table 2: Examples of place types b which were misclassified by 1-NN because they are most similar to a place type a from a different category.

Place b	Place a	Category found
marina	pier	Travel & Transport
music school	jazz club	Arts & Entertainment
campground	hostel	Travel & Transport
bike shop	bike rental	Travel & Transport
medical center	medical school	College & University
legal services	tech startup	Professional & Other
candy store	ice cream shop	Food
art gallery	sculpture garden	Parks & Outdoor
skate park	board shop	Shops & Services
veterinarian	photography lab	Shops & Services

In practice we will need a measure for the degree of betweenness of any three points (a, b, c). The following measure was found to give good results in (Derrac and Schockaert 2014b):

$$\mathit{Btw}(a,b,c) = \begin{cases} \|\overrightarrow{bp}\| & \text{if } \cos(\overrightarrow{ac}, \overrightarrow{ab}) \geq 0 \\ & \quad \text{and } \cos(\overrightarrow{ca}, \overrightarrow{cb}) \geq 0 \\ +\infty & \text{otherwise} \end{cases}$$

where p is the orthogonal projection of b on the line connecting a and c. Note that the condition $\cos(\overrightarrow{ac}, \overrightarrow{ab}) \geq 0$ and $\cos(\overrightarrow{ca}, \overrightarrow{cb}) \geq 0$ is satisfied iff p lies on the line segment between a and c. Also note that higher values for $\mathit{Btw}(a,b,c)$ correspond to weaker betweenness relations, and in particular that a score of 0 denotes perfect betweenness.

The resulting betweenness relations can then be used for automating interpolative inference. In this paper, we will focus on the use of interpolation in a standard classification setting, as an alternative to k-NN. In particular, assume that we need to classify an entity b to one of the categories $C_1, ..., C_n$ and that for each category C_i we have a set Y_i of training items available. Using interpolation, we assign b to the category C_i minimising $\min_{a,c \in Y_i} \mathit{Btw}(a,b,c)$. In practice, we could also consider the k strongest betweenness triples (a, b, c) and use a voting procedure, as for k-NN.

We compared the performance of this betweenness based classifier against 1-NN on three types of classification problems. In particular, we consider a binary classification problem for each of the categories of place types used by

[4] https://snap.stanford.edu/data/web-CellarTracker.html

Table 3: Results obtained for place types.

	Foursquare		GeoNames		OpenCYC	
	Acc.	F1	Acc.	F1	Acc.	F1
Btw	0.950	0.730	0.872	0.345	0.955	0.400
a fortiori	0.938	0.724	0.852	0.328	0.945	0.404
1-NN	0.934	0.715	0.852	0.323	0.945	0.372
SVM	0.936	0.676	0.876	0.362	0.930	0.355

Table 4: Results obtained for wine varieties.

	20		50		100	
	Acc.	F1	Acc.	F1	Acc.	F1
Btw	0.884	0.527	0.888	0.553	0.882	0.543
a fortiori	0.885	0.564	0.883	0.570	0.874	0.554
1-NN	0.880	0.559	0.875	0.550	0.869	0.546
SVM	0.839	0.492	0.862	0.516	0.867	0.564

GeoNames, for each of the top-level categories used by Foursquare, and for the 93 largest sub-categories of *site* in OpenCYC. We did consider k-NN classifiers with different values for k, but found $k = 1$ to be a suitable choice here. The results are shown in Table 3 (using 5-fold cross-validation). The results for SVM have been obtained using the LIBSVM implementation, considering a Gaussian kernel. The optimum value for the C parameter was set for each category by using grid search (holding out one third of the training data for testing the different values of C). The *a fortiori* method will be explained in the next section.

We can observe that the betweenness based classifier consistently outperforms 1-NN, although the differences are relatively small. In many cases, both classifiers will make the same decisions, which is unsurprising given that when a and b are highly similar, the score *Btw*(a, b, c) will be very low. As a result, when a sufficiently similar training item exists, betweenness based classification often degenerates to a similarity based classification. However, in contrast to k-NN, the betweenness based classifier can make reliable classification decisions even when no highly similar training items exist. Table 1 contains examples of place types which were correctly classified by the betweenness classifier but incorrectly by 1-NN (for the Foursquare categories); Table 2 shows which training item misled the 1-NN classifier. Essentially, there are two types of mistakes in Table 2. As an example of the first type, since there is no place type in the training data which is highly similar to *veterinarian*, 1-NN relied on the most similar item, which was *photography lab*. In contrast, the betweenness classifier successfully recognised that *veterinarian* is conceptually between *animal shelter* and *emergency room*, from which the correct category was obtained. A second type of mistake the 1-NN classifier makes is illustrated by *music school*. While *music school* and *jazz club* are indeed similar (both being music related), they are not similar in the relevant aspects (as they serve a rather different function). This is reflected in the fact that *music school* is not between any two place types from the *Arts & Entertainment* category.

