Identifying hypernyms in distributional semantic spaces

Alessandro Lenci
University of Pisa, Dept. of Linguistics
via S. Maria 36
I-56126, Pisa, Italy
alessandro.lenci@ling.unipi.it

Giulia Benotto
University of Pisa, Dept. of Linguistics
via S. Maria 36
I-56126, Pisa, Italy
mezzanine.g@gmail.com

Abstract

In this paper we apply existing directional similarity measures to identify hypernyms with a state-of-the-art distributional semantic model. We also propose a new directional measure that achieves the best performance in hypernym identification.

1 Introduction and related works

Distributional Semantic Models (DSMs) measure the semantic similarity between words with proximity in distributional space. However, semantically similar words in turn differ for the type of relation holding between them: e.g., *dog* is strongly similar to both *animal* and *cat*, but with different types of relations. Current DSMs accounts for these facts only partially. While they may correctly place both *animal* and *cat* among the nearest distributional neighbors of *dog*, they are not able to characterize the different semantic properties of these relations, for instance the fact that hypernymy is an asymmetric semantic relation, since being a dog entails being an animal, but not the other way round.

The purpose of this paper is to explore the possibility of identifying hypernyms in DSMs with *directional (or asymmetric) similarity measures* (Kotlerman et al., 2010). These measures all rely on some variation of the **Distributional Inclusion Hypothesis**, according to which if u is a semantically narrower term than v, then a significant number of salient distributional features of u is included in the feature vector of v as well. Since hypernymy is an asymmetric relation and hypernyms are semantically broader terms than their hyponyms, then we can predict that directional similarity measures are better suited to identify terms related by the hypernymy relation.

Automatic identification of hypernyms in corpora is a long-standing research line, but most methods have adopted semi-supervised, pattern-based approaches (Hearst, 1992; Pantel and Pennacchiotti, 2006). Fully unsupervised hypernym identification with DSMs is still a largely open field. Various models to represent hypernyms in vector spaces have recently been proposed (Weeds and Weir, 2003; Weeds et al., 2004; Clarke, 2009), usually grounded on the Distributional Inclusion Hypothesis (for a different approach based on representing word meaning as "regions" in vector space, see Erk (2009a; 2009b)). The same hypothesis has been adopted by Kotlerman *et al.* (2010) to identify (substitutable) lexical entailments" . Within the context of the Textual Entailment (TE) paradigm, Zhitomirsky-Geffet and Dagan (2005; 2009) define *(substitutable) lexical entailment* as a relation holding between two words, if there are some contexts in which one of the words can be substituted by the other and the meaning of the original word can be inferred from the new one. Its relevance for TE notwithstanding, this notion of lexical entailment is more general and looser than hypernymy. In fact, it encompasses several standard semantic relations such as synonymy, hypernymy, metonymy, some cases of meronymy, etc.

Differently from Kotlerman *et al.* (2010), here we focus on applying directional, asymmetric similarity measures to identify hypernyms. We assume the classical definition of a hypernymy, such that Y is

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 75–79,
Montréal, Canada, June 7-8, 2012.

an hypernym of X if and only if X is a kind of Y, or equivalently every X is a Y.

2 Directional similarity measures

In the experiments reported in section 3 we have applied the following directional similarity measures (F_x is the set of distributional features of a term x, $w_x(f)$ is the weight of the feature f for x):

WeedsPrec (**M1**) - this is a measure that quantifies the weighted inclusion of the features of a term u within the features of a term v (Weeds and Weir, 2003; Weeds et al., 2004; Kotlerman et al., 2010):

$$WeedsPrec(u,v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)} \quad (1)$$

cosWeeds (**M2**) - this measure corresponds to the geometrical average of *WeedsPrec* and the symmetric similarity between u and v, measured by their vectors' cosine:

$$cosWeeds(u,v) = \sqrt{M1(u,v) * cos(u,v)} \quad (2)$$

This is actually a variation of the *balPrec* measure in Kotlerman *et al.* (2010), the difference being that cosine is used as a symmetric similarity measure instead of the *LIN* measure (Lin, 1998).

