AHyDA: Automatic Hypernym Detection with feature Augmentation

Ludovica Pannitto
University of Pisa
ellepannitto@gmail.com

Lavinia Salicchi
University of Pisa
lavinia.salicchi@libero.it

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

English. Several unsupervised methods for hypernym detection have been investigated in distributional semantics. Here we present a new approach based on a smoothed version of the distributional inclusion hypothesis. The new method is able to improve hypernym detection after testing on the BLESS dataset.

Italiano. *Sulla base dei metodi non supervisionati presenti in letteratura, affrontiamo il task di riconoscimento di iperonimi nello spazio distribuzionale. Introduciamo una nuova misura direzionale, basata su un'espansione dell'ipotesi di inclusione distribuzionale, che migliora il riconoscimento degli iperonimi, testandola sul dataset* BLESS.

1 Introduction and related works

Within the Distributional Semantics framework, semantic similarity between words is usually expressed in terms of proximity in a semantic space, where the dimensions of the space represent, at some level of abstraction, the contexts in which the words occur.

Our intuitions about the meaning of words allow inferences of the kind expressed in example (1), and we expect Distributional Semantic Models (DSMs) to support such inferences:

(1) a. Wilbrand *invented* TNT → Wilbrand *uncovered* TNT
b. A *horse ran* → An *animal moved*

The type of relation between semantically similar lexemes may differ significantly, but DSMs only account for a generic notion of semantic relatedness. Furthermore, not all lexical relations are symmetrical (see example (2)), while most of the similarity measures defined in distributional semantics are, like the cosine.

(2) a. I saw a *dog* → I saw an *animal*
b. I saw an *animal* ↛ I saw a *dog*

Hypernymy is an asymmetric relation. Automatic hypernym identification is a very well-known task in literature, which has mostly been addressed with semi-supervised, pattern-based approaches (Hearst, 1992; Pantel and Pennacchiotti, 2006). Various unsupervised models have been proposed (Weeds and Weir, 2003; Weeds et al., 2004; Clarke, 2009; Lenci and Benotto, 2012; Santus et al., 2014), based on the notion of **Distributional Generality** (Weeds et al., 2004) and on the **Distributional Inclusion Hypothesis** (DIH) (Geffet and Dagan, 2005) which has been derived from it.

1.1 The pitfalls of the DIH

The DIH aims at providing a distributional correlate of the extensional definition of hyponymy in terms of set inclusion: x is a hyponym of y iff the extension of x (i.e. the set of entities denoted by x) is a subset of the extension of y. The DIH turns this into the assumption that a significant number of the most salient contexts of x should also appear among the salient contexts of y. While this is consistent with the logical inferences licensed by hyponymy (cf. (2)), it does not take into account the actual usage of hypernyms with respect to hyponyms. Consider for instance the following examples:

(3) a. A *horse* gallops $\xrightarrow{?}$ An *animal* gallops
b. A *dog* barks $\xrightarrow{?}$ An *animal* barks

These inferences are truth-conditionally valid: whenever the antecedent is true, the consequent is also true. However, they are not equally "pragmatically" sound. In fact, the fact that one uses

	horse	*dog*	*animal*
gallop	216	–	7
bark	–	869	16

Table 1: Co-occurrence frequency distribution extracted from the ukWaC corpus

a sentence like *A dog barks* does not entail that in the same situation one would have also used the sentence *An animal barks*. The latter sentence would be pragmatically appropriate only in cases in which one knows that something is barking, without knowing which animal is producing this sound. However, the latter condition hardly applies, since barking is a very typical feature of dogs: knowing that something is barking typically entails knowing that it is a dog, since we know that barking is something dogs do. The same argument also applies to the case of *horse* and *galloping*.

The problem of the DIH is that the assumption it rests on, namely that the most typical contexts of the hyponym are also typical contexts of the hypernym, is not borne out in practical language usage because of pragmatic constraints. The most typical contexts of an hyponym are not necessarily the typical contexts of its hypernym. This is also proved by a simple inspection of corpus data, as reported in Table 1. Despite *animal* $(161,107)$ is more frequent than *dog* $(128,765)$ and *horse* $(90,437)$, its co-occurrence with *bark* and *gallop* is much lower than the ones of the hyponyms: *bark* and *gallop* are not typical contexts of *animal*.

If the inferences in (3) are pragmatically odd, the following ones are instead fully acceptable:

(4) a. A *horse* gallops → An *animal* moves
b. A *dog* barks → An *animal* calls

Salient features of the *hypernym* are indeed supposed to be semantically more general than the salient features of the *hyponym*. Santus et al. (2014) tried to capture this fact by abandoning the DIH and introducing an entropy-based measure to estimate of informativeness of the hypernym and hyponym contexts, under the assumption that the former have a higher entropy, because they are more general (e.g. *move* vs. *gallop*).

