Squeezing Semantics out of Contexts

Automatic Identification of Semantic Relations in DSMs

Enrico Santus, Alessandro Lenci
Qin Lu and Chu-Ren Huang

The Hong Kong Polytechnic University
香港理工大學

In Supremæ Dignitatis 1343
Università di Pisa

Distributional Semantics and Semantic Relations

- In order to overcome the Knowledge Acquisition Bottleneck
  - Systems for the automatic development & update of knowledge resources have assumed a key role in NLP.
    - **Semantic relations** have been identified as a main building block of such resources (Herger, 2014)

- Distributional semantics is often used at this scope as the *Distributional Hypothesis* (Harris, 1954) puts in relation the syntagmatic and the paradigmatic axes, that is by observing the position of words we can predict their relations (de Saussure, 1916):

  - At least some aspects of word meaning depend on word distribution
  - Words occurring in similar contexts tend to have similar meanings

Under its UMBRELLA…

- If *distributional similarity* allows us to identify paradigmatically related words,
  - its definition is so loose that under its umbrella fall different kinds of paradigmatic relations:
    - hypernyms
    - co-hyponyms
    - antonyms
    - meronyms

So, we suggest that…

1. A careful analysis of word distribution – *possibly inspired by the linguistic and cognitive literature* – allows the identification of properties that are relevant to discriminate semantic relations

2. Not all contexts are equal: the most salient contexts (**PPMI/PLMI**) of the target words are more informative
   - Principle of Cognitive Economy (Collins and Quillians, 1969)
   - Feature Saliency (Smith et al., 1974)

To support our claim…

- We present 3 unsupervised methods:
  - APSyn (Average Precision for Synonymy)
  - APAnt (Average Precision for Antonymy)
  - SLQS: An Entropy-based Measure for Hypernymy

SYNONYMY

MAXIMUM DEGREE OF SIMILARITY

Synonymy and APSyn

- Synonymy: nearly the same meaning
  - Similarity: main organizer of the semantic lexicon (Landauer and Dumais, 1997).

- APSyn (Average Precision for Synonymy)
  - Not only similar words occur in similar contexts, but they tend to share their most relevant contexts in higher proportion than any other relation.

$$APSyn(w_1, w_2) = \sum_{f \in N(F_1) \cap N(F_2)} \frac{1}{(rank_1(f) + rank_2(f))/2}$$

Evaluating APSyn

- Synonym Identification: TOEFL and the ESL test
  - Question word + four choices: algorithm should assign higher score the most similar choice (near-synonym).
    - TOEFL (Test of English as Foreign Language): 80 questions
    - ESL (English as Second Language): 50 questions

- Similarity Estimation: SimLex, WordSim, and MEN
  - Word pairs with similarity score: Spearman Correlation between the algorithm score and the goldstandard one.

(Santus et al., 2016b; 2016d; 2016e)

Synonymy Identification
TOEFL and ESL

- DSM → standard window-based DSM counting co-occurrences of lemmatized content words
  - Windows: 2, 3, 5 and 10
  - Association Measure: PLMI and PPMI
  - Corpora: **ukWaC** and **WaCkypedia** corpora (*2.7B words*; Baroni et al., 2009)

- Dataset → Question-choices transformed into pairs to which APSyn, Cosine and Co-Occurrence scores are assigned
  - TOEFL: Average non-English US college applicant in the TOEFL is 64.5%

ESL on 2-PPMI DSM		
Measure	**N**	**Score**
APSyn	100	73%
Cosine	---	46%
Co-Occurrence	---	43%

TOEFL on 2-PPMI DSM		
Measure	**N**	**Score**
APSyn	100	70%
Cosine	---	58%
Co-Occurrence	----	45%

Similarity Estimation
SimLex, WordSim and MEN

- DSM → standard window-based DSM counting co-occurrences of lemmatized content words
  - Windows: 2, 3
  - Association Measure: Frequency, PLMI and PPMI
  - Corpora: **RCV1** (150M) and **Wikipedia** (820M)

- Dataset → Calculate the Spearman correlation between APSyn scores and gold standard.
  - State-of-the-art results are taken from Hill et al. (2015)

