AHyDA - CLiC

Automatic Hypernym Detection with feature Augmentation

Ludovica Pannitto, Lavinia Salicchi, Alessandro Lenci
Computational Linguistics Laboratory (CoLing Lab), Università di Pisa

Dec 12th 2017

Inference in Semantic Space, Erk 2009

(1) Google *acquired* YouTube $\rightarrow$ Google *bought* YouTube

(2) A *horse ran* $\rightarrow$ An *animal moved*

- Can we exploit distributional information in order to support lexical inference?
- Is distributional information enough?

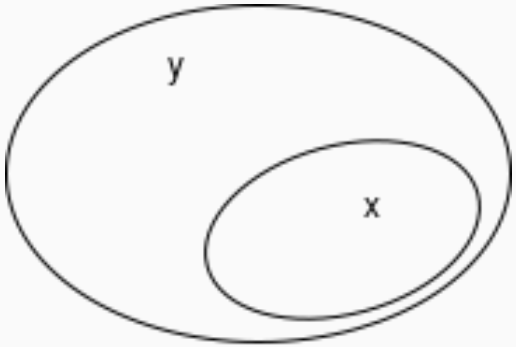
Some inferences are based on **hypernym** - hyponym relation detection

Distributional Inclusion and Directional Similarity Measures

Current Approaches to Hypernym Detection

- Semi-supervised, **pattern-based approaches** (Pantel e Pennacchiotti 2006)
- Fully unsupervised hypernym identification with DSMs. Representation of hypernyms in **vector spaces** usually grounded on the **Distributional Inclusion Hypothesis** (Weeds e Weir 2003, Weeds, Weir e McCarthy 2004, Clarke 2009)
  - **Chasing Hypernyms in Vector Spaces with Entropy**, Santus et al. 2014: Introduction of SLQS, a new entropy-based measure for the unsupervised identification of hypernym and its directionality in Distributional Semantic Models.
- Supervised **Machine Learning** approaches (Weeds et al. 2014, Roller, Erk e Boleda 2014, Kruszewski, Paperno e Baroni 2015, Shwartz, Goldberg e Dagan 2016)

Distributional Inclusion Hypothesis



$$x = \begin{bmatrix} f_1 & w_x(f_1)) \\ \dots & \dots \\ f_k & w_x(f_k) \\ \dots & \dots \end{bmatrix} \sqsubseteq y = \begin{bmatrix} f_1 & w_y(f_1) \\ \dots & \dots \\ f_i & w_y(f_i) \\ \dots & \dots \\ f_k & w_y(f_k) \\ \dots & \dots \end{bmatrix}$$

- **Distributional Inclusion Hypothesis** (Geffet e Dagan 2005):
  - if $x \rightarrow y$ then (most of) the characteristic features of x are expected to appear with y
  - If (most of) the characteristic features of x appear with y then we expect that $x \rightarrow y$

The Pitfalls of the Distributional Inclusion Hypothesis

(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.

The Pitfalls of the Distributional Inclusion Hypothesis

The assumption made by Distributional Inclusion Hypothesis – 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.

	horse	*dog*	*animal*
gallop	216	–	7
bark	–	869	16
	90,437	128,765	161,107

Co-occurrence frequency distribution extracted from the ukWaC corpus. Frequency of lexical items is reported in the last row of the table.

