Chasing Hypernyms in Vector Spaces with Entropy

Enrico Santus
Dept. of Chinese and Bilingual Studies
The Hong Kong Polytechnic University, Hong Kong
e.santus@connect.polyu.hk

Alessandro Lenci
CoLing Lab – Dept. of Philology, Literature, and Linguistics
University of Pisa, Italy
alessandro.lenci@ling.unipi.it

Qin Lu
Dept. of Computing
The Hong Kong Polytechnic University, Hong Kong
csluqin@comp.polyu.edu.hk

Sabine Schulte im Walde
Inst. for Natural Language Processing
University of Stuttgart
Germany
schulte@ims.uni-stuttgart.de

Abstract

In this paper, we introduce *SLQS*, a new entropy-based measure for the unsupervised identification of hypernymy and its directionality in Distributional Semantic Models (DSMs). *SLQS* is assessed through two tasks: (i.) identifying the hypernym in hyponym-hypernym pairs, and (ii.) discriminating hypernymy among various semantic relations. In both tasks, *SLQS* outperforms other state-of-the-art measures.

1 Introduction

In recent years, Distributional Semantic Models (DSMs) have gained much attention in computational linguistics as unsupervised methods to build lexical semantic representations from corpus-derived co-occurrences encoded as distributional vectors (Sahlgren, 2006; Turney and Pantel, 2010). DSMs rely on the *Distributional Hypothesis* (Harris, 1954) and model lexical semantic similarity as a function of distributional similarity, which is most commonly measured with the *vector cosine* (Turney and Pantel, 2010). DSMs have achieved impressive results in tasks such as synonym detection, semantic categorization, etc. (Padó and Lapata, 2007; Baroni and Lenci, 2010).

One major shortcoming of current DSMs is that they are not able to discriminate among different types of semantic relations linking distributionally similar lexemes. For instance, the nearest neighbors of *dog* in vector spaces typically include hypernyms like *animal*, co-hyponyms like *cat*, meronyms like *tail*, together with other words semantically related to *dog*. DSMs tell us how similar these words are to *dog*, but they do not give us a principled way to single out the items linked by a specific relation (e.g., hypernyms).

Another related issue is to what extent distributional similarity, as currently measured by DSMs, is appropriate to model the semantic properties of a relation like hypernymy, which is crucial for Natural Language Processing. Similarity is by definition a symmetric notion (*a* is similar to *b* if and only if *b* is similar to *a*) and it can therefore naturally model symmetric semantic relations, such as synonymy and co-hyponymy (Murphy, 2003). It is not clear, however, how this notion can also model hypernymy, which is asymmetric. In fact, it is not enough to say that *animal* is distributionally similar to *dog*. We must also account for the fact that *animal* is semantically broader than *dog*: every *dog* is an *animal*, but not every *animal* is a *dog*.

In this paper, we introduce *SLQS*, a new entropy-based distributional measure that aims to identify hypernyms by providing a distributional characterization of their *semantic generality*. We assess it with two tasks: (i.) the identification of the broader term in hyponym-hypernym pairs (*directionality task*); (ii.) the discrimination of hypernymy among other semantic relations (*detection task*). Given the centrality of hypernymy, the relevance of the themes we address hardly needs any further motivation. Improving the ability of DSMs to identify hypernyms is in fact extremely important in tasks such as Recognizing Textual Entailment (RTE) and ontology learning, as well as to enhance the cognitive plausibility of DSMs as general models of the semantic lexicon.

2 Related work

The problem of identifying asymmetric relations like hypernymy has so far been addressed in distributional semantics only in a limited way (Kotlerman et al., 2010) or treated through semi-supervised approaches, such as pattern-based approaches (Hearst, 1992). The few works that have attempted a completely unsupervised approach to the identification of hypernymy in corpora have mostly relied on some versions of the *Distributional Inclusion Hypothesis* (DIH; Weeds and Weir, 2003; Weeds et al., 2004), according to which the contexts of a narrow term are also shared by the broad term.