Similarity Estimation on RCV1 (win3)				
Measure	**Matrix/N**	**SimLex**	**WordSim**	**MEN**
APSyn	**500**	**0.32**	**0.468**	**0.478**
Cosine	**SVD-PPMI-300**	**0.382**	**0.47**	**0.538**
Cosine	**SVD-PLMI-300**	0.21	0.29	0.286
Mikolov et al.	---	*0.282*	*0.442*	*0.433*

Similarity Estimation on Wikipedia (win3)				
Measure	**Matrix/N**	**SimLex**	**WordSim**	**MEN**
APSyn	**500**	**0.423**	**0.653**	**0.773**
Cosine	**SVD-PPMI-300**	**0.464**	**0.562**	**0.779**
Cosine	**SVD-PLMI-300**	0.329	0.408	0.563
Huang et al.	---	*0.098*	*0.30*	*0.433*
Coll & West	---	*0.268*	*0.494*	*0.575*
Mikolov et al.	---	*0.414*	*0.655*	*0.699*

Discussion

- TOEFL and ESL:
  - No SoA (100% on TOEFL and 82% on ESL) but simple measure that largely outperforms:
    - Random Baseline, Vector Cosine and Co-occurrence on the same DSM
    - US non-English college applicants (i.e. 64.5% on the TOEFL)

- SimLex, WordSim and MEN:
  - Outperforms the Vector Cosine in all settings except SVD-PPMI-300 (best setting)
  - Outperforms the performance of word embedding models, as reported by Hill et al. (2015)

- Observations:
  - **Scalability concerns**: using the same corpus of Baroni et al. (2014), we obtain:
    - **0.72** on WordSim (vs. 0.62 counting and 0.75 predicting)
    - **0.77** on MEN (vs. 0.72 counting and 0.80 predicting).
  - ***N*** is relatively stable till N=1000
    - Performance drops for higher values
      - The most relevant contexts are in fact informative!

ANTONYMY

WHEN SIMILARITY BECOMES OPPOSITION

Antonymy and APAnt

- Antonymy: opposition
  - Distributionally similar to synonyms ☹
  - Cruse (1986) → *paradox of simultaneous similarity and difference*:
    - Antonyms are similar in every dimension of meaning except one
      - e.g. size → dwarf and giant

- APAnt (Average Precision for Antonymy)
  - If perfect synonyms → identical distribution
  - Antonyms → similar distribution except for some dimensions
    - ASSUMPTION: These dimensions are the most relevant ones (i.e. size for dwarf and giant)

- Easiest way to test it: inverse of APSyn
  - 1/Vector Cosine = Non-similarity
  - 1/APSyn = Non-sharing the most relevant contexts

$$APAnt(w_1, w_2) = \frac{1}{APSyn(w_1, w_2)}$$

$$APSyn(w_1, w_2) = \sum_{f \in N(F_1) \cap N(F_2)} \frac{1}{(rank_1(f) + rank_2(f))/2}$$

Evaluating APAnt

- Antonym retrieval task:
  - Given a word, find its antonym(s) among several related words
    - antonyms, synonyms, hypernyms and co-hyponyms

- Datasets (word-pairs labeled with SemRel):
  - Lenci/Benotto (Benotto, 2015)
  - EVALution (Santus et al., 2015)
  - BLESS (Baroni and Lenci, 2011)

Results

- DSM:
  - Windows: 2 and 5
  - Association Measure: Frequency, PLMI and PPMI
  - Corpora: **ukWaC** and **WaCkypedia** corpora (*2.7B words*; Baroni et al., 2009)

- Baselines: Vector Cosine and Co-Occurrence Frequency
  - Co-occurrence hypothesis (Charles and Miller, 1989): antonyms occur in the same sentence more often than by chance.