Smoothed Distributional Inclusion Hypotesis

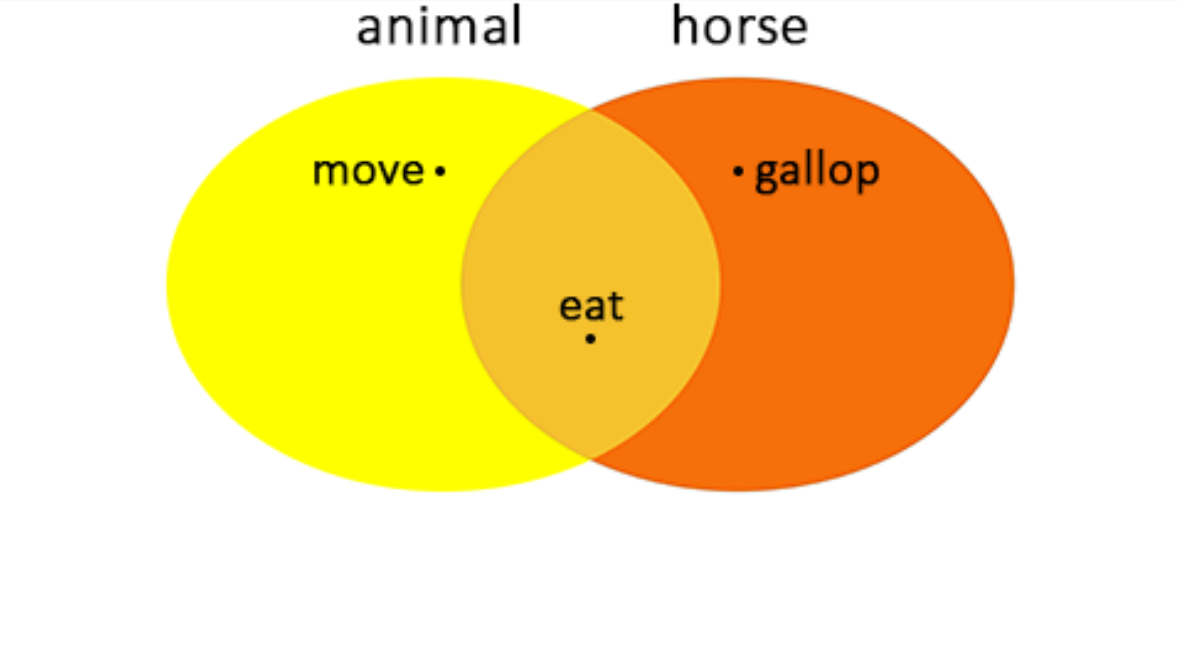
Smoothed Feature Intersection

(4) a. A *horse* gallops $\rightarrow$ An *animal* moves

b. A *dog* barks $\rightarrow$ An *animal* calls

- Salient features of the *hypernym* are supposed to be semantically more general than the salient features of the *hyponym*
- 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.

