Resolving PP-attachment ambiguity by distributional semantic modeling in the context of parsing of French

Presentation of the master thesis research

work supervised by dr. C. Cerisara (LORIA)

Simon Šuster, RUG

s.suster@rug.nl

September 7, 2012

Outline

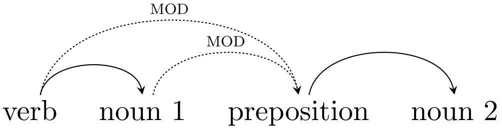
- Syntactic parsing
- PP-attachment (PPA) disambiguation
- Distributional semantic modeling
- Experiment and results
- Integration with parsing
- Conclusion

(Dependency) Parsing

- Parsing: strict parsing and disambiguation
- Disambiguation can be done by probabilistic modeling to select the most plausible parse
- Dependency parsing
  - intuitive predicate-argument representation
  - (suitable for languages with less fixed word order)
  - accurate results for many languages [Kübler et al., 2009]
- Data-driven vs. grammar-driven parsing (or something in between)
- Data-driven supervised parsers select an optimal parse given the model learnt from treebanks and the sentence
- Graph-based (MST, **MATE**) vs. transition-based dependency parsers (Malt)
- A parser performs differently well on different structural problems

PP-attachment disambiguation

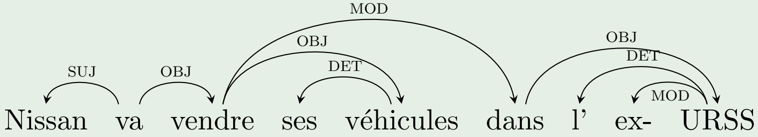
- Notoriously difficult, but much researched
- Factors: lexical preferences, subcategorization frames, fixed phrases, semantic and pragmatic knowledge
- The problem of choosing the right attachment site:



(a simplification with only two competing sites)

Example

(French Treebank)



A brief overview of the PPA-disambiguation research

- Co-occurence strength [Hindle and Rooth, 1993]: the importance of the preposition
- Supervised learning on a PPA-dataset [Ratnaparkhi et al., 1994]: isolated view
- Inclusion of semantic information:
  - mapping WordNet concepts onto 4-tuples (88.1% acc.) [Stetina and Nagao, 1997]
  - parser training on semantic classes [Agirre et al., 2008]
  - nearest-neighbours with distr. sim. between 4-tuples [Zhao and Lin, 2004]
- PPA disambiguation in the context of parsing [Atterer and Schütze, 2007]: situated view
  - retrieve PPA cases based on parser's output
  - evaluate attacher against the parser
- **French** Feature-rich parsing correction [Henestroza and Candito, 2011]: no semantic information

Distributional Semantic Models (DSM)

- Comp. models using distributional patterns to derive representations of meaning of ling. units
- Spatial proximity = semantic similarity
- Our distributional hypothesis:
  words with similar distributional properties have similar meanings

Distributional Semantic Models (DSM)

- Comp. models using distributional patterns to derive representations of meaning of ling. units
- Spatial proximity = semantic similarity
- Our distributional hypothesis:
  words with similar distributional properties have similar meanings
- DSMs are implemented as matrices, parametrized through [Evert and Lenci, 2009, Turney and Pantel, 2010]:
  - target elements (rows)
  - contexts (dimensions)
  - relation between targets and contexts
  - weights for matrix values
  - dimensionality reduction
  - distance measures between vectors

Experiment data

- French Treebank (12k sent.) for testing, in CoNLL format
  - extracted 3398 PPA instances not including the preposition "de"
- Gigaword French corpus (36m sent.) for model construction
- MATE parser [Bohnet, 2010]
- Gold and parser statistics on the PPA (French Treebank):

Total sentences	120
PPA per sentence	1 per 1.39
verbal/nominal att. ratio	0.44
verbal/nominal att. ratio, "de"-only	0.054
verbal/nominal att. ratio, non-"de"	**0.786**
Parser ER	0.19
Parser ER, "de"-only	0.054
Parser ER, non-"de"	**0.31**

Experiment description

- Skewed class distribution
- Disambiguation as detection: a true positive is a correct nominal detection above some threshold
- DSM-obtained information (ratio) is seen as a confidence measure in determining the attachment site
- Detection is done on the cases *retrieved* as ambiguous
    - PPA case: construction V N1 P N2, where N1 is the direct object of V, and where P is not "de"
    - POS-, dependency- and lexicon-driven retrieval
    - precision: $0.886 \pm 0.057_{95\%\,CI}$; recall: $0.738 \pm 0.079_{95\%\,CI}$

