IJCoL
Italian Journal of Computational Linguistics

4-2 | 2018
Emerging Topics at the Fourth Italian Conference on Computational Linguistics (Part 2)

Refining the Distributional Inclusion Hypothesis for Unsupervised Hypernym Identification

Ludovica Pannitto, Lavinia Salicchi and Alessandro Lenci

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/ijcol/506
DOI: 10.4000/ijcol.506
ISSN: 2499-4553

Publisher
Accademia University Press

Printed version
Number of pages: 45-55

Electronic reference
Ludovica Pannitto, Lavinia Salicchi and Alessandro Lenci, "Refining the Distributional Inclusion Hypothesis for Unsupervised Hypernym Identification", *IJCoL* [Online], 4-2 | 2018, Online since 01 December 2018, connection on 28 January 2021. URL: http://journals.openedition.org/ijcol/506 ; DOI: https://doi.org/10.4000/ijcol.506

Refining the Distributional Inclusion Hypothesis for Unsupervised Hypernym Identification

Ludovica Pannitto*
Università di Pisa
Università di Trento

Lavinia Salicchi**
Università di Pisa

Alessandro Lenci†
Università di Pisa

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.

1. Introduction and related works

Our intuitions about the meaning of words allow inferences of the kind expressed in example (1) and we want any model of lexical meaning to support such inferences.

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

Words belonging to the same semantic domain are organized into networks of paradigmatic relations such as *synonymy*, *antonymy*, *hypernymy*, *meronymy*, and these are related to the possible inferences that a speaker is able to make when understanding a sentence. The formalization of such relations has been central to both linguistic and computer science research, as they provide a valuable resource for many Natural Language Processing tasks such as *word-sense disambiguation* or *query expansion*. As symbolic models were dominating in linguistics, paradigmatic semantic relations have long been modeled as hierarchies in semantic networks like WordNet (Fellbaum 1998). In computational semantics, several unsupervised methods for the automatic detection of paradigmatic relations have been investigated: here we present a new approach based on a smoothed version of the Distributional Inclusion Hypothesis. This is an extension of the **Distributional Hypothesis**, which claims that lexemes with similar distributional properties have similar meanings. This assumption, which was implicitly introduced in Harris' and Firth's works in the 50s, is the grounding idea of **Distributional Semantics**: semantically similar words could be detected in similar environments, that is similar contexts. Therefore, the semantic similarity between words can be represented in terms

* E-mail: ludovica.pannitto@unitn.it
** E-mail: lavinia.salicchi@libero.it
† E-mail: alessandro.lenci@unipi.it

of their proximity in a semantic space, where the dimensions of the space correspond, at some level of abstraction, to the contexts in which the words occur.

This theoretical framework is computationally implemented in Distributional Semantic Models (DSMs) (Lenci 2018), which build vector spaces from large training corpora to represent linguistic co-occurrences, and so the distributional information, of a given word. To measure the distributional similarity, as an estimate of semantic similarity, several measures have been proposed, the most common one being the cosine.

Our semantic competence licences inferences of the kind expressed in (1), and we expect Distributional Semantic Models (DSMs) to account for such inferences. 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)), differently from most similarity measures used in distributional semantics. Cosine similarity (equation 1), for example, which is one of the most widely employed measures in vector space, quantifies the similarity of two non-zero vectors in terms of the angle between them.

$$cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \left\|\vec{b}\right\|} = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}} \tag{1}$$

As shown in Equation 1, cosine similarity is a symmetric measure, which makes it unsuitable for modelling asymmetric relations such as hypernymy (see example (2)).

(2) a. I saw a *dog* $\rightarrow$ I saw an *animal*
b. I saw an *animal* $\nrightarrow$ I saw a *dog*

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) and Weeds, Weir, and McCarthy (2004) used the Information Retrieval concepts of precision and recall as metrics to identify hypernyms, while Clarke (2009) presented a context-theoretic framework referring to the composition of the meaning of words. These papers introduced measures that quantify the inclusion of a word features among the features of another word, a method which represents the starting point of our proposal. Lenci and Benotto (2012) compared these measures and proposed a variant called invCL (cf. Section 4.3). Another interesting work is Santus et al. (2014), which introduced a new entropy-based measure for the unsupervised identification of hypernym and its directionality in DSMs, starting from the assumption that the typical contexts of a word are more informative than its hypernym contexts. In Weeds et al. (2014) a supervised Machine Learning approach based on linear SVMs has been employed to distinguish co-hyponyms and hypernyms.

The unsupervised hypernym detection task has typically been accomplished relying on the notion of **Distributional Generality** (Weeds, Weir, and McCarthy 2004) and on the **Distributional Inclusion Hypothesis** (DIH) (Geffet and Dagan 2005), which represents its extension and generalization. The intuition is that, since the *hyponym* x is a semantically narrower term than the *hypernym* y, then a number of salient distributional features of x is included in the feature vector of y.

Here we focus on the possibility of identifying hypernyms with directional similarity measures. In the next section we highlight some problems affecting the current versions of the DIH. In Section 3, we introduce AHyDA, a variant of the DIH that aims at

Table 1
Co-occurrence frequency distributions extracted from the ukWaC corpus

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

addressing these problems. In section 4 we describe the experiments in which AHyDA is compared with other state-of-the-art directional measures for hypernym identification.