The betweenness based classifier also outperforms SVM on the Foursquare and OpenCYC classes, but not for GeoNames. This seems related to the observation that some of the categories used by GeoNames are more artificial (e.g. feature code L encompasses *naval base*, *housing development*, *park* and *wildlife reserve*) and therefore less likely to correspond to convex regions in the semantic space, while the betweenness classifier directly relies on the assumption that categories are convex. For example, for feature type L, SVM achieves an F1 score of 0.308 compared to 0.207 for 1-NN and 0.167 for *Btw*. In contrast, for feature type V, which is a much more well-defined feature type covering *forest*, *heath* and closely related types, SVM achieves an F1 score of 0.180 compared to 0.324 for 1-NN and 0.455 for *Btw*. This also illustrates that the relatively small difference between the results for 1-NN and *Btw* hides considerable variation for individual categories (with *Btw* often performing substantially better for natural categories).

Direction and a fortiori reasoning

Our hypothesis is that for two entities a and b, the direction of the vector $\overrightarrow{ab}$ encodes how b differs from a. To test this hypothesis, we evaluate the performance of a classifier which is based on a form of a fortiori reasoning. We consider a binary classification problem, with P the set of positive training items and N the set of negative training items. If $x \in N$ and $y \in P$ then we assume that the direction of $\overrightarrow{xy}$ in the semantic space is towards category membership, and conversely that the direction of $\overrightarrow{yx}$ is towards non-membership. This leads to the following measures of positive and negative support for a test item b:

$$\mathit{pos}(b) = \max_{x \in N} \max_{y \in P} \max_{a \in P} \cos(\overrightarrow{xy}, \overrightarrow{ab})$$

$$\mathit{neg}(b) = \max_{x \in P} \max_{y \in N} \max_{a \in N} \cos(\overrightarrow{xy}, \overrightarrow{ab})$$

Intuitively, if $\overrightarrow{xy}$ points towards category membership and a is already a member of the category, then *a fortiori* we would also expect b to belong to the category if $\overrightarrow{xy}$ is approximately parallel to $\overrightarrow{ab}$, and conversely for *neg*. We assume that b is a positive instance iff *pos*$(b) >$ *neg*(b). While a naive implementation of this classifier would have a cubic time complexity, when using a KD tree, the time it takes to classify an item will only be quadratic in the number of training items.

The results of this classifier for the place type experiments are shown in Table 3. As for the betweenness classifier, we find a consistent but small improvement over 1-NN. In an additional experiment, we looked at the semantic space of wines. To obtain binary classification problems, we considered the 14 largest categories from a wine taxonomy[5]. The results, for semantic spaces of dimension 20, 50 and 100, are shown in Table 4. Here, the best results are obtained by the a fortiori classifier in 50D. The a fortiori classifier outperforms 1-NN across all dimensions, while the betweenness classifier only outperforms 1-NN for 50D.

In (Derrac and Schockaert 2014a) we proposed a different type of classifier, also based on the idea of a fortiori reasoning and the hypothesis that directions in the semantic

[5] http://winefolly.com/review/different-types-of-wine/, accessed April 2014.

space encode relative properties. In particular, we proposed a method to identify the most salient properties of the domain under consideration, and to identify which directions correspond to these properties. Initially each term occurring in the text collection is assumed to potentially correspond to an important property. For each term, we train a linear SVM classifier, separating those entities in the semantic space that contain the term in at least one of their associated texts from the other entities. We then use the Kappa score for selecting those terms for which this classifier is sufficiently accurate. The most salient properties are essentially taken as those for which the Kappa score is maximal, and the corresponding direction is defined by the perpendicular to the hyperplane that was found by the SVM classifier. For example, in this way the method discovered (in an unsupervised way) that being scary and being romantic are salient properties in a semantic space of movies. The corresponding directions are displayed in Figure 1, which shows the projection of a 100-dimensional semantic space of movies on the 2-dimensional plane spanned by the two directions. Each of the salient directions induces a ranking of the entities under consideration. The classifier we proposed in (Derrac and Schockaert 2014a) uses a modification of the well-known FOIL algorithm to learn classification rules from the rankings corresponding to the 200 most salient properties. These rules are of the form "if x is a Thriller and y is more violent than x, then y is also a Thriller". While the FOIL based classifier achieved very good results in the movies domain, we found it to be uncompetitive for the place types and wines domains. The main reason seems to be that the number of underlying entities is much smaller in these latter cases: while the movies space was induced from data about 15000 movies, we only considered 330 wine varieties and 1383 place types, which may not be enough to reliable identify directions of salient properties in a high-dimensional space. In contrast, the betweenness and a fortiori based classifiers do not scale to training sets of more than a few thousand items.

