Derivational Smoothing for Syntactic Distributional Semantics

Sebastian Padó*, Jan Šnajder†, and Britta Zeller*

*Institute for Computational Linguistics, Heidelberg University
†Faculty of Electrical Engineering and Computing, Zagreb University

The 51st Annual Meeting of the
Association for Computational Linguistics
August 6, 2013

EXCITEMENT

hrzz
Hrvatska zaklada za znanost

Distributional Semantics

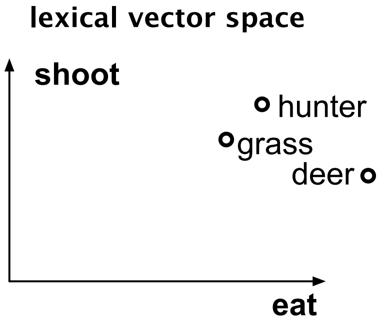
- Representation of word meaning as vectors
  - Vector components: co-occurrences with context features
  - Firth (1957): *You shall know a word by the company it keeps*

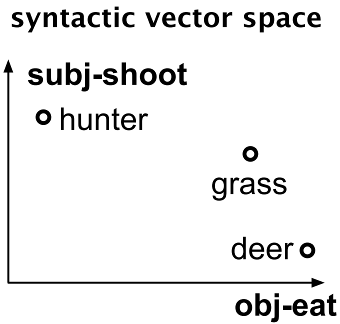
Peter convinced himself to write reports $\Rightarrow$

	report
Peter	1
convince	1
write	1

- Vector similarity approximates semantic similarity
  - Simple, unsupervised induction of word meaning
  - Used in variety of tasks (Turney and Pantel, 2010)

Main Context Choices

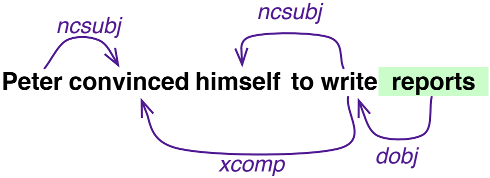




- Lexical (word) context captures topical similarity
- Syntactic (word-relation) context captures relational similarity
  - Can model fine-grained information (Baroni and Lenci, 2010)
  - More appropriate for free word order languages

A problem for syntactic vector spaces: Sparsity

- Syntactic vector spaces are very sparse
  - Even if constructed from very large corpora
- Reason: Less cooccurrences



$\Rightarrow$

	report
write	1

- Many word pairs receive semantic similarities of zero
  - Real dissimilarity or missing data?

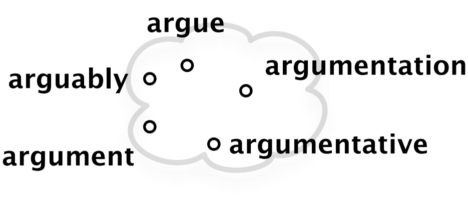
Derivation Smoothing

The question

Where can we get semantic relatedness information to smooth distributional similarity?

The answer: Derivational morphology

- Consider derivational families:



- Words that are derived from one another have similar meaning
- Available from resources like CatVar (Habash and Dorr, 2003)

Derivational Smoothing

- If vectors are sparse, do not compute semantic similarity directly
- Instead, back off to less sparse members of derivational families

sim(arguably, debatably) = 0

sim(argue, debate) > 0

$\big\downarrow$ **back-off**

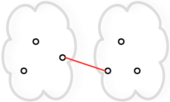
smoothed-sim(arguably, debatably) =

f(arguably , debatably)

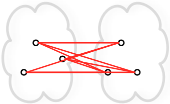
- (Similar to back-off to less sparse $n - 1$ grams in LMs)

Derivational parameters: Two parameters

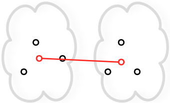
1. Smoothing trigger: When is a vector considered too sparse?
   - Smooth always
   - Smooth only if $sim(l_1, l_2) = 0$ (or undefined)
2. Smoothing scheme: How to bring in derivational family
   - maxSim: Consider most similar pair between families