2. 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. sentences in example (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 $\overset{?}{\rightarrow}$ An *animal* gallops
b. A *dog* barks $\overset{?}{\rightarrow}$ 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 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 actual 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 in the ukWaC corpus *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 $\rightarrow$ An *animal* moves
b. A *dog* barks $\rightarrow$ 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. For example, contexts like *move* and *call*, which could be salient contexts of *animal*, are semantically more general and consequently less informative than *gallop* and *neigh*, typical contexts of *horse*. As entropy can be used to measure informativeness (Shannon 1948), Santus and colleagues used it to propose a new method called SLQS (Equation 2) and defined as the reciprocal difference between the semantic generality E_{w_1} and E_{w_2} of two terms w_1 and w_2. Each E_{w_i} is based on the word most associated contexts entropy. The formula is asymmetric: $SLQS > 0$ if $E_{w_1} < E_{w_2}$, $SLQS < 0$ if $E_{w_1} > E_{w_2}$. So, if $SLQS(w_1, w_2) > 0$, w_1 is semantically less general than w_2. Referring to the previous example, we expect $SLQS(horse, animal)$ to be negative, and $SLQS(animal, horse$ to be positive.

$$SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}} \tag{2}$$

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 domain, 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.

3. AHyDA: Smoothing the DIH

All the measures implementing the DIH are based on computing the (weighted) intersection of the distributional features (i.e., the typical contexts) of the hyponym and the hypernym. This is then typically normalized with respect 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.11 Note that features need not be pure lexical items. In general, we define a feature f as a pair (w, σ) where w is typically a lexical item, and σ is any additional contextual information, in the present case a syntactic pattern occurring between x and w, as explained in section 4.1. The core novelty of AHyDA is to define a set of shared features between the hyponym and the hypernym that, differently from standard set intersection, relies on the expansion of each feature of the hyponym.

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 2, 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 $g = (y, \sigma)$ such that y is a neighbor of *gallop* and it's 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.

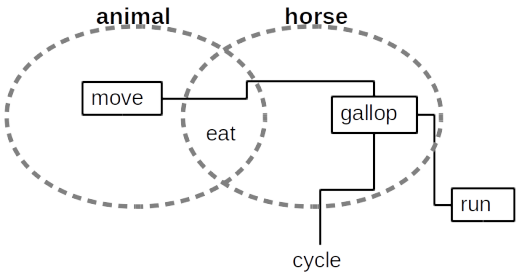

Figure 1
An example of smoothed intersection. Black arrows indicate semantic similarity with *gallop*, countoured items are the ones included in $N_{\text{horse}}(gallop)$: *move* and *run* are included because they are features of a neighbor of *horse* (not shown in the picture), while *cycle* is excluded because no neighbor of *horse* has *cycle* as a feature. Therefore, *gallop* ends up being in the set of shared features of *animal* and *horse*, along with *eat*, thanks to the overlap between the features of *animal* and $N_{\text{horse}}(gallop)$.

Mathematically, for each feature $f = (w_f, \sigma_f)$ in F_x, we define its set of neighbouring features $N_x(f)$ as follows:

$$N_x(f) = \{g = (w_g, \sigma_f) | \exists y.(d\,(w_f, w_g) < \alpha \wedge d\,(x, y) < \beta \wedge g \in F_y)\} \tag{3}$$

where $d(x, y)$ is any distance measure in the semantic space, α and β are empirically set threshold values. Given a feature f, its expanded set $N_x\,(f)$ is therefore generated by looking for features g such that:

- the lexical item w_g is similar, in the semantic space, to the lexeme in the feature f (i.e., w_f);
- $g = (w_g, \sigma_f)$ is a feature of some neighbor of the target x.

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*). $N_x(f)$ is used to define the set of shared features between F_x and F_y as:

$$SharedF(F_x, F_y) = \{f | f \in F_x \wedge N_x\,(f) \cap F_y \neq \emptyset\} \tag{4}$$

While (*gallop*, *sbj*) does not belong to $F_{horse} \cap F_{animal}$ because it is not a feature of *animal*, and does not contribute to any of the measures defined in section 4.3, it is instead included in $SharedF(F_{horse}, F_{animal})$, because there is some feature in $N(gallop, sbj)_{horse}$ that is also included in F_{animal}.

AHyDA is thus defined as follows:

$$AHyDA\,(x,y) = \frac{|SharedF(F_x, F_y)|}{|F_x|} \tag{5}$$

Importantly, AHyDA only considers the average cardinality of the sets, 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.

4. Experiments and Evaluation

4.1 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 $((u, \sigma, v), \kappa)$, such that u and v are lexical items, σ is a syntagmatic co-occurrence link between u and v and κ is the *Local Mutual Information* (Evert 2008) computed on link type frequency. Hence, each lexical item u is represented in terms of features of the kind (v, σ).

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.

4.2 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. Table 2 contains some examples of BLESS tuples for the noun *beaver*.