One of the first proposed measures formalizing the DIH is *WeedsPrec* (Weeds and Weir, 2003; Weeds et al., 2004), which quantifies the weights of the features *f* of a narrow term *u* that are included into the set of features of a broad term *v*:

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

where F_x is the set of features of a term *x*, and $w_x(f)$ is the weight of the feature *f* of the term *x*. Variations of this measure have been introduced by Clarke (2009), Kotlerman et al. (2010) and Lenci and Benotto (2012).

In this paper, we adopt a different approach, which is not based on DIH, but on the hypothesis that hypernyms are semantically more general than hyponyms, and therefore tend to occur in less informative contexts than hypernyms.

3 *SLQS*: A new entropy-based measure

DIH is grounded on an "extensional" definition of the asymmetric character of hypernymy: since the class (i.e., extension) denoted by a hyponym is included in the class denoted by the hypernym, hyponyms are expected to occur in a subset of the contexts of their hypernyms. However, it is also possible to provide an "intensional" definition of the same asymmetry. In fact, the typical characteristics making up the "intension" (i.e., concept) expressed by a hypernym (e.g., *move* or *eat* for *animal*) are semantically more general than the characteristics forming the "intension" of its hyponyms (e.g., *bark* or *has fur* for *dog*). This corresponds to the idea that superordinate terms like *animal* are less informative than their hyponyms (Murphy, 2002). From a distributional point of view, we can therefore expect that the most typical linguistic contexts of a hypernym are less informative than the most typical linguistic contexts of its hyponyms. In fact, contexts such as *bark* and *has fur* are likely to co-occur with a smaller number of words than *move* and *eat*. Starting from this hypothesis and using entropy as an estimate of context informativeness (Shannon, 1948), we propose *SLQS*, which measures the semantic generality of a word by the entropy of its statistically most prominent contexts.

For every term w_i we identify the *N* most associated contexts *c* (where *N* is a parameter empirically set to 50)[1]. The association strength has been calculated with *Local Mutual Information* (LMI; Evert, 2005). For each selected context *c*, we define its entropy *H(c)* as:

[1] *N=50* is the result of an optimization of the model against the dataset after trying the following suboptimal values: 5, 10, 25, 75 and 100.

$$H(c) = -\sum_{i=1}^{n} p(f_i|c) \cdot log_2\big(p(f_i|c)\big)$$

where $p(f_i|c)$ is the probability of the feature f_i given the context c, obtained through the ratio between the frequency of $<c, f_i>$ and the total frequency of c. The resulting values $H(c)$ are then normalized in the range 0-1 by using the Min-Max-Scaling (Priddy and Keller, 2005): $H_n(c)$. Finally, for each term w_i we calculate the median entropy E_{wi} of its N contexts:

$$E_{w_i} = Me_{j=1}^{N}\left(H_n(c_j)\right)$$

E_{w_i} can be considered as a *semantic generality index* for the term w_i: the higher E_{w_i}, the more semantically general w_i is. *SLQS* is then defined as the reciprocal difference between the semantic generality E_{w_1} and E_{w_2} of two terms w_1 and w_2:

$$SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}}$$

According to this formula, $SLQS<0$, if $E_{w_1}>E_{w_2}$; $SLQS\simeq 0$, if $E_{w_1}\simeq E_{w_2}$; and $SLQS>0$, if $E_{w_1}<E_{w_2}$. *SLQS* is an asymmetric measure because, by definition, $SLQS(w_1,w_2)\neq SLQS(w_2,w_1)$ (except when w_1 and w_2 have exactly the same generality). Therefore, if $SLQS(w_1,w_2)>0$, w_1 is semantically less general than w_2.

4 Experiments and evaluation

4.1 The DSM and the dataset

For the experiments, we used a standard window-based DSM recording co-occurrences with the nearest 2 content words to the left and right of each target word. Co-occurrences were extracted from a combination of the freely available ukWaC and WaCkypedia corpora (with 1.915 billion and 820 million words, respectively) and weighted with LMI.