ClarkeDE (**M3**) - a close variation of M1, proposed by Clarke (2009):

$$ClarkeDE(u,v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)} \quad (3)$$

invCL (**M4**) - this a new measure that we introduce and test here for the first time. It takes into account not only the inclusion of u in v, but also the *non-inclusion* of v in u, both measured with *ClarkeDE*:

$$invCL(u,v) = \sqrt{M3(u,v) * (1 - M3(v,u))} \quad (4)$$

The intuition behind *invCL* is that, if v is a semantically broader term of u, then the features of u are included in the features of v, but crucially the features of v are also *not* included in the features of u. For instance, if *animal* is a hypernym of *lion*, we can expect i.) that a significant number of the *lion*-contexts are also *animal*-contexts, and ii.) that a significant number of *animal*-contexts are not *lion*-contexts. In fact, being a semantically broader term of *lion*, *animal* should also be found in contexts in which animals other than lions occur.

3 Experiments

The main purpose of the experiments reported below is to investigate the ability of the directional similarity measures presented in section 2 to identify the hypernyms of a given target noun, and to discriminate hypernyms from terms related by symmetric semantic relations, such as coordinate terms.

We have represented lexical items with distributional feature vectors extracted from the *TypeDM* tensor (Baroni and Lenci, 2010). TypeDM is a particular instantiation of the *Distributional Memory* (DM) framework. In DM, distributional facts are represented as a *weighted tuple structure* T, a set of weighted word-link-word tuples $\langle\langle w_1, l, w_2\rangle, \sigma\rangle$, such that w_1 and w_2 are content words (e.g. nouns, verbs, etc.), l is a syntagmatic co-occurrence links between words in a text (e.g. syntactic dependencies, etc.), and σ is a weight estimating the statistical salience of that tuple. The TypeDM word set contains 30,693 lemmas (20,410 nouns, 5,026 verbs and 5,257 adjectives). The TypeDM link set contains 25,336 direct and inverse links formed by (partially lexicalized) syntactic dependencies and patterns. The weight σ is the *Local Mutual Information* (LMI) (Evert, 2005) computed on link type frequency (negative LMI values are raised to 0).

3.1 Test set

We have evaluated the directional similarity measures on a subset of the BLESS data set (Baroni and Lenci, 2011), consisting of tuples expressing a **relation** between a target concept (henceforth referred to as **concept**) and a relatum concept (henceforth referred to as **relatum**). BLESS includes 200 distinct English concrete nouns as target concepts, equally divided between living and non-living entities, and grouped into 17 broader classes (e.g., BIRD, FRUIT, FURNITURE, VEHICLE, etc.).

For each concept noun, BLESS includes several

relatum words, linked to the concept by one of 5 semantic relations. Here, we have used the BLESS subset formed by 14,547 tuples with the relatum attested in the TypeDM word set, and containing one of these relations: COORD: the relatum is a noun that is a co-hyponym (coordinate) of the concept: $\langle alligator, coord, lizard\rangle$; HYPER: the relatum is a noun that is a hypernym of the concept: $\langle alligator, hyper, animal\rangle$; MERO: the relatum is a noun referring to a part/component/organ/member of the concept, or something that the concept contains or is made of: $\langle alligator, mero, mouth\rangle$; RANDOM-N: the relatum is a random noun holding no semantic relation with the target concept: $\langle alligator, random-n, message\rangle$.

Kotlerman *et al.* (2010) evaluate a set of directional similarity measure on a data set of valid and invalid (substitutable) lexical entailments (Zhitomirsky-Geffet and Dagan, 2009). However, as we said above, lexical entailment is defined as an asymmetric relation that covers various types of classic semantic relations, besides hypernymy . The choice of BLESS is instead motivated by the fact that here we focus on the ability of directional similarity measure to identify hypernyms.

3.2 Evaluation and results

For each word x in the test set, we represented x in terms of a set F_x of distributional features $\langle l, w_2\rangle$, such that in the TypeDM tensor there is a tuple $\langle\langle w_1, l, w_2\rangle, \sigma\rangle$, $w_1 = x$. The feature weight $w_x(f)$ is equal to the weight σ of the original DM tuple. Then, we applied the 4 directional similarity measures in section 2 to BLESS, with the goal of evaluating their ability to discriminate hypernyms from other semantic relations, in particular co-hyponymy. In fact, differently from hypernyms, coordinate terms are not related by inclusion. Therefore, we want to test whether directional similarity measures are able to assign higher scores to hypernyms, as predicted by the Distributional Inclusion Hypothesis. We used the *Cosine* as our baseline, since it is a symmetric similarity measure and it is commonly used in DSMs.