Smoothed Feature Intersection

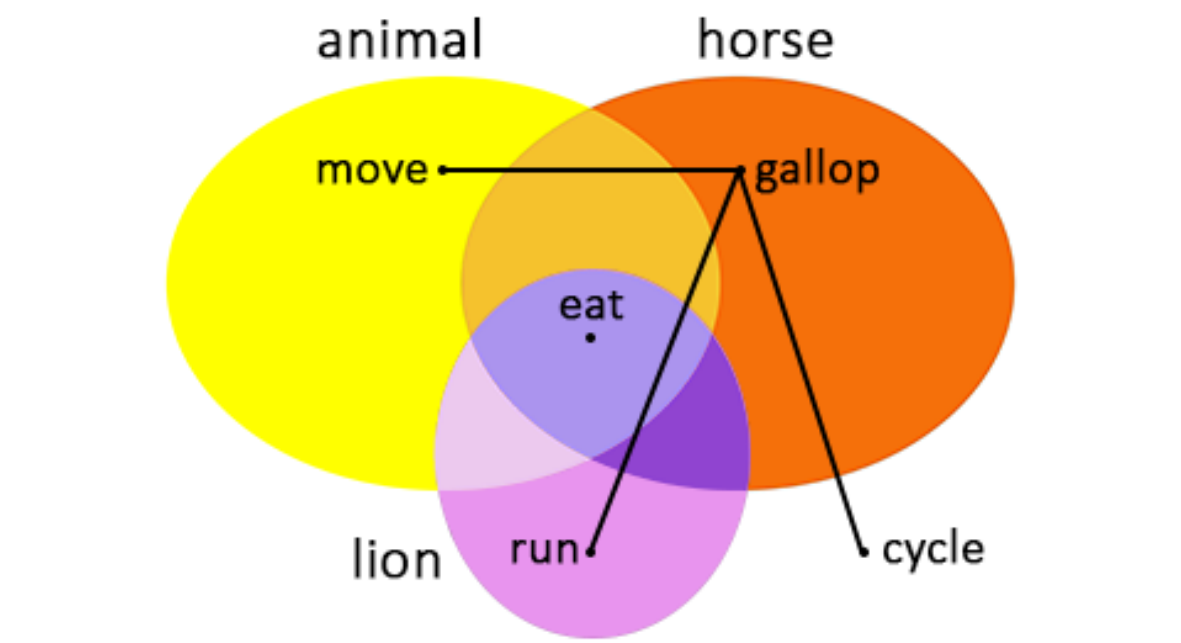


$$animal = \{eat, move, ...\}$$

$$horse = \{eat, gallop...\}$$

$$animal \cap horse = \{eat, ...\}$$

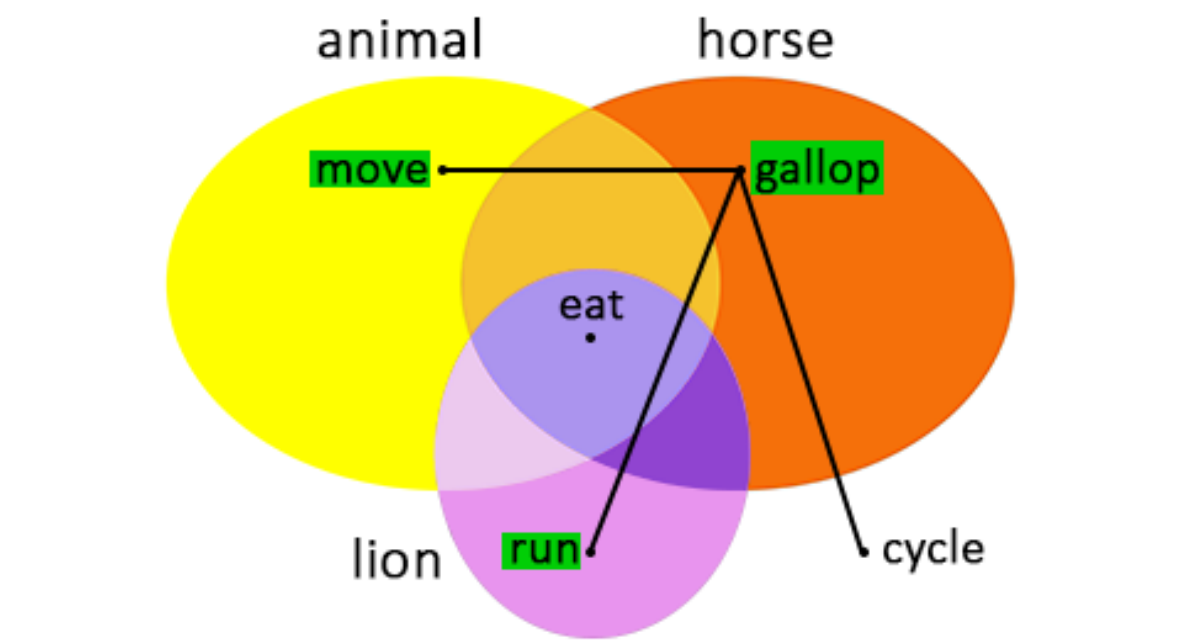
Smoothed Feature Intersection



$$lion = \{eat, run...\}$$

$$N(gallop) = \{move, run, cycle...\}$$

Smoothed Feature Intersection



$$lion = \{eat, run...\}$$

$$N(gallop) = \{move, run, cycle...\}$$

$$N_{horse}(gallop) = \{gallop, run, move, ...\}$$

$$horse' \hat{\cap} animal = \{\textbf{move}, eat, ...\}$$

Fomalizing the Smoothed Feature Intersection

Being F_x the set of features for any lexical item x, we define a smoother version of F_x as follows:

$$F_x{}' = \{(f, N_x(f)) \, \forall f \in F_x\} \tag{1}$$

where $N_x(f)$ is a set of features which are similar to some feature of x, and shared by a lexical item similar to x.