- Task:
  - Assign scores to all pairs
  - Sort pairs decreasingly by first word and score
  - Calculate MAP to evaluate the ranking (Kotlerman et al., 2010)

Antonymy Discrimination (win 5)					
Measure	**Matrix/N**	**Antonyms**	**Synonym**	**Hypernyms**	**Co-Hyponyms**
Number of Pairs	**---**	***2545***	***2190***	***4261***	***3231***
APAnt	**250**	**0.28**	**0.18**	**0.43**	**0.17**
Cosine	**PPMI**	**0.20**	**0.20**	**0.31**	**0.29**
Co-occurrence	**Frequency**	0.23	0.19	0.36	0.23

Discussion

- APAnt outperforms **vector cosine** and **co-occurrence** baseline in discriminating antonyms from synonyms (they have a slight preference for synonyms)
  - True → NOUN (best), VERBS and ADJ (worst)
  - True → when Co-Hypo and Hyper are involved
    - A preference for hypernyms has also been noticed and this might mean that also hypernyms do not share their most relevant contexts.

HYPERNYMY

SIMILAR, BUT MORE GENERAL

Hypernymy and SLQS

- Hypernymy: Is-A relationship
  - Properties: Similarity and generality
  - Fundamental relationship to structure taxonomies (Cruse, 1986)

- The lower the position in the taxonomy, the more informative a word is (Murphy, 2002)

- *Distributional Informativeness Hypothesis*: the semantic generality of a word can be inferred by the informativeness of its most related contexts

  _____ barks → dog

  _____ eats → ???

- SLQS: Inversed ratio between the *median entropy of the top N most related contexts of W1 and W2*

$$H(c) = -\sum_{i=1}^{n} p(f_i|c) \cdot log_2(p(f_i|c)) \qquad E_{w_i} = Me_{j=1}^{N}\left(H_n(c_j)\right) \qquad SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}}$$

Evaluating SLQS

- Dataset: BLESS (Baroni and Lenci, 2011)

- Two tasks:
  - Directionality identification (1277 hyper pairs)
    - 87% versus:
      - 63.04% of WeedsPrec (Distributional Inclusion Hypothesis)
      - 66.09% of Frequency (more general = more frequent)

  - Hypernymy discrimination (1277 hyper, coord, mero, random)
    - Assign the scores, sort, and calculate AP (Kotlerman et al., 2010)

Hypernymy Discrimination (win 2)					
Measure	**Matrix/N**	**Hyper**	**Coord**	**Mero**	**Random**
SLQS * Cosine	**50**	0.59	0.27	0.35	0.24
Frequency Baseline	**Frequency**	0.40	0.51	0.38	0.17
WeedsPrec	**Frequency**	0.50	0.35	0.39	0.21
Cosine	**Frequency**	0.48	0.46	0.31	0.21

Discussion

- SLQS outperforms WeedsPrec, Cosine and Frequency baselines (competitive).
  - Better vs. symmetric relations (i.e. coordinates)
  - Worse vs. meronyms and randoms

- In a recent study (Shwartz et al., 2016):
  - Despite lower performance than supervised methods, it is stable in transfer learning

ROOT9

- Classifier: RandomForest, SMO and Logistic
- Datasets: EVALution, Weeds
- Features: only 9 features
  - **Frequency**: W1, W2, TopNContexts W1, TopNContexts W2
  - **Entropy**: W1, W2, TopNContexts W1, TopNContexts W2
  - **APSyn**

- Results:
  - ROOT9 is very competitive with the supervised models presented by Weeds et al. (2014)
    - Outperformed on all datasets only by the SVNcat

- **Do these 9 distributional properties summarize 1000 features?**

Conclusions

- Distributional Semantics should get inspired by the linguistic and cognitive theory – and, in turn, it should support it by providing evidence.

- The proposed measures are not meant to be closed boxes, but they aim at providing insights and stimulate further speculations:
  - There seems to be room for improvement.
  - The top *N* contexts are informative

- In the future, there might be the need to focus on the properties of subsets of contexts rather than on the full word distribution

Thank you

Enrico Santus – The Hong Kong Polytechnic University
Alessandro Lenci – University of Pisa
Qin Lu – The Hong Kong Polytechnic University
Chu-Ren Huang – The Hong Kong Polytechnic University

Co-authors:
Emmanuele Chersoni, Sabine Schulte im Walde, Dominik Schlechtweg,
Vered Shwartz and Frances Yung