   - avgSim: Consider average similarity of all pairs



   - centSim: Consider similarity of family centroids



Experiments

Language choice: German

- Resource situation comparable to English, but not quite as good
- Derivation important process of word formation

Distributional models

- Base Model: German Distributional Memory DM.DE (Padó and Utt, 2012)
  - 900M-token SDEWAC web corpus (Faaß *et al.*, 2010)
- DERIVBASE derivational families (Zeller *et al.*, 2013)
  - Rule-based resource for German, focus on precision
  - 18.000 non-singleton families covering 60.000 lemmas
- Baseline: Bag-of-words models (same corpus)

Evaluation

Task 1: Synonym choice

- 980 targets with four candidates each (Reader's Digest)
  "Which term is *antiquated* most similar to?
  (a) venerable, (b) old, (c) unusable, (d) outdated?"
- Prediction: candidate with max cosine similarity to target
- Evaluation: Accuracy (%) + Coverage (%)

Task 2: Word similarity prediction

- 350 pairwise judgments on 5-point scale (Zesch *et al.*, 2007)
  (*monkey, macaque*) $\Rightarrow$ 4
  (*office, tiger*) $\Rightarrow$ 1
- Prediction: Cosine similarity
- Evaluation: Correlation (Pearson's r) + Coverage (%)

Results: Synonym choice

Model		Acc. %	Cov. %
DM.DE, unsmoothed		53.7	80.8
DM.DE, smooth always	avgSim	46.0	86.6
	maxSim	50.3	86.6
	centSim	49.1	86.6
DM.DE, smooth if *sim* = 0	avgSim	52.6	86.6
	maxSim	51.2	86.6
	centSim	51.3	86.6
BOW "baseline"		**56.9**	**98.5**

- Gain in coverage (+6%), but small loss in accuracy (-1%)
    - BoW "baseline" performs best
- Conservative trigger (smooth if necessary) works best

Results: Semantic similarity

Model		r	Cov. %
DM.DE, unsmoothed		.44	58.9
DM.DE, smooth always	avgSim	.30	88.0
	maxSim	.43	88.0
	centSim	.44	88.0
DM.DE, smooth if $sim = 0$	avgSim	.43	88.0
	maxSim	.42	88.0
	centSim	**.47**	88.0
BOW baseline		.36	**94.9**

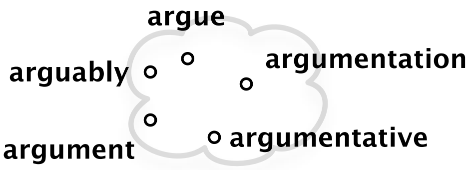
- Again, conservative trigger works best
- Big increase in coverage (+30%), small increase in correlation

Task Comparison

Result change through smoothing

Task	Quality	Coverage
Synonym choice	-0.09 % Acc.	$+6\%$
Semantic similarity	$+0.03$ Corr.	$+30\%$

- Semantic similarity benefits more from derivational smoothing than synonym choice
  - Derivational families contain *related words*, not *synonyms*



Summary

- Sparsity is a problem for syntax-based distributional models
  - "Derivational smoothing": Back off from rare word to derivational family
- Initial experiments
  - Conservative trigger (smooth only when sim=0) works best
  - Jury still out on smoothing scheme (combination method)
- Future work
  - More experiments on smoothing schemes
  - Use richer information about derivational families

References I

Baroni, M. and Lenci, A. (2010). Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, **36**(4).

Faaß, G., Heid, U., and Schmid, H. (2010). Design and application of a gold standard for morphological analysis: SMOR as an example of morphological evaluation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press.

Habash, N. and Dorr, B. (2003). A categorial variation database for English. In *Proceedings of the NAACL/HLT*, pages 17–23.

Padó, S. and Utt, J. (2012). A distributional memory for German. In *Proceedings of KONVENS*, Vienna, Austria.

References II

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, **37**(1), 141–188.

Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL*, Sofia, Bulgaria.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of NAACL/HLT*, pages 205–208.