Improving PP attachment in a hybrid dependency parser using semantic, distributional, and lexical resources

University of Zurich[UZH]

EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

COST PARSEME WG 3, Athens

Gerold Schneider
gschneid@es.uzh.ch

Introduction

ATTACHMENT OF PREPOSITIONAL PHRASES is a major source of ambiguity for parsers. Noun-PP and particularly verb-PP relations are multi-word constructions for which considerable amounts of resources exist. Our dependency parser (Schneider, 2008) is hybrid:

- It uses a hand-written *competence* grammar and statistical *performance* disambiguation learnt from the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993)
- Maximum Likelihood Estimation (MLE) probability model for the tri- and bi-lexical performance disambiguation estimates the probability of the dependency relation R at distance (in chunks) $dist$, given the lexical head of the governor (a) and the lexical head of the dependent (b) and description noun in PPs (c):

$$p(R, dist|a, b, c) \cong \frac{f(R, a, b, c)}{f((\sum R), a, b, c)} \cdot \frac{f(R, dist)}{fR} \tag{1}$$

- Sparse data: back-off architecture similar to (Collins and Brooks, 1995), but extending from PP-attachment to most of its dependency relations, and including simple semantic classes from WordNet (Miller et al., 1990)

The parser's label set is close to and can be mapped to GREVAL (Carroll, Minnen, and Briscoe, 2003) and the Stanford scheme (Haverinen et al., 2008).
We improve its performance with the following statistical multi-word resources.

Multi-Word Terminology

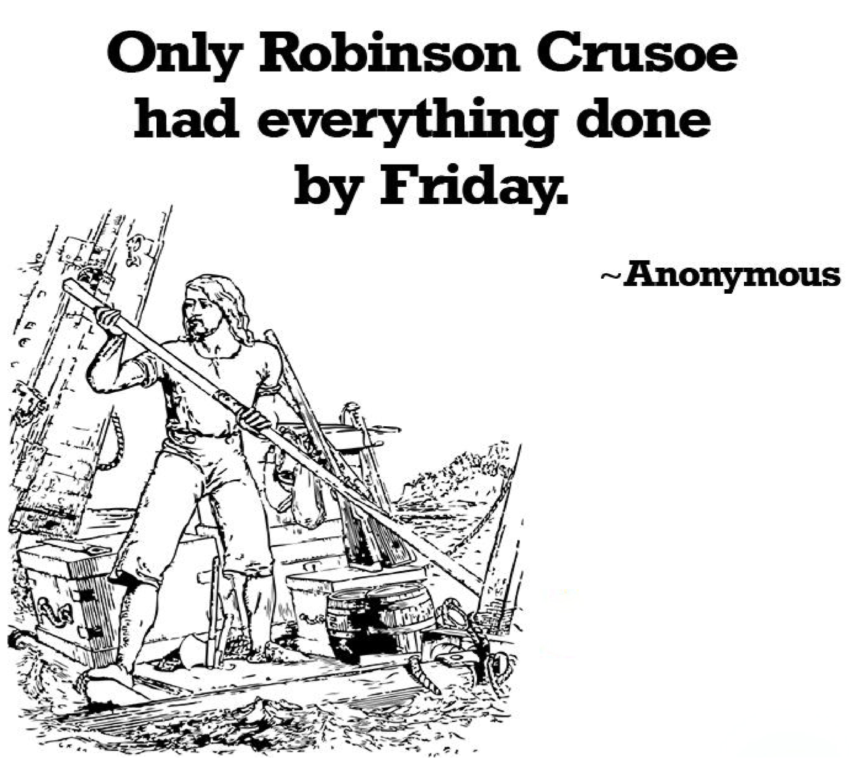
OUR PARSER Pro3Gres uses chunker pre-processing, it only parses between chunk heads. Multi-word terms (MWT) can be treated like chunks, e.g. by replacing the MWT by its head in a pre-processing step.