Exp 1 DSM-based detection

Exp 1b Integration of DSM-based detection into parsing

DSM-based disambiguation

Intuition *Semantic similarity between elements in an ambiguous case indicates the attachment site*

- The more N1 and N2 (or N1 and entire PP) are semantically similar, and the less V and N2 (or V and entire PP) are semantically similar, the more likely the nominal attachment

Example

"eat salad with croutons"
→ *salad* & *croutons* semantically more similar than *eat* & *croutons*

"eat salad with fork"
→ the opposite is true

DSM-based disambiguation

Intuition *Semantic similarity between elements in an ambiguous case indicates the attachment site*

- The more N1 and N2 (or N1 and entire PP) are semantically similar, and the less V and N2 (or V and entire PP) are semantically similar, the more likely the nominal attachment

- Decision based on V and N2 compared to N1 and N2:

$$A = nom \text{ if } \frac{Cos(n1, n2)}{Cos(v, n2)} > \delta \tag{1}$$

- Decision based on V and PP compared to N1 and PP (PP is composed P and N2) (cf. [Mitchell and Lapata, 2008]):

$$A = nom \text{ if } \frac{Cos(n1, f(p, n2))}{Cos(v, f(p, n2))} > \delta, \text{ where } f \in \{add., mult.\} \tag{2}$$

DSM-based disambiguation

Parameters of our DSMs

- 2,816 by 10,000 matrix from the 447M-word Gigaword
- Rows are lemmas from the test 4-tuples and dimensions are 10,000 most frequent non-function words (lemmas)
- Relation: window of max. -3+3 words
- Weights: log-frequency, PMI, positive-PMI, local-PMI
- Dimensionality reduction: by constraints on the number of rows/dimensions (74% zero elements); SVD to 300 dimensions (92% of the variance)
- Similarity metric: Cosine

DSM-based disambiguation

Main findings

- Signif. better with PPMI (and LPMI) than plain PMI, log-freq. or plain freq.
- SVD to 300 dim. improves results significantly
- Composition by the P+N2 addition yields better results than N2-only semantic representation
- Adding P to both V/N1 and N2 yields even superior results
- Multiplication results worse than the baseline of always choosing V att.

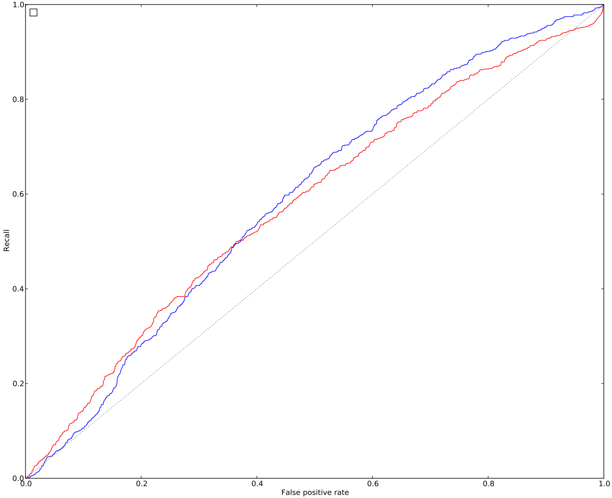


Figure: ROC curve for DSM-based detection (PPMI, 300-dim. SVD), addition of P+N2 (red), and addition of P+N2 and V/N1+P (blue). Difference between both: d=-0.011, p=0; compared to baseline (dotted line): $d_{red} = 0.08$, AUC=0.58, p=0; $d_{blue} = 0.091$, AUC=0.591, p=0

Integration with parsing I

- MATE parser baseline UAS 86.93%
- *Constrained* parsing on the preannotated dependencies (att. decisions) leads to an optimal result (not true for post-festum approaches like parsing correction)
- Certain thresholds lead to an improvement, but impact small

Threshold	Deps pre-annotated	UAS
0.0462	62	0.863
0.5326	62	0.865
1.0190	62	0.868
1.5055	62	0.871
1.9919	62	**0.8726**
2.4784	62	**0.8726**
2.9648	62	0.8722

Table: Parsing improvement with a DSM-driven detector for PP-attachment on a 200-sent. test corpus