Consequently, we redefine the set intersection operation

$$F_x{}' \hat{\cap} F_y = \{f | f \in F_x \wedge N_x(f) \cap F_y \neq \emptyset\} \tag{2}$$

AHyDA

Employing the smoothed feature intersection, we defined a new measure as follows:

$$AHyDA(x, y) = \frac{\sum_{f \in F_x} |F'_x \cap F_y|}{|F_x|} \tag{3}$$

- AHyDA only considers the average cardinality of the intersection, without looking at the feature weights
- The formula is asymmetric: it is suitable to capture the asymmetric nature of hypernym

Experiments

Distributional Space - Baroni e Lenci 2010

Lexical items represented with distributional feature vectors extracted from TypeDM tensor.

1. Sparse space, where any x is represented by its set of features F_x
   - F_x is a set of pairs composed by a **lexical item** (f_w) and a **syntactic pattern** occurring between x and f_w
   - Employed to retrieve features and their weights
2. Dense space, obtained via SVD (300 dim) from the sparse space
   - Used to retrieve neighbors during the smoothing operation

Dataset BLESS - Baroni e Lenci 2011

- Tuples representing a relation between a **target** concept and a **relatum** concept
- 5 semantic relations are represented + random control elements
- We take into account 3 relations involving nouns (COORD, HYPER, MERO) + RANDOM-N relation

eagle-n

COORD	HYPER	MERO	RANDOM-N
crow	animal	beak	shopping
dove	bird	claw	stuff
falcon	chordate	eye	generation
...	...	...	...

Directional Similarity Measures

WeedsPrec - Weeds and Weir, 2003; Weeds, Weir and McCarthy 2004; Koleman et al. 2010

$$WeedsPrec(x,y) = \frac{\sum_{f \in F_x \cap F_y} w_x(f)}{\sum_{f \in F_x} w_x(f)} \tag{4}$$

ClarkeDE - Clarke, 2009

$$ClarkeDE(x,y) = \frac{\sum_{f \in F_x \cap F_y} min(w_x(f), w_y(f))}{\sum_{f \in F_x} w_x(f)} \tag{5}$$

invCL - Lenci and Benotto, 2012

$$invCL(x,y) = \sqrt{ClarkeDE(x,y)(1 - ClarkeDE(x,y))} \tag{6}$$

Evaluation

- **Method 1 - Boxplots** Given the similarity scores of a target with all its relata, pick the relatum with the highest score for each relation
  Standardize scores for each target (transform into *z-scores*: $x \mapsto \frac{x-\mu}{\sigma}$)
  Summarize the distribution over the dataset
- **Method 2 - Mean Average Precision**:
  Given the ranked sequence of items $x_1, ..., x_N$, it is defined by:

$$AP = \sum_{k=1}^{N} P(k)\Delta_r(k) \tag{7}$$

  where $P(k)$ is the precision evaluated on the sequence $x_1, ..., x_k$ and $\Delta_r(k)$ is the change in recall from step $k-1$ to step k

Evaluation

- The distribution of concepts in BLESS is not uniform.
- 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.

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

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

Results

Preliminary Results - MAP

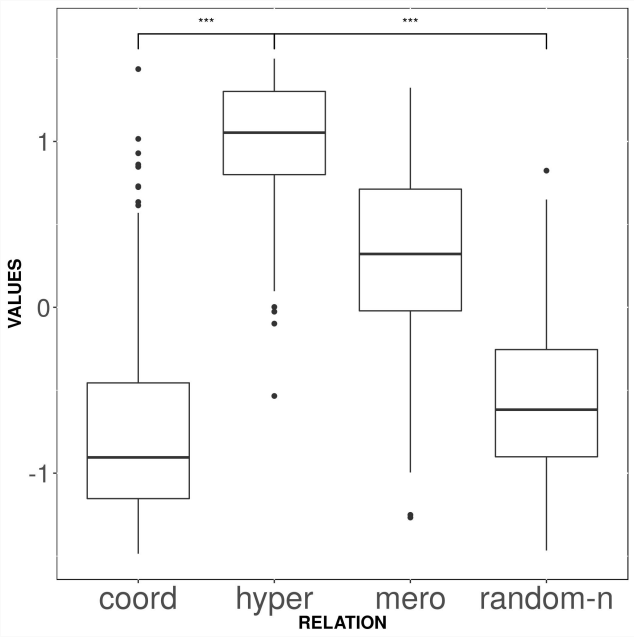
- Mean AP values for each semantic relation achieved by the selected similarity scores, without feature augmentation