- On in-domain text (Penn, GREVAL):
  - with standard NER (LT-TTT2, Grover (2008)): worse to similar, most multi-word terms are shorter than chunks. Re-chunking on term heads leads to similar results.
- On out-of-domain text (Biomedical):
  - with domain NER: Better than chunker (Weeds et al., 2007), as it corrects many tagging errors, which are frequent (e.g. protein names)
  - with domain-trained tagger: similar to slightly lower performance $\rightarrow$ statistical $>$ lexical resources

Semantic expectations

THE ORIGINAL PARSER models probabilities using only those syntactic relations that are in competition. E.g. objects (e.g. *meet president*) and nominal adjuncts (e.g. *meet Friday*) are modeled as being in competition, but not subjects and objects.

The original parser models syntactic competition. We now add semantic competition: every relation is in competition with every other relation.
A sentence like *the rabbit chased the dog* now gets a lower probability than *the dog chased the rabbit* $\leftarrow$ rabbits are very unlikely to be subjects of active instances of *chase*.
This improves PP-attachment F-score from 71.9% to 72.4%.



PP interactions

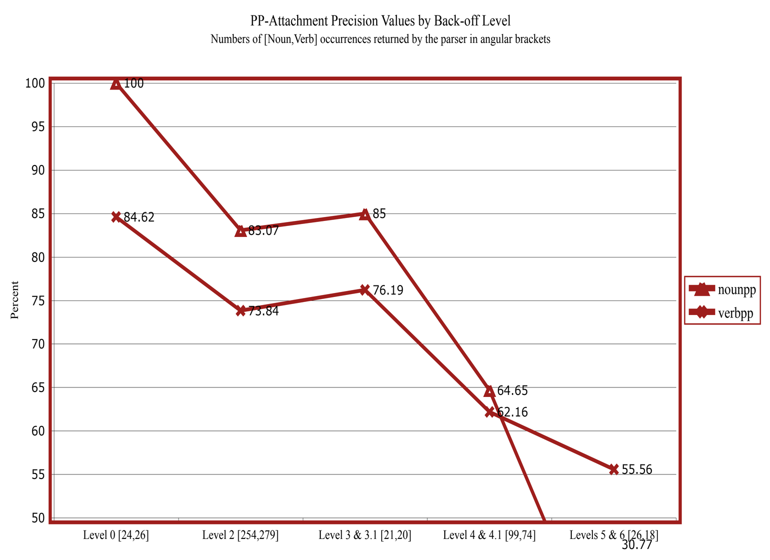
LOCAL PROBABILITIES ONLY are used by the original Pro3Gres parser. Although locality extends further in Dependency Grammar than in constituency grammar (where trees are more nested) and although there are global restrictions in the hand-written grammar, this is a shortcoming.
Now: probability that PP_2 is a dependent of PP_1 ($PP_1 < PP_2$)

$$p(verb < (PP_1 < PP_2)) = \frac{\#(verb < (PP_1 < PP_2))}{\#(verb < (PP_1 < PP_2)) + \#((verb < PP_1) < PP_2))} \tag{2}$$

This improves PP-attachment F-score marginally, from 71.9% to 72.1%.

Self-Training

SELF-TRAINING can improve results where sparseness is worse than error rate. Pro3Gres has a strong correlation between backoff level and parser accuracy. Fully lexicalized decisions have much higher performance than those further down the back-off chain.



Level 0: head + preposition + description noun, level 2: verb + preposition, level 3: head class + preposition + noun, level 4: head class + preposition + description-noun class, level 5: preposition + description-noun class, level 6: preposition only.

We need methods to reduce sparseness $\rightarrow$ more decisions can be taken at early backoff levels.

Self-training was thought to be unable to lead to better performance (Charniak, 1997; Steedman et al., 2003). Bacchiani et al. (2006) have shown that self-training can improve parsing out-of-domain texts, and is therefore a suitable approach for domain adaptation. (McClosky, Charniak, and Johnson, 2006) was the first approach to show that the use of a re-ranker (Charniak and Johnson, 2005) can also improve in-domain parsing.
We present an approach which does not need a re-ranker but marginally improves performance (from 71.9% to 72.0%): use parsed BNC probabilities where Penn TB only has low backoff counts.

Distributional Semantics

NON-NEGATIVE MATRIX FACTORIZATION (Lee and Seung, 2001) is a vector space model similar to LSA. It boosts plausible but unseen combinations.
It never uses negative weights $\rightarrow$ suitable for treating probabilities
For our verb/noun-PP attachment probability matrix

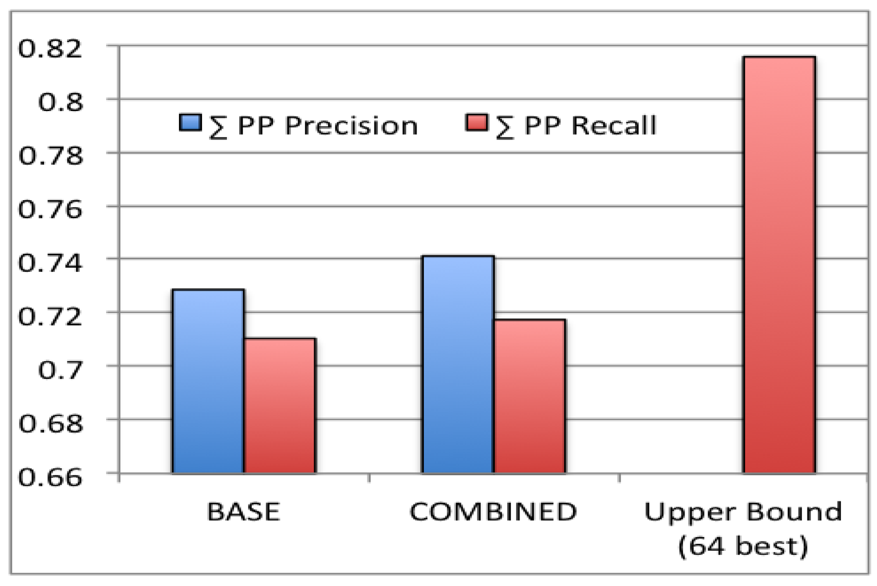
- Initial verb-prep and noun-prep matrices are filled with attachment probabilities. Null counts are given p=0.2
- We use a version for multiple PPs (verb-prep1-prep2 ; noun-prep1-prep2)
- The approximated matrix contains non-sparse (non-zero) probabilities for every verb/noun-prep1-prep2 combination.

This improves F-score from 71.9% to 72.4%.

Combined Model

WE COMBINE the described improvements. F-Score increases from 71.9% to 72.9%. This is modest, but the upper bound is low due to

- lemmatisation, tagging and chunking errors
- mapping to gold standard representation, e.g. grammar assumptions



From BASE to COMBINED

We also tried many lexical resources, but got no improvement $\rightarrow$ implicit in stats

References

Bacchiani, Michiel, Michael Riley, Brian Roark, and Richard Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20(1):41–68.

Carroll, John, Guido Minnen, and Edward Briscoe. 2003. Parser evaluation: using a grammatical relation annotation scheme. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer, Dordrecht, pages 299–316.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. of the 15th Annual Conference on Artificial Intelligence (AAAI-97)*, Stanford, USA.

Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Collins, Michael and James Brooks. 1995. Prepositional attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA.

Grover, Claire. 2008. LT-TTT2 example pipelines documentation. Technical report, Edinburgh Language Technology Group,.

Haverinen, Katri, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2008. Accurate conversion of dependency parses: targeting the Stanford scheme. In Tapio Salakoski, Dietrich Rebholz-Schuhmann, and Sampo Pyysalo, editors, *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 133–136, Turku, Finland. Turku Centre for Computer Science (TUCS).

Lee, Daniel D. and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pages 556–562.

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

McClosky, David, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia, July. Association for Computational Linguistics.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Steedman, Mark, Steven Baker, Jeremiah Crim, Stephen Clark, Julia Hockenmaier, Rebecca Hwa, Miles Osborne, Paul Ruhlen, and Anoop Sarkar. 2003. Semi-supervised training for statistical parsing. Technical Report CLSP WS-02 Final report, John Hopkins University.

Weeds, Julie, James Dowdall, Gerold Schneider, Bill Keller, and David Weir. 2007. Using distributional similarity to organise BioMedical terminology. In Fidelia Ibekwe-SanJuan, Anne Condamines, and M. Teresa Cabré Castellví, editors, *Application-Driven Terminology Engineering*. Benjamins, Amsterdam/Philadelphia.