Integration with parsing II

- With 2 separate thresholds for nominal and verbal att.
- Cosine ratio as a kind of confidence measure: we can keep only the most reliable dependencies
- $A_{nominal}$ if $ratio_{Cos} > \delta_{nominal}$, A_{verbal} if $ratio_{Cos} < \delta_{verbal}$

	δ_{ver}	δ_{nom}	N. of attachment cases	Correct att. by the parser	Correct att. by the DSM-driven detector
	1.078217	1.7003012	44	31	33
	1.078217	3.2809374	39	26	32
	1.078217	1.3897803	46	32	34
	1.078217	1.9061964	43	30	33
	0.9382211	2.546115	36	23	27
	0.9382211	1.5775068	39	25	28
	0.9382211	2.680369	34	21	26
	0.9382211	1.3190644	41	26	28
	⋮	⋮	⋮	⋮	⋮
Avg. accuracy				0.69	0.769

Conclusion I

- DSM-driven PP-disambiguation in the context of parsing
- Encouraging results: semantic information leads to a small improvement
- Most useful when we have very high semantic similarity between two elements on the one hand, and very low similarity of the competing two elements on the other hand
- May prove powerful for semi-supervised approaches that need additional information not already modeled by the parser:
  1. Establish dependency relations by DSM
  2. Constrained parsing; we obtain text that is parsed entirely
  3. Retraining of the parser
- Integration with parsing promising for future research (preferably expanding the problem to other types of structural ambiguity)

Conclusion II

- The question still remains: what types of structural ambiguity with what kind of semantic information?
- Explore semantic composition in more detail: use external criteria in deciding which elements to compose (e.g. sub-categorization frames) instead of a naive composition
- Tensors as multiway objects are an alternative for combining more than two vector representations [van de Cruys, 2010]

Conclusion III

- Specific syntactic ambiguity, specific language with specific occurrence rates, specific parser (performance)
- For Dutch, an attempt by [van Herwijnen et al., 2003], using memory-based learning with lexical and cooccurrence-strength features: an isolated perspective
- Is there space for improvement of Alpino on this particular problem (or other syntactic ambiguities)?
  - An error analysis is needed
  - General occurrence statistics of the problem for Dutch

Bibliography I

Agirre, E., Baldwin, T., and Martínez, D. (2008).
Improving parsing and pp attachment performance with sense information.
In McKeown, K., Moore, J. D., Teufel, S., Allan, J., and Furui, S., editors, *ACL*, pages 317–325. The Association for Computer Linguistics.

Atterer, M. and Schütze, H. (2007).
Prepositional phrase attachment without oracles.
Computational Linguistics, 33(4):469–476.

Bohnet, B. (2010).
Top accuracy and fast dependency parsing is not a contradiction.
In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.

Evert, S. and Lenci, A. (2009).
Foundations of distributional semantic models.
Tutorial at ESSLLI 2009, Bordeaux.

Bibliography II

Henestroza, E. and Candito, M. (2011).
Parse correction with specialized models for difficult attachment types.
In *EMNLP*, pages 1222–1233. ACL.

Hindle, D. and Rooth, M. (1993).
Structural ambiguity and lexical relations.
Computational Linguistics, 19:103–120.

Kübler, S., McDonald, R. T., and Nivre, J. (2009).
Dependency Parsing.
Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Mitchell, J. and Lapata, M. (2008).
Vector-based models of semantic composition.
In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Bibliography III

Ratnaparkhi, A., Reynar, J., and Roukos, S. (1994).
A maximum entropy model for prepositional phrase attachment.
In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 250–255, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stetina, J. and Nagao, M. (1997).
Corpus based PP attachment ambiguity resolution with a semantic dictionary.
In Zhou, J. and Church, K. W., editors, *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 66–80, Beijing, China. ACL.

Turney, P. D. and Pantel, P. (2010).
From frequency to meaning: vector space models of semantics.
Journal of Artificial Intelligence Research, 37:141–188.

van de Cruys, T. (2010).
Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text.
Groningen dissertations in linguistics.

Bibliography IV

van Herwijnen, O., Terken, J., van den Bosch, A., and Marsi, E. (2003).
Learning pp attachment for filtering prosodic phrasing.
In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 139–146, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhao, S. and Lin, D. (2004).
A nearest-neighbor method for resolving pp-attachment ambiguity.
In *IJCNLP*, pages 545–554.