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

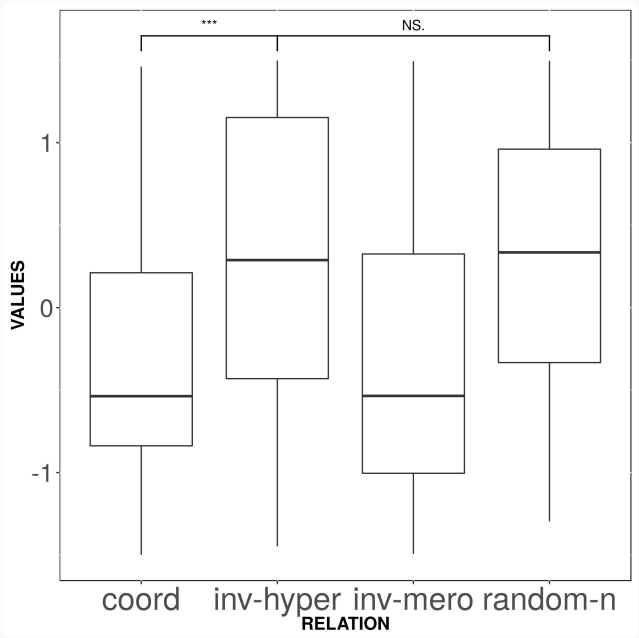
- Mean AP values for each semantic relation achieved by AHyDA and the other similarity scores, **with feature augmentation**

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

Results - Average



Average score produced with AHyDA. Here *hypernyms* are neatly set apart from *co-hyponyms*, whereas the distance with *meronyms* and with the control group, *randoms*, is less significative.



Average scores produced by AHyDA when applied to the reverse hypernym pair. AHyDA produces basically the same results as random pairs, capturing the asymmetric nature of hypernymy.

Closing remarks

Conclusions and Open Issues

- Smoothed Feature Intersection and AHyDA improve the distance between hypernyms and co-hyponyms in the semantic space
- **Not all hypernyms in BLESS share the same status**:
  some are logic entailments (e.g. *eagle* $\rightarrow$ *bird*)
  others depict taxonomic relations (e.g. *alligator* $\rightarrow$ *chordate*)
- Some words are **prototypical hypernyms**, while others are not (Levy et al. 2015).

Ongoing work focuses on refining the way in which the smoothing is performed, and testing its performance on other datasets of semantic relations.

Thank you :)

References

Baroni, Marco e Alessandro Lenci (2010). "Distributional memory: A general framework for corpus-based semantics". In: *Computational Linguistics* 36.4, pp. 673–721.

— (2011). "How we BLESSed distributional semantic evaluation". In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pp. 1–10.

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

Erk, Katrin (2009). "Supporting inferences in semantic space: representing words as regions". In: *Proceedings of the Eighth International Conference on Computational Semantics*. Association for Computational Linguistics, pp. 104–115.

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

Kruszewski, Germán, Denis Paperno e Marco Baroni (2015). "Deriving Boolean structures from distributional vectors". In: *Transactions of the Association for Computational Linguistics* 3, pp. 375–388.

Levy, Omer et al. (2015). "Do Supervised Distributional Methods Really Learn Lexical Inference Relations?" In: *HLT-NAACL*, pp. 970–976.

Pantel, Patrick e 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*. Association for Computational Linguistics, pp. 113–120.

Roller, Stephen, Katrin Erk e Gemma Boleda (2014). "Inclusive yet Selective: Supervised Distributional Hypernymy Detection." In: *COLING*, pp. 1025–1036.

Santus, Enrico et al. (2014). "Chasing Hypernyms in Vector Spaces with Entropy." In: *EACL*, pp. 38–42.

Shwartz, Vered, Yoav Goldberg e Ido Dagan (2016). "Improving hypernymy detection with an integrated path-based and distributional method". In: *arXiv preprint arXiv:1603.06076*.

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

Weeds, Julie, David Weir e Diana McCarthy (2004). "Characterising measures of lexical distributional similarity". In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 1015.

Weeds, Julie et al. (2014). "Learning to distinguish hypernyms and co-hyponyms". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University e Association for Computational Linguistics, pp. 2249–2259.