In this paper, we address the same issue by amending the DIH, to make it more consistent with the actual distributional properties of hyponyms and hypernyms. Therefore, we introduce **AHyDA** (Automatic Hypernym Detection with feature Augmentation), a smoothed version of the DIH: given a context feature f that is salient for a lexical item x, we expect *co-hyponyms* of x to have some feature g that is similar to f, and an *hypernym* of x to have a number of these clusters of features. To remain in the animal sounds area, we expect a *dog* to *bark* and a *duck* to *quack* and an *animal* to produce either of those sounds or to co-occur with a more general sound-emission verb.

2 AHyDA: Smoothing the DIH

All the measures implementing the DIH are based on computing the (weighted) intersection of the features of the hyponym and the hypernym. This is then typically divided by the hyponym features. AHyDA essentially proposes a new way to compute the intersection of the hyponym and hypernym contexts. Given a lexical item x, we call F_x the set of its distributional features. Note that features need not be pure lexical items. In general, we define f as a pair (f_w, f_r) where f_w is typically a lexical item, and f_r is any additional contextual information, in the present case a pattern occurring between x and f_w, as explained in section 3.1. The core novelty of AHyDA is to use a smoothed version of F_x, called F'_x.

The idea is shown in figure 1, which provides a simplified graphical example of the intersection operation. Consider a case where the target *horse* has some feature with *gallop* as a lexical item, for example a feature $f = (gallop, sbj)$ meaning that *horse* is a possible subject of *gallop*. Given what we have said in Section 1.1, we do not expect *animal* to share this *horse*-specific property. So, instead of looking for this particular feature among the ones of *animal*, we generate a new set $N_{horse}(gallop)$ of features $g = (g_w, f_r)$ such that g_w is a neighbor of *gallop* and is a feature (with the same syntactic relation *sbj*) of some neighbor of *horse*. Suppose that *run*, *move*, and *cycle* are neighbors of *gallop*. As *run* and *move* are also features of some neighbor of *horse* (e.g., *lion*), we would have $N_{horse}(gallop) = \{gallop, run, move\}$. Conversely, since *cycle* is not a feature of a close neighbor of *horse*, it would not be included in the expanded feature set.

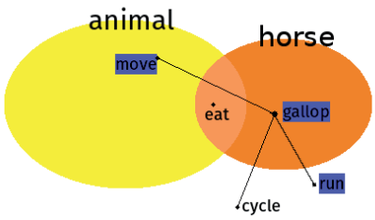

Figure 1: An example of smoothed intersection. Black arrows indicate semantic similarity with *gallop*, items with the blue background are the ones included in $N_{horse}(gallop)$.

Mathematically, we define the expanded feature set F'_x as follows:

$$F'_x = \{(f, N_x(f)) \forall f \in F_x\} \quad (1)$$

$$N_x(f) = \{g | g = (g_w, f_r)\} \quad (2)$$

where the following conditions hold for g:

$$d(f_w, g_w) < k \wedge \exists y | d(x, y) < h \wedge g \in F_y \quad (3)$$

where $d(x, y)$ is any distance measure in the semantic space, k and h are empirically set threshold values.
$N_x(f)$ is generated by looking for features g that are similar to f_w, We then check whether this new feature is shared by some neighbor of the target x, and eventually include g in $N_x(f)$. This allows us to redefine the intersection operation between F'_x and F_y as:

$$F_x{}' \hat{\cap} F_y = \{f | f \in F_x \wedge N_x(f) \cap F_y \neq \emptyset\} \quad (4)$$

When expanding a feature f into $N_x(f)$, we expect to find in $N_x(f)$ features that express the same "property" in different ways. We expect these features to be shared by hypernyms more than co-hyponyms, because hypernyms are supposed to collect features from all their hyponyms, while co-hyponyms lack those of other co-hyponyms (e.g. lions *run* but do not *gallop*). AHyDA is thus defined as follows:

$$AHyDA(x, y) = \frac{\sum_{f \in F_x} |F'_x \cap F_y|}{|F_x|} \quad (5)$$

Importantly, AHyDA only considers the average cardinality of the intersections, without looking at the feature weights. Moreover, the formula is asymmetric (like the others implementing the DIH), and therefore it is suitable to capture the asymmetric nature of hypernymy.

3 Experiments and Evaluation

3.1 Distributional Space

Each lexical item u is represented with distributional features extracted from the *TypeDM* tensor (Baroni and Lenci, 2010). In TypeDM, distributional co-occurrences are represented as a *weighted tuple structure*, a set of $((u, l, v), \sigma)$, such that u and v are lexical items, l is a syntagmatic co-occurrence link between u and v and σ is the *Local Mutual Information* (Evert, 2005) computed on link type frequency. Hence, each lexical item u is represented in terms of features of the kind (l, v).

