Bibliography

[Aon98] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. SRA: Description of the IE2 system used for MUC. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. 1998. URL `http://citeseer.ist.psu.edu/aone98sra.html`.

[Ass05] Fidelis Assis, William Yerazunis, Christian Siefkes, and Shalendra Chhabra. CRM114 versus Mr. X: CRM114 notes for the TREC 2005 spam track. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*. 2005. URL `http://crm114.sourceforge.net/NIST_TREC_2005_paper.pdf`.

[Bag97] Amit Bagga and Joyce Yue Chai. A trainable message understanding system. In *CoNLL*, pp. 1–8. 1997. URL `http://citeseer.ist.psu.edu/amit97trainable.html`.

[Blo05] Stephan Bloehdorn, Philipp Cimiano, Andreas Hotho, and Steffen Staab. An ontology-based framework for text mining. *Zeitschrift für Computerlinguistik und Sprachtechnologie*, 20(1):87–112, 2005. URL `http://www.uni-koblenz.de/~staab/Research/Publications/2005/LDV_Forum_20_1-OntoTextMining.pdf`.

[Cal98a] Mary E. Califf and Raymond J. Mooney. Relational learning of pattern-match rules for information extraction. In *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 6–11. Menlo Park, CA, 1998. URL `http://citeseer.ist.psu.edu/39151.html`.

[Cal98b] Mary Elaine Califf. *Relational Learning Techniques for Natural Language Extraction*. Ph.D. thesis, University of Texas at Austin, 1998. URL `http://www.cs.utexas.edu/users/ml/papers/rapier-dissertation-98.pdf`.

[Cal03] Mary E. Califf and Raymond J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, 4:177–210, 2003. URL `http://www.jmlr.org/papers/volume4/califf03a/califf03a.pdf`.

[Cam98] Robert D. Cameron. *REX: XML Shallow Parsing with Regular Expressions*. Tech. Rep. 1998-17, School of Computing Science, Simon Fraser University, 1998. URL `http://www.cs.sfu.ca/~cameron/REX.html`.

[Car04] Andrew J. Carlson, Chad M. Cumby, Nicholas D. Rizzolo, Jeff L. Rosen, and Dan Roth. *SNoW User Manual. Version: January, 2004*. Tech. rep., UIUC, 2004. URL `http://l2r.cs.uiuc.edu/~cogcomp/software/snow-userguide.ps.gz`.

[Cha99] Joyce Yue Chai and Alan W. Biermann. The use of word sense disambiguation in an information extraction system. In *AAAI/IAAI*. 1999. URL `http://citeseer.ist.psu.edu/chai99use.html`.

[Chi02] Hai Leong Chieu and Hwee Tou Ng. A maximum entropy approach to information extraction from semi-structured and free text. In *AAAI 2002*. 2002. URL `http://citeseer.ist.psu.edu/chieu02maximum.html`.

[Cir01] Fabio Ciravegna. $(LP)^2$, an adaptive algorithm for information extraction from Web-related texts. In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. Seattle, USA, 2001. URL `http://www.smi.ucd.ie/ATEM2001/proceedings/ciravegna-atem2001.pdf`.

[Cir02] Fabio Ciravegna and Alberto Lavelli. LearningPinocchio: Adaptive information extraction for real world applications. In *Proceedings of the 2nd Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND 2002)*. Frascati, Italy, 2002. URL `http://www.dcs.shef.ac.uk/~fabio/paperi/romand2002.zip`.

[Cob02] Grégory Cobéna, Talel Abdessalem, and Yassine Hinnach. *A Comparative Study for XML Change Detection*. Gemo Report 221, INRIA, 2002. URL `ftp://ftp.inria.fr/INRIA/Projects/gemo/gemo/GemoReport-221.pdf`.

[Coh99] William W. Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2):141–173, 1999. URL `http://www-2.cs.cmu.edu/~wcohen/postscript/tois-sigir.pdf`.

[Col05] Marc E. Colosimo, Alexander A. Morgan, Alexander S. Yeh, Jeffrey B. Colombe, and Lynette Hirschman. Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, 6(Suppl 1):S12, 2005. URL `http://www.biomedcentral.com/content/pdf/1471-2105-6-S1-S12.pdf`.

[Cor05] Gordon Cormack and Thomas Lynam. Trec 2005 spam track overview. In E. M. Voorhees and Lori P. Buckland, eds., *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*. National Institute of Standards and Technology (NIST), 2005. URL `http://plg.uwaterloo.ca/~gvcormac/trecspamtrack05/trecspam05paper.pdf`.

[CRM] CRM114: The controllable regex mutilator. `http://crm114.sourceforge.net/`.

[Cun02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. 2002.

[Dag97] Ido Dagan, Yael Karov, and Dan Roth. Mistake-driven learning in text categorization. In Claire Cardie and Ralph Weischedel, eds., *Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing*, pp. 55–63. Association for Computational Linguistics, Providence, US, 1997. URL `http://citeseer.ist.psu.edu/552405.html`.

[Dem02] George Demetriou and Robert Gaizauskas. Utilizing text mining results: The PastaWeb system. In *Proceedings of the Association for Computational Linguistics Workshop on Natural Language Processing in the Biomedical Domain*, pp. 77–84. 2002.

[Eik99] Line Eikvil. *Information Extraction from World Wide Web – A Survey*. Tech. Rep. 945, Norwegian Computing Center, 1999. URL `http://citeseer.ist.psu.edu/eikvil99information.html`.

[Fel98] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Fin98] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998. URL `http://citeseer.ist.psu.edu/article/fine98hierarchical.html`.

[Fin03] Aidan Finn and Nicholas Kushmerick. Active learning selection strategies for information extraction. In *Proceedings of the International Workshop on Adaptive Text Extraction and Mining*. 2003. URL `http://www.dcs.shef.ac.uk/~fabio/ATEM03/finn-ecml03-atem.pdf`.

[Fin04a] Aidan Finn and Nicholas Kushmerick. Information extraction by convergent boundary classification. In *AAAI-2004 Workshop on Adaptive Text Extraction*

and Mining. San Jose, USA, 2004. URL http://www.ai.sri.com/~muslea/atem-04/finn.pdf.

[Fin04b] Aidan Finn and Nicholas Kushmerick. Multi-level boundary classification for information extraction. In *ECML 2004*, pp. 111–122. 2004.

[Fin06] Aidan Finn. *A Multi-Level Boundary Classification Approach to Information Extraction*. Ph.D. thesis, University College Dublin, 2006.

[Fre98a] Dayne Freitag. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. thesis, Carnegie Mellon University, 1998. URL http://www-2.cs.cmu.edu/afs/cs/user/dayne/www/ps/diss-freitag.ps.

[Fre98b] Dayne Freitag. Toward general-purpose learning for information extraction. In Christian Boitet and Pete Whitelock, eds., *Proc. 36th Annual Meeting of the Association for Computational Linguistics*, pp. 404–408. San Francisco, CA, 1998. URL http://citeseer.ist.psu.edu/freitag98toward.html.

[Fre99] Dayne Freitag and Andrew K. McCallum. Information extraction with HMMs and shrinkage. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*. 1999. URL http://citeseer.ist.psu.edu/freitag99information.html.

[Fre00a] Dayne Freitag and Nicholas Kushmerick. Boosted wrapper induction. In *AAAI/IAAI*, pp. 577–583. 2000. URL http://citeseer.ist.psu.edu/freitag00boosted.html.

[Fre00b] Dayne Freitag and Andrew K. McCallum. Information extraction with HMM structures learned by stochastic optimization. In *AAAI/IAAI*, pp. 584–589. 2000. URL http://citeseer.ist.psu.edu/freitag00information.html.

[Für99] Johannes Fürnkranz. Separate-and-conquer rule learning. *Artificial Intelligence Review*, 13(1):3–54, 1999. URL http://citeseer.ist.psu.edu/26490.html.

[Gra03] Paul Graham. Better Bayesian filtering. In *MIT Spam Conference*. 2003. URL http://www.paulgraham.com/better.html.

[Han02] Siegfried Handschuh, Steffen Staab, and Fabio Ciravegna. S-CREAM: Semi-automatic creation of metadata. In Asuncion Gomez-Perez and V. Richard Benjamins, eds., *Proc. 13th International Conference on Knowledge Engineering and Management*. 2002. URL http://www.aifb.uni-karlsruhe.de/~sst/Research/Publications/ekaw2002scream-sub.pdf.

[HTM] *HTML 4.01 Specification*. URL http://www.w3.org/TR/html4/. W3C Recommendation, 24 December 1999.

[JTi] JTidy. http://jtidy.sourceforge.net/.

[Kah03] Heiko Kahmann. *Erstellung eines Systems zur effizienten Unterstützung eines Anwenders bei der manuellen, modellbasierten Faktenextraktion und zur Qualitätssicherung von Ergebnissen automatischer Extraktionssysteme*. Diplomarbeit, Freie Universität Berlin, 2003.

[Kau02] David Kauchak, Joseph Smarr, and Charles Elkan. *Sources of Success for Information Extraction Methods*. Tech. Rep. CS2002-0696, UC San Diego, 2002. URL http://www-cse.ucsd.edu/users/elkan/BWI.pdf.