Conclusions

We have proposed two inference methods that rely on relations derived from a semantic space: interpolation, which uses geometric betweenness, and a fortiori inference, which uses direction relations. Like similarity-based reasoning, these methods are based on commonsense reasoning, and could thus readily provide intuitive explanations of why a given conclusion is considered plausible. We envisage that a combination of different forms of commonsense reasoning will be needed in practice. For example, interpolation is only meaningful for natural categories, since it relies on the assumption that categories correspond to convex regions;k-NN classifiers are likely to be better suited for more artificial categories. When sufficient training data is available, SVM classifiers or the FOIL based methods from (Derrac and Schockaert 2014a) are likely to be more suitable. Note, however, that the former does not easily allow us to provide intuitive explanations, which would be a key drawback in applications such as question answering.

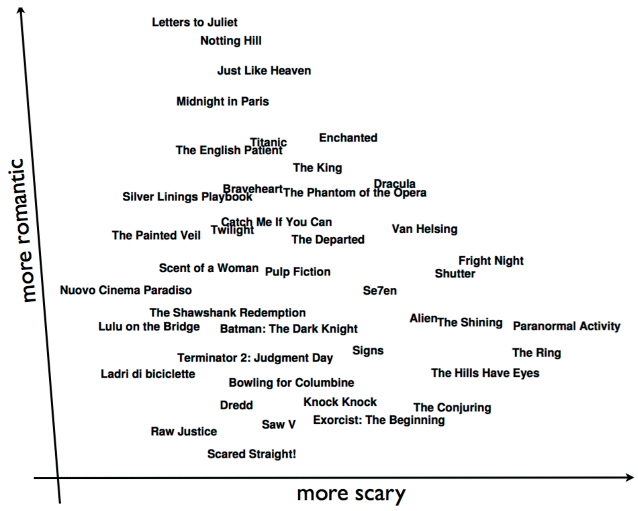

Figure 1: Directions modelling the relative properties *scary* and *romantic* (showing a two-dimensional projection of a 100-dimensional space).

References

Beltagy, I.; Chau, C.; Boleda, G.; Garrette, D.; Erk, K.; and Mooney, R. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proc. 2nd Joint Conf. on Lexical and Computational Semantics*, 11–21.

d'Amato, C.; Fanizzi, N.; Fazzinga, B.; Gottlob, G.; and Lukasiewicz, T. 2010. Combining semantic web search with the power of inductive reasoning. In *Proc. Conference on Scalable Uncertainty Management*. 137–150.

Derrac, J., and Schockaert, S. 2014a. Characterising semantic relatedness using interpretable directions in conceptual spaces. In *Proc. ECAI*. 243–248.

Derrac, J., and Schockaert, S. 2014b. Enriching taxonomies of place types using Flickr. In *Proc. FOIKS*. 174–192.

Freitas, A.; da Silva, J. C.; Curry, E.; and Buitelaar, P. 2014. A distributional semantics approach for selective reasoning on commonsense graph knowledge bases. In *Proc. 19th International Conference on Applications of Natural Language to Information Systems*. 21–32.

Gärdenfors, P., and Williams, M. 2001. Reasoning about categories in conceptual spaces. In *Proc. IJCAI*, 385–392.

Lao, N.; Mitchell, T.; and Cohen, W. W. Random walk inference and learning in a large scale knowledge base. In *Proc. EMNLP*, 529–539.

Speer, R.; Havasi, C.; and Lieberman, H. 2008. Analogyspace: reducing the dimensionality of common sense knowledge. In *Proc. AAAI*, 548–553.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.

West, R.; Gabrilovich, E.; Murphy, K.; Sun, S.; Gupta, R.; and Lin, D. 2014. Knowledge base completion via search-based question answering. In *Proc. WWW*, 515–526.