To assess *SLQS* we relied on a subset of *BLESS* (Baroni and Lenci, 2011), a freely-available dataset that includes 200 distinct English concrete nouns as target concepts, equally divided between living and non-living entities (e.g. `BIRD`, `FRUIT`, etc.). For each target concept, *BLESS* contains several *relata*, connected to it through one relation, such as co-hyponymy (`COORD`), hypernymy (`HYPER`), meronymy (`MERO`) or no-relation (`RANDOM-N`).[2]

Since *BLESS* contains different numbers of pairs for every relation, we randomly extracted a subset of 1,277 pairs for each relation, where 1,277 is the maximum number of `HYPER`-related pairs for which vectors existed in our DSM.

4.2 Task 1: Directionality

In this experiment we aimed at identifying the hypernym in the 1,277 hypernymy-related pairs of our dataset. Since the `HYPER`-related pairs in *BLESS* are in the order hyponym-hypernym (e.g. *eagle-bird*, *eagle-animal*, etc.), the hypernym in a pair (w_1,w_2) is correctly identified by *SLQS*, if $SLQS(w_1,w_2) > 0$. Following Weeds et al. (2004), we used word frequency as a baseline model. This baseline is grounded on the hypothesis that hypernyms are more frequent than hyponyms in corpora. Table 1 gives the evaluation results:

	SLQS	WeedsPrec	BASELINE
POSITIVE	1111	805	844
NEGATIVE	166	472	433
TOTAL	**1277**	**1277**	**1277**
PRECISION	87.00%	63.04%	66.09%

Table 1. Accuracy for Task 1.

As it can be seen in Table 1, *SLQS* scores a precision of 87% in identifying the second term of the test pairs as the hypernym. This result is particularly significant when compared to the one obtained by applying WeedsPrec (+23.96%). As it was also noticed by Geffet and Dagan (2005) with reference to a previous similar experiment performed on a different corpus (Weeds et al., 2004), the WeedsPrec precision in this task is comparable to the naïve baseline. *SLQS* scores instead a +20.91%.

[2] In these experiments, we only consider the *BLESS* pairs containing a noun relatum.

4.3 Task 2: Detection

The second experiment aimed at discriminating HYPER test pairs from those linked by other types of relations in *BLESS* (i.e., MERO, COORD and RANDOM-N). To this purpose, we assumed that hypernymy is characterized by two main properties: (i.) the hypernym and the hyponym are distributionally similar (in the sense of the *Distributional Hypothesis*), and (ii.) the hyponym is semantically less general than the hypernym. We measured the first property with the *vector cosine* and the second one with *SLQS*.

After calculating *SLQS* for all the pairs in our datasets, we set to zero all the negative values, that is to say those in which – according to *SLQS* – the first term is semantically more general than the second one. Then, we combined *SLQS* and *vector cosine* by their product. The greater the resulting value, the greater the likelihood that we are considering a hypernymy-related pair, in which the first word is a hyponym and the second word is a hypernym.

To evaluate the performance of *SLQS*, we used *Average Precision* (AP; Kotlerman et al., 2010), a method derived from Information Retrieval that combines precision, relevance ranking and overall recall, returning a value that ranges from 0 to 1. AP=1 means that all the instances of a relation are in the top of the rank, whereas AP=0 means they are in the bottom. AP is calculated for the four relations we extracted from *BLESS*. *SLQS* was also compared with *WeedsPrec* and *vector cosine*, again using frequency as baseline. Table 2 shows the results:

	HYPER	COORD	MERO	RANDOM
Baseline	0.40	0.51	0.38	0.17
Cosine	0.48	0.46	0.31	0.21
WeedsPrec	0.50	0.35	0.39	0.21
*SLQS * Cosine*	**0.59**	**0.27**	**0.35**	**0.24**

Table 2. AP values for Task 2.

The AP values show the performances of the tested measures on the four relations. The optimal result would be obtained scoring 1 for HYPER and 0 for the other relations.