In addition to the sparse space, we also produced a dense space of 300 dimensions reducing the matrix with Singular Value Decomposition (SVD). This additional space was used to retrieve neighbors during the smoothing operation, as it allowed us to perform faster and more accurate calculations for cosines. The sparse space was instead employed to retrieve features and get their weights.

3.2 Data set

Evaluation was carried on a subset of the BLESS dataset (Baroni and Lenci, 2011), consisting of tuples expressing a relation between nouns.

BLESS includes 200 English concrete nouns as target concepts, equally divided between living and non-living entities. For each concept noun, BLESS includes several relatum words, linked to the concept by one of the following 5 relations: COORD (i.e. co-hyponyms), HYPER (i.e. hypernyms), MERO (i.e. meronyms), ATTRI (i.e. attributes), EVENT (i.e. verbs that define events related to the target). BLESS also includes the relations RANDOM-N, RANDOM-J, RANDOM-V, which relate the targets to control tuples with random noun, adjective and verb relata, respectively.

By restricting to *noun-noun* tuples, we got a subset containing these relations: COORD, HYPER, MERO, RANDOM-N. We preprocessed the dataset in order to exclude lexical items that are not included in TypeDM. As reported in table 2, the distribution (minimum, mean and maximum) of the relata of all BLESS concepts is not even, and therefore we took this into account while

relation	*min*	*avg*	*max*
coord	6	17.1	35
hyper	2	6.7	15
mero	2	14.7	53
ran-n	16	32.9	67

Table 2: Distribution (minimum, mean and maximum) of the relata of all BLESS concepts

evaluating our results.

3.3 Evaluation

We compared AHyDA with a number of directional similarity measures tested on BLESS, with the goal of evaluating their ability to discriminate hypernyms from other semantic relations, in particular co-hyponyms. Given a lexical item x, F_x is the set of its distributional features, $w_x(f)$ is the weight of the feature f for the term x:

WeedsPrec - quantifies the weighted inclusion of the features of a term x within the features of a term y (Weeds and Weir, 2003; Weeds et al., 2004; Kotlerman et al., 2010)

$$\textit{WeedsPrec}(x,y) = \frac{\sum_{f \in F_x \cap F_y} w_x(f)}{\sum_{f \in F_x} w_x(f)} \quad (6)$$

ClarkeDE - a variation of *WeedsPrec*, proposed in (Clarke, 2009)

$$\textit{ClarkeDE}(x,y) = \frac{\sum_{f \in F_x \cap F_y} min(w_x(f), w_y(f))}{\sum_{f \in F_x} w_x(f)} \quad (7)$$

invCL - a new measure introduced in (Lenci and Benotto, 2012), to take into account not only the inclusion of x in y but also the non-inclusion of y in x. The measure is defined as a function of *ClarkeDE* (CD).

$$\textit{invCL}(x,y) = \sqrt{\textit{CD}(x,y)(1 - \textit{CD}(x,y))} \quad (8)$$

We used the **cosine** as a baseline, since it is a symmetric similarity measure and is commonly used to evaluate semantic similarity/relatedness in DSMs. In the definition of $N_x(f)$, the target and feature neighbors are identified with the cosine, setting the k and h parameters to 0.8 and 0.9 respectively.

To avoid biases due to the relata distribution among concepts, for each target x, we computed the *minimum* and *maximum* number of items holding a relation with x, and performed $\frac{maximum}{minimum}$ random samples where each relation is presented with *minimum* relata, and then averaged the results.
For example, consider the situation where x has 3 hypernyms, 6 co-hyponyms, 6 meronyms and 12 random nouns. In this situation, the *minimum* number of relata for x would be 3, while the *maximum* would be 12. Therefore, we would perform 4 random sampling for each relation, averaging the results in order to obtain a singular measurement for each relation in the end.

We adopted the same evaluation methods described in Lenci and Benotto (2012): plotting the distribution of scores per relation across the BLESS concepts, and calculating Average Precision (AP).

3.4 Results

Table 3 summarizes the Average Precision obtained by AHyDA, the other DIH-based measures, and the cosine. Although AHyDA's improvement is not big in hypernym detection, *co-hyponyms* get lower values of AP, thus showing that smoothing the intersection allows a better discrimination between the two classes. It is worth remarking that the values for the other measures are generally higher than those reported by Lenci and Benotto (2012), because of the evaluation on the balanced random samples of relations we have adopted. We also reported, in table 4, the AP values obtained through the standard measures, without employing the feature augementation procedure. Altough values for hypernyms do not change much, the main differences are in the *coord* values, which are generally higher without feature augmentation. As mentioned in section 3.1, the results for all the measures are obtained using the sparse space. The reduced space was employed to compute the *Cosine* baseline.