We adopt two different evaluation methods. The first is based on the methodology described in Baroni and Lenci (2011). Given the similarity scores for a concept with all its relata across all relations in our test set, we pick the relatum with the highest score (nearest neighbour) for each relation. In this way, for each of the 200 BLESS concepts, we obtain 4 similarity scores, one per relation. In order to factor out concept-specific effects that might add to the overall score variance, we transform the 8 similarity scores of each concept onto standardized z scores (mean: 0; s.d: 1) by subtracting from each their mean, and dividing by their standard deviation. After this transformation, we produce a **boxplot** summarizing the distribution of scores per relation across the 200 concepts.

Boxplots for each similarity measure are reported in Figure 1. They display the median of a distribution as a thick horizontal line within a box extending from the first to the third quartile, with whiskers covering 1.5 of the interquartile range in each direction from the box, and values outside this extended range – extreme outliers – plotted as circles (these are the default boxplotting option of the R statistical package). To identify significant differences between relation types, we also performed pairwise comparisons with the Tukey Honestly Significant Difference test, using the standard $\alpha = 0.05$ significance threshold.

In the boxplots we can observe that all measures (either symmetric or not) are able to discriminate truly semantically related pairs from unrelated (i.e. random) ones. Crucially, *Cosine* shows a strong tendency to identify coordinates among the nearest neighbors of target items. This is actually consistent with its being a symmetric similarity measure. Instead, directional similarity measures significantly promote hypernyms over coordinates. The only exception is represented by *cosWeeds*, which again places coordinates at the top, though now the difference with hypernyms is not significant. This might be due to the cosine component of this measure, which reduces the effectiveness of the asymmetric *WeedsPrec*. The difference between coordinates and hypernyms is slightly bigger in *invCL*, and the former appear to be further downgraded than with the other directional measures. From the boxplot analysis, we can therefore conclude that similarity measures based on the Distributional Inclusion Hypothesis do indeed improve hypernym identification in context-feature semantic spaces, with respect to other types of semantic relations, such as COORD.

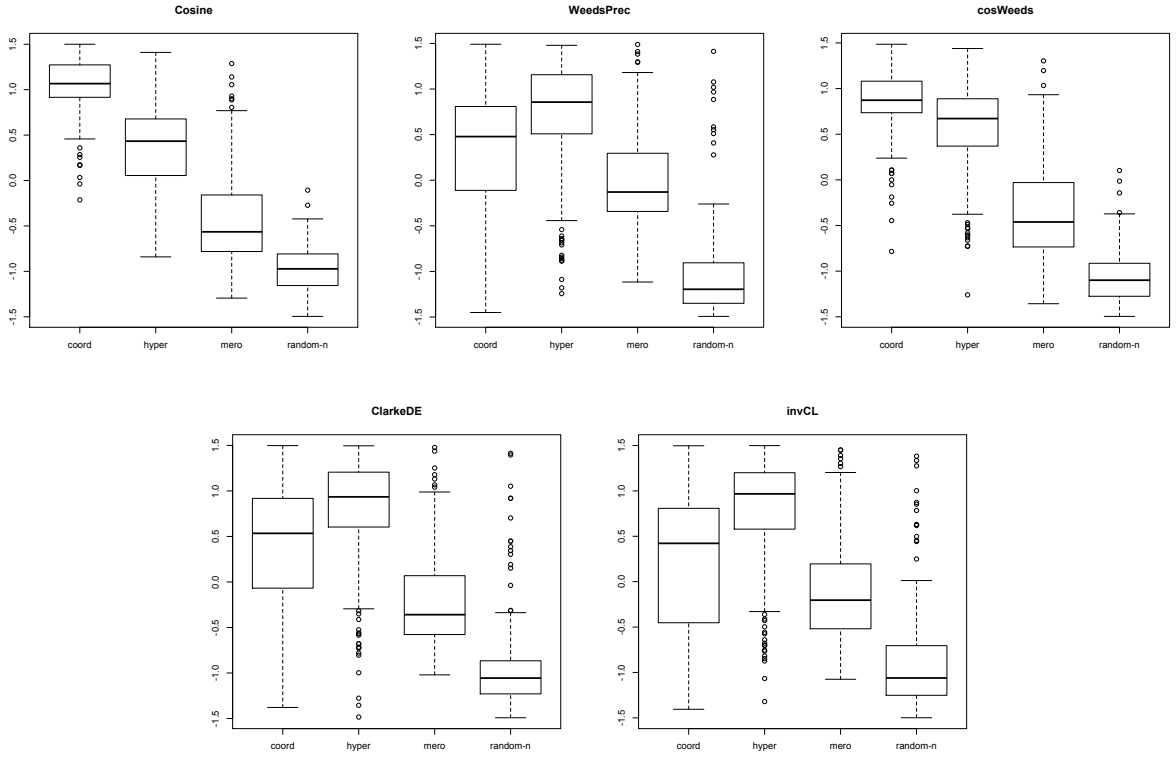

Figure 1: Distribution of relata similarity scores across concepts (values on ordinate are similarity scores after concept-by-concept z-normalization).