The product between *SLQS* and *vector cosine* gets the best performance in identifying HYPER (+0.09 in comparison to *WeedsPrec*) and in discriminating it from COORD (-0.08 than *WeedsPrec*). It also achieves better results in discriminating MERO (-0.04 than *WeedsPrec*). On the other hand, it seems to get a slightly lower precision in discriminating RANDOM-N (+0.03 in comparison to *WeedsPrec*). The likely reason is that unrelated pairs might also have a fairly high semantic generality difference, slightly affecting the measure's performance. Figure 1 gives a graphic depiction of the performances. *SLQS* corresponds to the black line in comparison to the *WeedsPrec* (black borders, grey fill), the *vector cosine* (grey borders) and the baseline (grey fill).

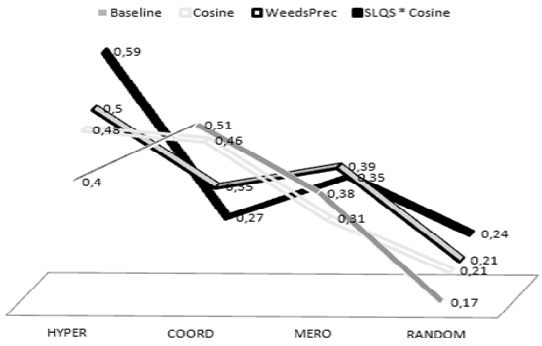

Figure 1. AP values for Task 2.

5 Conclusions and future work

In this paper, we have proposed *SLQS*, a new asymmetric distributional measure of semantic generality which is able to identify the broader term in a hypernym-hyponym pair and, when combined with *vector cosine*, to discriminate hypernymy from other types of semantic relations. The successful performance of *SLQS* in the reported experiments confirms that hyponyms and hypernyms are distributionally similar, but hyponyms tend to occur in more informative contexts than hypernyms. *SLQS* shows that an "intensional" characterization of hypernymy can be pursued in distributional terms. This opens up new possibilities for the study of semantic relations in DSMs. In further research, *SLQS* will also be tested on other datasets and languages.

References

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

Baroni, Marco and Lenci, Alessandro. 2011. "How we BLESSed distributional semantic evaluation". *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*. Edinburg, UK. 1-10.

Clarke, Daoud. 2009. "Context-theoretic semantics for natural language: An overview". *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Athens, Greece. 112-119.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Geffet, Maayan and Dagan, Idan. 2005. "The Distributional Inclusion Hypotheses and Lexical Entailment". *Proceedings of 43rd Annual Meeting of the ACL*. Michigan, USA. 107-114.

Harris, Zellig. 1954. "Distributional structure". *Word*, Vol. 10 (23). 146-162.

Hearst, Marti A. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora". *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France. 539-545.

Kotlerman, Lili, Dagan, Ido, Szpektor, Idan, and Zhitomirsky-Geffet, Maayan. 2010. "Directional Distributional Similarity for Lexical Inference". *Natural Language Engineering*, Vol. 16 (4). 359-389.

Lenci, Alessandro and Benotto, Giulia. 2012. "Identifying hypernyms in distributional semantic spaces". *SEM 2012 – The First Joint Conference on Lexical and Computational Semantics*. Montréal, Canada. Vol. 2. 75-79.

Murphy, Gregory L.. 2002. *The Big Book of Concepts*. The MIT Press, Cambridge, MA.

Murphy, M. Lynne. 2003. *Lexical meaning*. Cambridge University Press, Cambridge.

Padó, Sebastian and Lapata, Mirella. 2007. "Dependency-based Construction of Semantic Space Models". *Computational Linguistics*, Vol. 33 (2). 161-199.

Priddy, Kevin L. and Keller, Paul E. 2005. *Artificial Neural Networks: An Introduction*. SPIE Press - International Society for Optical Engineering, October 2005.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. dissertation, Department of Linguistics, Stockholm University.

Shannon, Claude E. 1948. "A mathematical theory of communication". *Bell System Technical Journal*, Vol. 27. 379-423 and 623-656.

Turney, Peter D. and Pantel, Patrick. 2010. "From Frequency to Meaning: Vector Space Models of Semantics". *Journal of Articial Intelligence Research*, Vol. 37. 141-188.

Weeds, Julie and Weir, David. 2003. "A general framework for distributional similarity". *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Sapporo, Japan. 81-88.

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