As regards the AP values for hypernyms, we must notice that not all hypernyms in BLESS share the same status: some of them are what we would consider logic entailments (e.g. *eagle* $\rightarrow$ *bird*), others depict taxonomic relations (e.g. *alligator* $\rightarrow$ *chordate*), some are not true logic entailments (e.g. *hawk* $\stackrel{?}{\rightarrow}$ *predator*)

Figure 2 shows the average score produced with the new measure. Here *hypernyms* are neatly set apart from *co-hyponyms*, whereas the distance with *meronyms* and with the control group, *randoms*, is less significative.

measure	*coord*	*hyper*	*mero*	*ran-n*
Cosine	0.77	0.31	0.21	0.14
WeedsPrec	0.29	0.50	0.32	0.16
ClarkeDE	0.31	0.52	0.24	0.14
invCL	**0.28**	**0.52**	0.32	0.17
AHyDA	**0.20**	**0.49**	0.33	0.23

Table 3: Mean AP values for each semantic relation achieved by AHyDA and the other similarity scores

measure	*coord*	*hyper*	*mero*	*ran-n*
Cosine	0.77	0.32	0.21	0.14
WeedsPrec	0.34	0.51	0.28	0.15
ClarkeDE	0.36	0.51	0.27	0.16
invCL	**0.31**	**0.51**	0.29	0.16

Table 4: Mean AP values for each semantic relation achieved by the cited similarity scores, without employing feature augmentation

Figure 3 shows the average scores produced by AHyDA when applied to the reverse hypernym pair. It is interesting to notice that in this case AHyDA produces basically the same results as random pairs. This suggests that AHYDA correctly predicts that hyponyms entail hypernyms, but not vice versa, thereby capturing the asymmetric nature of hypernymy.

4 Conclusion

The Distributional inclusion hypothesis has proven to be a viable approach to hypernym detection. However, its original formulation rests on an assumption that does not take into consideration the actual usage of hypernyms in texts. In this paper we have shown that, by adding some further pragmatically inspired constraints, a better discrimination can be achieved between co-hyponyms and hypernyms. Our ongoing work focuses on refining the way in which the smoothing is performed, and testing its performance on other datasets of semantic relations.

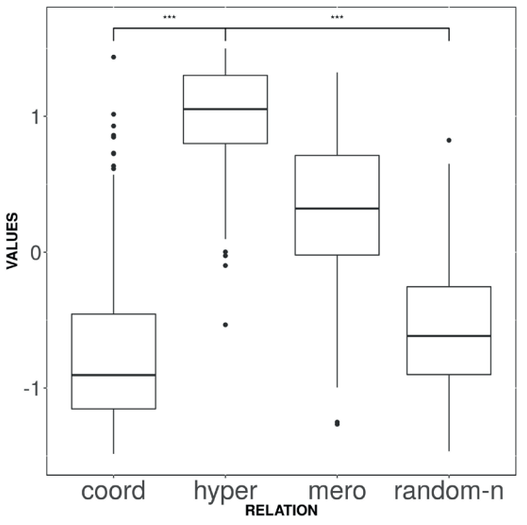

Figure 2: Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores)

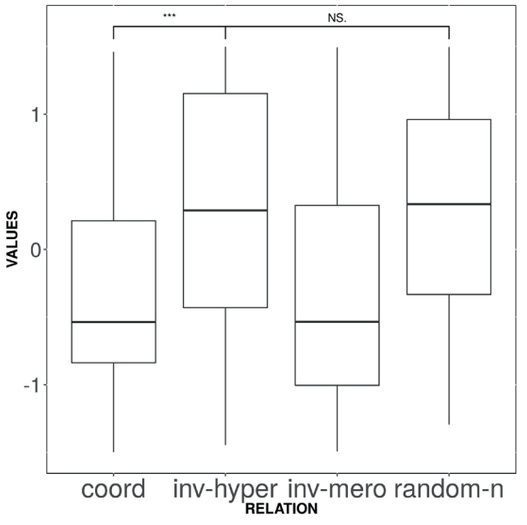

Figure 3: Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores), when tested on the inverse inclusion (i.e. *hypernym* does not entail *hyponym*)

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 *Pro-*

ceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pages 1–10. Association for Computational Linguistics.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 112–119. Association for Computational Linguistics.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of 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*, pages 75–79. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *EACL*, pages 38–42.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88. Association for Computational Linguistics.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics.