

Machine Learning Approaches to Analysis of Corpora

Eric Atwell

CVL: Computer Vision and Language research group
School of Computing – University of Leeds
Leeds LS2 9JT, England
eric@comp.leeds.ac.uk

Résumé – Abstract

Corpus-based Machine Learning of linguistic annotations has been a key topic for all areas of Natural Language Processing. The first part of this tutorial deals with supervised learning from hand-annotated corpora, and presents a survey along three dimensions of classification. First we outline different linguistic level of analysis: Tokenisation, Part-of-Speech tagging, Parsing, Semantic analysis and Discourse annotation. Secondly, we introduce alternative approaches to Machine Learning applicable to linguistic annotation of corpora: N-gram and Markov models, Neural Networks, Transformation-Based Learning, Decision Tree learning, and Vector-based classification. Thirdly, we examine a range of Machine Learning systems for arguably the most challenging level of linguistic annotation, discourse analysis; these illustrate the various Machine Learning approaches.

The second part of the tutorial focusses on Unsupervised learning. In recent years there have been significant advances in the field of Unsupervised Grammar Inference (UGI) for Natural Languages such as English or Dutch. The tutorial presents a broad range of UGI implementations, where we can begin to see how the theory has been put in to practice. Several mature systems are emerging, built using complex models and capable of deriving natural language grammatical phenomena. The range of systems is classified into: systems for categorizing words into Part-of-Speech classes based on context-based clustering and/or constraints; models based on Categorical Grammar (GraSp, CLL, EMILE); Memory Based Learning models (FAMBL, RISE); Evolutionary computing models (ILM, LAgts); and string-pattern searches (ABL, GB). An objectively measurable statistical comparison of performance Of the systems reviewed is not yet feasible. However, their merits and shortfalls are discussed, as well as a look at what the future has in store for Unsupervised Grammar Inference.

The third part of the tutorial presents a series of personal research perspectives on English Corpus Linguistics, focussing on Natural Language Learning techniques for deriving language models from corpora, and applications of these corpus-derived language models. In summary, the main contributions are:

- ❖ An analysis of the Constituent-Likelihood language model used in the grammatical Part-of-Speech (PoS) tagging of the Lancaster-Oslo/Bergen Corpus, to derive a new formalism for representing parse-trees in a corpus as a sequence of “hypertags”, thus allowing parsing to be cast computationally as a natural extension of tagging (Atwell et al 84, Atwell 87a, Atwell 88a);
- ❖ A comparative evaluation of Markov model and neural network techniques for parsing via hypertagging (Atwell 93a);
- ❖ Techniques for unsupervised learning of English grammatical categories from “raw” (un-tagged) corpus text (Atwell 87b, Atwell and Drakos 87, Hughes and Atwell 94);
- ❖ Applications of the Constituent-Likelihood PoS-tagging technique to detection of grammatical errors in English word-processed text (Atwell 87c, Atwell and Elliott 87, Atwell 90);
- ❖ Techniques for learning a context-free parser from a Treebank (parsed corpus), by converting each parse-tree into a set of Prolog grammar rules (Atwell 88b, Atwell and Lajos 93);
- ❖ Techniques for supervised learning of multi-level linguistic annotation (Atwell 93b, Atwell 96a);
- ❖ Techniques for comparative evaluation of rival corpus-based grammatical annotation schemes (Atwell 96b, Atwell and Sutcliffe 97, Atwell et al 2000a);
- ❖ An analysis of applications of English corpus linguistics and natural language learning in TESOL, teaching English to speakers of other languages (Atwell 86, Atwell 99, Atwell et al 2000b);
- ❖ Applications of English corpus linguistics and natural language learning in design of a message corpus in preparation for Extra-Terrestrial contact (Atwell and Elliott 2001a,b).

Our overall aim is to provide an ontology or framework for further development of research applying Machine Learning to Corpus Linguistics.

The tutorial will cover material from a number of sources, principally:

Xunlei Rose Hu and Eric Atwell , 2003. A survey of machine learning approaches to analysis of large corpora. In: Simov K, Osenova P (eds.) Proceedings of SProLaC: Workshop on Shallow Processing of Large Corpora, held in conjunction with the Corpus Linguistics 2003 conference, UCREL technical paper number 17. UCREL, Lancaster University, pp.45-52.

and: **Andrew Roberts and Eric Atwell, 2002. Unsupervised Grammar Inference Systems for Natural Language. Research Report number 2002.20. School of Computing, University of Leeds**

and: **Eric Atwell, 2004. English Corpus Linguistics and Natural Language Learning. Unpublished PhD thesis, School of Computing, University of Leeds**

Keywords – Mots Clés

Corpus, marquage des catégories grammaticales, regroupement, unification, catégories de mots, type/occurrence, évaluation.

