Reconsidering non-trivial DOP estimators

Andreas van Cranenburgh Khalil Sima'an
Institute for Logic, Language and Computation, Univ. of Amsterdam
{a.w.vancranenburgh,k.simaan}@uva.nl

Introduction

A data-oriented parsing model (DOP; Scha, 1990) consists of a set of tree fragments extracted from a treebank; an estimator assigns a probability to each fragment. The traditional DOP estimation method based on relative frequencies is known to have several issues. A series of alternative estimators have been developed at the ILLC:

- DOP* (Zollmann and Sima'an, 2005)
- DOPα (Nguyen, 2004)
- Back-off DOP (Sima'an and Buratto, 2003)
- Push & pull (Zuidema, 2007)

Project proposal

The goal of the project could be one of the following:

- Establish empirically whether more sophisticated estimators are needed. There are theoretical results on this, but so far the relative frequency estimate or correction-factor approaches based on it obtain excellent performance in practice. This could work by identifying a linguistic phenomenon in the treebank and demonstrating that it is not generalized by a particular DOP model.
- Reformulate the estimator as a Double-DOP estimator, i.e., working only with a subset of fragments that occur at least twice, instead of assuming all fragments.
- Implement the estimator efficiently enough to evaluate it on the Wall Street Journal section of the Penn treebank. The implementation can be done as part of `disco-dop`[1] which already contains DOP implementations based on Goodman's DOP reduction and Double-DOP.

Deliverables

- Research report with theoretical or experimental results, ACL style, 9 pp.
- Code (where applicable).

References

Nguyen, Thuy Linh (2004). Rank consistent estimation: The DOP case. Master's thesis, University of Amsterdam. Available from: http://www.science.uva.nl/pub/theory/illc/researchreports/MoL-2004-06.text.pdf.

Scha, Remko (1990). Language theory and language technology; competence and performance. In de Kort, Q.A.M. and G.L.J. Leerdam, editors, *Computertoepassingen in de Neerlandistiek*, pages 7–22. LVVN, Almere, the Netherlands. Original title: Taaltheorie en taaltechnologie; competence en performance. Translation available at http://iaaa.nl/rs/LeerdamE.html.

Sima'an, Khalil and Luciano Buratto (2003). Backoff parameter estimation for the DOP model. *Machine Learning: ECML 2003*, pages 373–384. Available from: http://dare.uva.nl/record/126078.

Zollmann, Andreas and Khalil Sima'an (2005). A consistent and efficient estimator for Data-Oriented Parsing. *Journal of Automata Languages and Combinatorics*, 10(2/3):367. Available from: http://staff.science.uva.nl/~simaan/D-Papers/JALCsubmit.pdf.

Zuidema, Willem (2007). Parsimonious data-oriented parsing. In *Proceedings of EMNLP-CoNLL*, pages 551–560. Available from: http://aclweb.org/anthology/D/D07/D07-1058.

[1] http://github.com/andreasvc/disco-dop