The second type of evaluation we have performed is based on Kotlerman *et al.* (2010). The similarity measures have been evaluated with **Average Precision** (AP), a method derived from Information Retrieval and combining precision, relevance ranking and overall recall. For each similarity measure, we computed AP with respect to the 4 BLESS relations. The best possible score (AP = 1) for a given relation (e.g., HYPER) corresponds to the ideal case in which all the relata belonging to that relation have higher similarity scores than the relata belonging to the other relations. For every relation, we calculated the AP for each of the 200 BLESS target concepts.

In Table 1, we report the AP values averaged over the 200 concepts. On the one hand, these results confirm the trend illustrated by the boxplots, in particular the fact that directional similarity measures clearly outperform *Cosine* (or cosine-based measures such as *cosWeeds*) in identifying hypernyms, with no significant differences among them. However, a different picture emerges by comparing the AP values for HYPER with those for COORD. since in this case important distinctions among the directional measures emerge. In fact, even if *WeedsPrec* and *ClarkeDE* increase the AP for HYPER, still they assign even higher AP values to COORD. Conversely, *invCL* is the only measure that assigns to HYPER the top AP score, higher than COORD too.

measure	COORD	HYPER	MERO	RANDOM-N
Cosine	0.79	0.23	0.21	0.30
WeedsPrec	0.45	0.40	0.31	0.32
cosWeeds	0.69	0.29	0.23	0.30
ClarkeDE	0.45	0.39	0.28	0.33
invCL	**0.38**	**0.40**	0.31	0.34

Table 1: Mean AP values for each semantic relation reported by the different similarity scores.

The new directional similarity measure we have proposed in this paper, *invCL*, thus reveals a higher ability to set apart hypernyms from other relations, coordinates terms included. The latter are expected

to share a large number of contexts and this is the reason why they are strongly favored by symmetric similarity measures, such as *Cosine*. Asymmetric measures like *cosWeeds* and *ClarkeDE* also fall short of distinguishing hypernyms from coordinates because the condition of feature inclusion they test is satisfied by coordinate terms as well. If two sets share a high number of elements, then many elements of the former are also included in the latter, and vice versa. Therefore, coordinate terms too are expected to have high values of feature inclusions. Conversely, *invCL* takes into account not only the inclusion of u into v, but also the amount of v that is not included in u. Thus, *invCL* provides a better distributional correlate to the central property of hypernyms of having a broader semantic content than their hyponyms.

4 Conclusions and ongoing research

The experiments reported in this paper support the Distributional Inclusion Hypothesis as a viable approach to model hypernymy in semantic vector spaces. We have also proposed a new directional measure that actually outperforms the state-of-the-art ones. Focusing on the contexts that broader terms do not share with their narrower terms thus appear to be an interesting direction to explore to improve hypernym identification. Our ongoing research includes testing *invCL* to recognize lexical entailments and comparing it with the *balAPinc* measured proposed by Kotlerman *et al.* (2010) for this task, as well as designing new distributional methods to discriminate between various other types of semantic relations.

Acknowledgments

We thank the reviewers for their useful and insightful comments on the paper.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, Edinburgh, Scotland, UK: 1–10.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens, Greece: 112–119.

Katrin Erk. 2009a. Supporting inferences in semantic space: representing words as regions. In *Proceedings of the 8th International Conference on Computational Semantics*, Tilburg, January: 104–115.

Katrin Erk. 2009b. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, Boulder, Colorado: 57–65.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D. dissertation, Stuttgart University.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING 1992*, Nantes, France: 539–545.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04): 359–389.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the COLING-ACL 1998*, Montreal, Canada: 768–774.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the COLING-ACL 2006*, Sydney, Australia: 113–120.

Idan Szpektor and Ido Dagan. 2008. Learning Entailment Rules for Unary Templates. In *Proceedings of COLING 2008*, Manchester, UK: 849–856.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the EMNLP 2003*, Sapporo, Japan: 81–88.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING 2004*, Geneva, Switzerland: 1015–1021.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL 2005*, Ann Arbor, MI: 107–114.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3): 435-461.