Corpus, Part-of-Speech tagging, clustering, unification, word classes, type/token, evaluation

Références

Adriaans, P.W., 1992. *Language Learning from a Categorical Perspective*. Ph.D. thesis, Unversiteit van Amsterdam.

Adriaans, P.W. 1999. *Learning shallow context-free languages under simple distributions*. ILLC Report PP-1999-13, Institute for Logic, Language and Computation, Amsterdam.

Atwell, E, 1983. *Constituent-Likelihood Grammar*. ICAME Journal of the International Computer Archive of Modern English Vol.7 pp. 34-65.

Atwell, E, Leech, G & Garside, R, 1984. *Analysis of the LOB Corpus: progress and prospects* in Aarts, J & Meijs, W (editors), *Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research*, pp40-52, Rodopi.

Atwell E, 1986. *Beyond the micro: advanced software for research and teaching from computer science and artificial intelligence*. In Leech G and Candlin C (editors) *Computers in English language teaching and research: selected papers from the British Council Symposium*, pp167-183, Longman.

Atwell E, 1987a. *Constituent-likelihood grammar*. In Garside R, Sampson G and Leech G (editors) *The computational analysis of English: a corpus-based approach*, London, Longman.

Atwell E, 1987b. *A parsing expert system which learns from corpus analysis*. In Meijs W, (editor), *Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference on English Language Research on Computerised Corpora*, pp227-235, Amsterdam, Rodopi, 1987.

Atwell E, 1987c. *How to detect grammatical errors in a text without parsing it*. In Maegaard B (editor), *Proceedings of EACL: Third Conference of the European Chapter of the Association for Computational Linguistics*, New Jersey, ACL.

Atwell E and Drakos N, 1987. *Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text*. In Maegaard B (editor), *Proceedings of EACL: Third Conference of the European Chapter of the Association for Computational Linguistics*, New Jersey, ACL.

Atwell E and Elliot S, 1987. *Dealing with ill-formed English text*. In Garside R, Sampson G and Leech G (editors) *The computational analysis of English: a corpus-based approach*, London, Longman.

Atwell E, 1988a. *Grammatical analysis of English by statistical pattern recognition*. In Kittler J (editor), *Pattern Recognition: Proceedings of the 4th International Conference*, pp626-635, Berlin, Springer-Verlag.

Atwell E, 1988b. Transforming a Parsed Corpus into a Corpus Parser. In Kyto M, Ihalainen O and Rissanen M (editors), *Corpus Linguistics, Hard and Soft: Proceedings of the ICAME 8th International Conference on English Language Research on Computerised Corpora*, pp61-70, Rodopi.

Atwell E, 1990. Measuring grammaticality of machine-readable text. In Bahner W, Schildt J and Viehweger D (editors), *Proceedings of the Fourteenth International Congress of Linguists, Volume III*, pp2275-2277, Berlin, Akademie-Verlag.

Atwell E, 1993a. Corpus-based statistical modelling of English grammar. In Souter C and Atwell E (editors), *Corpus-based computational linguistics: proceedings of the 12th conference of the International Computer Archive of Modern English*, pp195-214, Amsterdam, Rodopi.

Atwell E, 1993b. Linguistic Constraints for Large-Vocabulary Speech Recognition. In Atwell E (editor), *Knowledge at Work in Universities: Proceedings of the second annual conference of the Higher Education Funding Councils' Knowledge Based Systems Initiative*, Leeds University Press.

Atwell E, and Lajos G, 1993. Knowledge and Constraint Management: Large Scale Applications. In Atwell E (editor), *Knowledge at Work in Universities: Proceedings of the second annual conference of the Higher Education Funding Councils' Knowledge Based Systems Initiative*, Leeds University Press.

Hughes J and Atwell E, 1994. The automated evaluation of inferred word classifications. In Cohn A (editor), *Proceedings of ECAI'94: 11th European Conference on Artificial Intelligence*, pp535-540, John Wiley.

Atwell E, 1996a. Machine Learning from Corpus Resources for Speech And Handwriting Recognition. In Thomas J, and Short M (editors), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, pp151-166, Longman.

Atwell E, 1996b. Comparative Evaluation of Grammatical Annotation Models. In Sutcliffe R, Koch H, and McElligott A (editors), *Industrial Parsing of Software Manuals*, pp25-46, Amsterdam, Rodopi.

Atwell E and Sutcliffe R, 1997. Industrial Parsing of Software Manuals: Empirical Qualitative Comparison of Parsers and Parsing Schemes. In Gaizauskas R (editor), *Evaluation in Speech and Language Technology: Proceedings of the SALT Club Workshop*", Sheffield University.

Atwell E, 1999. *The Language Machine*. London, British Council.

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S, 2000a. A comparative evaluation of modern English corpus grammatical annotation schemes. In *ICAME Journal*, v.24, pp7-23, International Computer Archive of Modern and medieval English, Bergen.

Atwell E, Howarth P, Souter C, Baldo P, Bisiani R, Bonaventura P, Herron D, Menzel W, Morton R, and Wick H, 2000b. User-Guided System Development in Interactive Spoken Language Education. In *Natural Language Engineering*, vol.6 no.3. Cambridge University Press.

Atwell, E. and Elliott, J. 2001a. A corpus for interstellar communication. In Rayson, P, Wilson, A, McEnery, T, Hardie, A, and Khoja, S. (Eds.), *Proceedings of CL2001: International Conference on Corpus Linguistics*. UCREL Technical Paper 13, Lancaster University, 2001, pp. 31-39.

Atwell E and Elliott J, 2001b. Corpus Linguistics and the design of a response message. In: *Proceedings of IAC'2001: the 51st International Astronautical Congress*, Toulouse, France.

Bates, E. and Goodman, J.C. 1997. On the Inseparability of Grammar and the Lex-icon: Evidence from Acquisition, Aphasia, and Real-time Processing. *Language and Cognitive Processes*, 12, pp. 507-584.

Belkin, M. and Goldsmith, J. 2002. Using eigenvectors of the bigram graph to infer morpheme identity. *Proceedings of the Morphology/Phonology Learning Workshop of ACL-02*. Association for Computational Linguistics.

Clustering of Word Types and Unification of Word Tokens into Grammatical Word-Classes

- Bod, R. 1993. *Using an annotated corpus as a stochastic grammar*. In Proceedings of EACL, the sixth conference of the European chapter of the Association for Computational Linguistics, pp.37-44.
- Brill, E. 1995. *Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging*. Computational Linguistics, volume 21(4), pages 543-566.
- Briscoe, E.J. 2000. Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device. *Language*, 76(2). pp. 245-296.
- Cohen, P. 1995. *Empirical methods for Artificial Intelligence*. MIT Press, Cambridge MA.
- Daelemans, W, Van den Bosch, A and Zavrel, J. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 11. pp. 11-43.
- Demetriou G and Atwell E. 2001. *A domain-independent semantic tagger for the study of meaning associations in English text*. In Harry Bunt, Ielka van der Sluis and Elias Thijsse (editors), Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4) pp.67-80. Tilburg, Netherlands.
- Domingos, P. 1995. The RISE 2.0 system: A case study in multistrategy learning. Technical Report 95-2, Department of Information and Computer Science, University of California.
- Dong, S and Searls, D.B.1994. Gene Structure Prediction by Linguistic Methods. *Genomics*.
- Elliott, J, Atwell, E and Whyte, W. 2001. Visualisation of Long Distance Grammatical Collocation Patterns in Language. In IV2001: 5th International Conference on Information Visualisation, London, UK.
- Finch, S. 1993. Finding structure in language. PhD thesis, Edinburgh University.
- Freltag, D. 1997. Using Grammatical Inference to Improve Precision in Information Extraction. In Working Papers of the ICML-97 Workshop on Automata Induction, Grammatical Inference and Language Acquisition.
- Gold, E.M. 1967. Language Identification in the Limit. *Information and Control*. 10. pp. 447-474.
- Henrichsen, P.J. 2002. GraSp: Grammar Learning from unlabelled speech corpora. In: Roth, D and Van den Bosch, A. (Eds), Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 22-28.
- Hong, T.W and Clark, K.L. 2001. Using Grammatical Inference to Automate Extraction from the Web. In Principles of Data Mining and Knowledge Discovery. pp. 216-227.
- Hughes, J and Atwell, E. 1994. The automated evaluation of inferred word classifications. In Cohn, A. (Ed), Proceedings of ECAI'94: 11th European Conference on Artificial Intelligence. John Wiley, pp.535-539.
- Kirby, S and Hurford, J. 2002. The emergence of linguistic structure: an overview of the iterated learning model. In: Cangelosi, A and Parisi, D (Eds.) *Simulating the Evolution of Language*. pp. 121-148.
- Kirby, S. 2002. Natural Language from Artificial Life. *Artificial Life*, 8(2). pp.185-215.
- Jurafsky, D and Martin, J.2000. *Speech and Language Processing*. Prentice-Hall.
- Leech, G, Garside, R & Atwell, E, 1983a. *Recent developments in the use of computer corpora in English language research*, in Transactions of the Philological Society, pp.23-40.
- Leech, G, Garside, R & Atwell, E, 1983b. *The Automatic Grammatical Tagging of the LOB Corpus* ICAME Journal of the International Computer Archive of Modern English Vol.7

Leech, G, Barnett, R and Kahrel, P. 1996. EAGLES Final Report and guide-lines for the syntactic annotation of corpora. European Expert Advisory Group on Language Engineering Standards (EAGLES) Report EAG-TCWG-SASG/1.5.

Mast, M, Kompe, R, Harbeck, S, Keissling, A, Niemann, H, Noth, E, Schukat-Talamazzini, E, and Warnke, V, 1996. *Dialog act classification with the help of prosody*. In Proceedings of ICLSP-96, Philadelphia, volume 3, pp1732-1735.

Paskin, M.A. 2001. Grammatical Bigrams. In Dietterich, T, Becker, S, and Gharahmani, Z (eds.), *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Reithinger, N, Engel, R, Kipp, M, and Klesen, M 1996. *Predicting dialogue acts for a speech-to-speech translation system*. In Proceedings ICLSP-96, Philadelphia, volume 2, pp.654-657.

Reithinger, N, and Klesen, M, 1997. *Dialogue act classification using language models*. In Proceedings of EUROSPEECH-97, volume 4, pp.2235-2238.

Roberts, A. 2002. Automatic acquisition of word classification using distributional analysis of content words with respect to function words. Technical Report, School of Computing, University of Leeds.

Samuel, K, Carberry, S, and Vijay-Shanker, K, 1998. *Dialogue act tagging with transformation-based learning*. In Proceedings of COLING/ACL-98, Montreal, volume 2, pp.1150-1156.

Steedman, M. 1989. Constituency and Coordination in a Combinatory Grammar. In: Baltin, M.R and Kroch, A.S (Eds.), *Alternative Conceptions of Phrase Structure*. University of Chicago. pp. 201-231.

Stolcke, A, Shriberg, E, Bates, R, Coccaro, N, Jurafsky, D, Martin, R, Meteer, M, Ries, K, Taylor, P, and Van Ess-Dykema, C. 1998. *Dialog Act Modeling for Conversational Speech*, in Chu-Carroll, J, and Green, N, (eds) *Applying machine learning to discourse processing: AAI Spring Symposium*, pp.98-105.

Taylor, P, King, S, Isard, S, and Wright, H, 1998. *Intonation and dialog context as constraints for speech recognition*. *Language and Speech*, volume 41(3-4), pages 489-508.

Van den Bosch, A. 1999. Careful abstraction from instance families in memory-based language learning. *Journal of Experimental and Theoretical Artificial Intelligence*, 11:3, special issue on Memory-Based Language Processing, pp. 339-368.

Van Zaanen, M and Adriaans, P.W, 2001. Comparing Two Unsupervised Grammar Induction Systems: Alignment-Based Learning vs. EMILE. Technical Report 2001.05, School of Computing, University of Leeds.

Van Zaanen, M. 2002. Bootstrapping Structure into Language: Alignment-Based Learning. PhD Thesis, School of Computing, University of Leeds.

Vervoort, M.R. 2000. Games, Ealks and Grammars. Ph.D thesis, Universiteit van Amsterdam.

Watkinson, S and Manandhar, S.A 2001. Psychologically Plausible and Computation-ally Effective Approach to Learning Syntax , CoNLL'01, the Workshop on Computational Natural Language Learning, ACL/EACL 2001.

Watkinson, S and Manandhar, S. 2001. Translating treebank annotation for evaluation. In: Proceedings of the Workshop on Evaluation Methodologies for Language and Dialogue Systems, ACL/EACL 2001.

Wood, M. 1993. *Categorial Grammars*. Routledge. London.

Woszczyna, M and Waibel, A, 1994. *Inferring linguistic structure in spoken language*. In ICSLP-94, Yokohama, pp.1363-1366.