We preprocessed the dataset in order to exclude lexical items that are not included in TypeDM. As reported in table 3, the distribution (minimum, mean and maximum) of the relata of all BLESS concepts is not even, and therefore we took this into account while evaluating our results.

4.3 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

Table 2
Examples of relata for the target noun *beaver* in BLESS.

coord	*hyper*	*mero*	*random-n*
bear	creature	muzzle	worker
cat	mammal	nose	rose
fox	rodent	tail	foliage

Table 3
Distribution (minimum, mean and maximum) of the relata of all BLESS concepts

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

semantic relations, in particular co-hyponyms.

Given a lexical item x, F_x is the set of its distributional features, $\kappa_x(f)$ 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, Weir, and McCarthy 2004; Kotlerman et al. 2010)

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

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

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

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, moving from the idea that a significant number of the hyponym-contexts are also hypernym-contexts, but a significant number of the hypernym-contexts are not hyponym-context. The measure is defined as a function of *ClarkeDE* (CD).

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

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 α

and β parameters to 0.8 and 0.9 respectively. The optimal settings of the parameter has been identified on a subset of BLESS used as development dataset.

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):

- for each target noun, given its scores agains all its relata in the dataset, which are normalized into *z-scores*, we pick the nearest neighbour of the target for each relation, thus obtaining 4 similarity scores for each BLESS concept. The distribution of the scores is then boxplotted;
- for each target noun, we rank its relata according to their scores agains the target: for every relation, we compute the average precision (AP) of the ranked list: the ideal case (AP = 1) for any relation is the case in which all the relata belonging to that relation are place in the top positions of the ranked list. For each relation, we calculated the AP for all BLESS targets and averaged them.

4.4 Results

Table 4 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. The "delta" between *hyper* and *coord* is an important diagnostic of the model's ability to set apart these two types of relations. Therefore, smoothing the feature 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 5, the AP values obtained through the standard measures, without employing the feature augementation procedure. Although 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 4.1, 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.

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

Table 4
Mean AP values for each semantic relation achieved by AHyDA and the other similarity scores. We evaluated the measures that relied on intersection (i.e., *WeedsPrec*, *ClarkeDE*, *invCL*) considering the features in the set *SharedF* when the formula presented the features in the intersection.

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 5
Mean AP values for each semantic relation achieved by the cited similarity scores, without employing feature augmentation

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

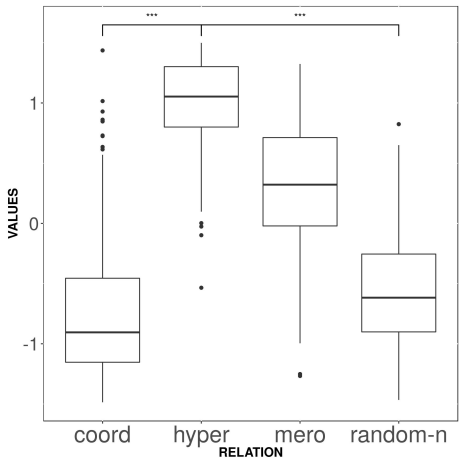

Figure 2
Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores)

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.

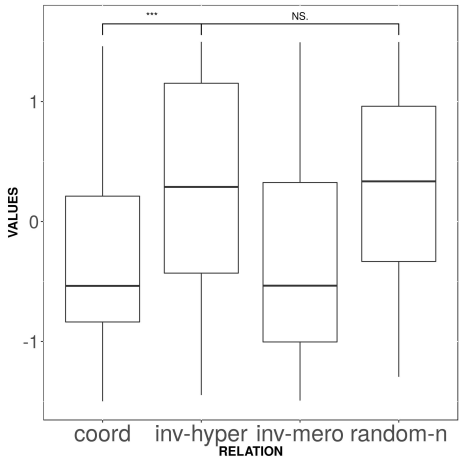

Figure 3
Distribution of relata similarity scores obtained with AHyDA (values are concept-by-concept z-normalized scores), when tested on the inverse inclusion (i.e. *hypernym* does not entail *hyponym*)

5. 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.

References

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

Baroni, Marco and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, Scotland, July 31. Association for Computational Linguistics.

Clarke, Daoud. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 112–119, Athens, Greece, March 31. Association for Computational Linguistics.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, Berlin, pages 1212–1248.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Geffet, Maayan 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, Ann Arbor, Michigan, USA, June 25-30. Association for Computational Linguistics.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics-Volume 2*, pages 539–545, Nantes, France, August 23-28. Association for Computational Linguistics.

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

Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Lenci, Alessandro 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, Montreal, Canada, June 7-8. Association for Computational Linguistics.

Pantel, Patrick 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, Sydney, Australia, July 17-21. Association for Computational Linguistics.

Santus, Enrico, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 38–42, Gothenburg, Sweden, April 26-30.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Weeds, Julie, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259, Dublin, Ireland, August 23-29. Dublin City University and Association for Computational Linguistics.

Weeds, Julie 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, Sapporo, Japan, July 11-12. Association for Computational Linguistics.

Weeds, Julie, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015, Geneva, Switzerland, August 23-27. Association for Computational Linguistics.