Unsupervised Language Learning 2014

1 Course details

- Lecturer: Jelle Zuidema
- Office: Science Park 107, F2.45 (entrance via SP105 NikHef)
- Email: w.h.zuidema@uva.nl (only for issues that can't wait until the next class)
- Credits: 6 ECTS
- Grade: Forum posts/discussions (25%), Practical assignments (25%), Presentation & Final report (50%)
- website: Blackboard & `http://www.illc.uva.nl/laco/clas/ull14`

2 Lectures & Reading Assignments

4/2 *Lecture:* Introduction, Generative Models

6/2 *Reading:* Griffiths and Yuille (2008) *Lecture:* Statistical Inference, Expectation-Maximization

11/2 Constituency Structure *Reading:* Borensztajn and Zuidema (2007) *Lecture:* Inside-Outside

13/2 *Reading:* chapter 11 from Manning & Schütze. *Lecture:* CCM

18/2 Dependency Structure *Reading:* Klein and Manning (2005, 2004) *Lecture:* UDOP, DMV

20/2 *Reading:* Johnson (2007) *Lecture:* Fold-Unfold, EVG

25/2 Semantics *Reading:* Headden III et al. (2009), *Guest Lecture:* Ivan Titov, Learning Shallow Semantics

27/2 Latent Variables Petrov et al. (2006) *Lecture:* State Splitting, PTSGs

4/3 *Reading:* Zuidema (2007) *Lecture:* Parsimonious DOP, *Guest Lecture:* Andreas van Cranenburgh (t.b.c.), DiscoDOP

6/3 Neural models *Reading:* Bengio (2008); Socher et al. (2012) *Guest Lecture:* Phong Le, Compositional Distributional Semantics, *Lecture:* Whither ULL?

3 Computer lab & Homework Assignments

Tutorials Two exercise sets will be distributed in week 1: on regular expressions and on grammar induction. You don't need to hand-in anything about these exercises.

Assignments are described below. Some additional tips & tricks will be distributed through the website. Hand in a single page with your results at the Thursday lecture at the start of class (i.e., the assignment from week 2 & 3 is due the Thursday of week 3).

Week 1. Download the Penn WSJ corpus (training set, all trees on one line) from the course website. Calculate the word frequency distribution, the subject vs. object NP length distribution, and the phrasal rule frequency distribution. Use the unix tools `grep`, `sed`, `gawk` and regular expressions to obtain the frequency counts you need, and familiarize yourself with these tools.

Week 2 & 3. Calculate the data likelihood of the trainset sentences (sec 02-21) from the Penn WSJ corpus according to three models: bigram over words, bigram over POS-tags and treebank PCFG. You can use BitPar as a fast parser to compute sentence probabilities, and sed and regular expressions to create the grammars you need.

Week 4. Implement the fold-unfold transform of Johnson (2007).

Week 5. Implement the DMV model of Klein and Manning (2004) using Johnson's transform and evaluate it on WSJ10 using a standard IO implementation (such as offered by BitPar).

Week 6-7. Miniproject: replicate a published paper on unsupervised language learning and try to extend it one step further. Present your results at the ULL workshop. Hand in a report consisting of a literature review, your model and results at the end of the block (maximum 8 pages).

4 Readings

T.L. Griffiths and A. Yuille. A primer on probabilistic inference. *The probabilistic mind: Prospects for Bayesian cognitive science*, pages 33–57, 2008.

G. Borensztajn and W. Zuidema. Bayesian model merging for unsupervised constituent labeling and grammar induction. *ILLC Preprint*, 2007.

D. Klein and C.D. Manning. Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, 38(9):1407–1419, 2005.

D. Klein and C.D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 478. Association for Computational Linguistics, 2004.

M. Johnson. Transforming projective bilexical dependency grammars into efficiently-parsable cfgs with unfold-fold. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 168, 2007.

W.P. Headden III, M. Johnson, and D. McClosky. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 101–109. Association for Computational Linguistics, 2009.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings ACL-COLING'06*, pages 443–440. Association for Computational Linguistics Morristown, NJ, USA, 2006.

Willem Zuidema. Parsimonious Data-Oriented Parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 551–560, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1058.

Yoshua Bengio. Neural net language models. Scholarpedia, 3(1):3881., 2008.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics, 2012.