[Kim04] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at JNLPBA. In *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (BioNLP/NLPBA 2004)*. 2004. URL http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/shared_task_intro.pdf.

[Koh03] Michael Kohlhase and Romeo Anghelache. Towards collaborative content

management and version control for structured mathematical knowledge. In *Second International Conference on Mathematical Knowledge Management (MKM 2003)*. 2003. URL `http://link.springer.de/link/service/series/0558/bibs/2594/25940147.htm`.

[Kom03] Kyriakos Komvoteas. *XML Diff and Patch Tool*. Master's thesis, Computer Science Department, Heriot-Watt University, Edinburgh, Scotland, 2003. URL `http://treepatch.sourceforge.net/report.pdf`.

[Laf01] John Lafferty, Andrew K. McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*. 2001. URL `http://citeseer.ist.psu.edu/lafferty01conditional.html`.

[Lap02] C. Laprun, J. Fiscus, J. Garofolo, and S. Pajot. A practical introduction to ATLAS. In *Proceedings of the Third International Conference on Language Resources and Evaluation(LREC)*. 2002. URL `http://www.nist.gov/speech/atlas/download/lrec2002-atlas.pdf`.

[Lav04a] A. Lavelli, M. Califf, F. Ciravegna, D. Freitag, C. Giuliano, N. Kushmerick, and L. Romano. A critical survey of the methodology for IE evaluation. In *4th International Conference on Language Resources and Evaluation (LREC)*. 2004. URL `http://tcc.itc.it/people/lavelli/papers/lavelli-lrec2004.ps.gz`.

[Lav04b] A. Lavelli, M.-E. Califf, F. Ciravegna, D. Freitag, C. Giuliano, N. Kushmerick, and L. Romano. IE evaluation: Criticisms and recommendations. In *AAAI-2004 Workshop on Adaptive Text Extraction and Mining*. San Jose, USA, 2004. URL `http://www.ai.sri.com/~muslea/atem-04/lavelli.pdf`.

[Lin01] Tancred Lindholm. *A 3-way Merging Algorithm for Synchronizing Ordered Trees—The 3DM Merging and Differencing Tool for XML*. Master's thesis, Helsinki University of Technology, Dept. of Computer Science, 2001. URL `http://www.cs.hut.fi/~ctl/3dm/thesis.pdf`.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[Mae01] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. In *International Workshop on Next Generation Geospatial Information*, vol. 16, pp. 72–79. 2001. URL `http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/ieee_semweb.pdf`.

[Man05] Inderjeet Mani, Zhangzhi Hu, Seok Bae Jang, Ken Samuel, Matthew Krause, Jon Phillips, and Cathy H. Wu. Protein name tagging guidelines: Lessons learned. *Comparative and Functional Genomics*, 6:72–76, 2005. URL `http://www3.interscience.wiley.com/cgi-bin/fulltext/109932700/PDFSTART`.

[McC00] Andrew K. McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy Markov models for information extraction and segmentation. In *ICML*. 2000. URL `http://www-2.cs.cmu.edu/afs/cs/user/dayne/www/ps/memm.ps`.

[McC02] Andrew Kachites McCallum. MALLET: A machine learning for language toolkit, 2002. URL `http://mallet.cs.umass.edu/`.

[McC03a] Andrew McCallum and Ben Wellner. Object consolidation by graph partitioning with a conditionally-trained distance metric. In *KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*. 2003. URL `http://csaa.byu.edu/kdd03-papers/mccallum-wellner.ps`.

[McC03b] Andrew K. McCallum and David Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In *IJCAI'03 Workshop on Learning Statistical Models from Relational*

Data. 2003. URL
http://www.cs.umass.edu/~mccallum/papers/iedatamining-ijcaiws03.pdf.

[Mil98] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group. Algorithms that learn to extract information—BBN: Description of the SIFT system as used for MUC. In *MUC-7*. 1998. URL
http://citeseer.ist.psu.edu/miller98algorithms.html.

[Mil00] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A novel use of statistical parsing to extract information from text. In *ANLP-NAACL*, pp. 226–233. 2000. URL http://citeseer.ist.psu.edu/miller00novel.html.

[Mun99] Marcia Munoz, Visin Punyakanok, Dan Roth, and Dav Zimak. *A Learning Approach to Shallow Parsing*. Tech. Rep. UIUCDCS-R-99-2087, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1999. URL http://citeseer.ist.psu.edu/333381.html.

[Mus01] Ion Muslea, Steven Minton, and Craig A. Knoblock. Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1/2):93–114, 2001. URL
http://citeseer.ist.psu.edu/muslea01hierarchical.html.

[Mus03] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active learning with strong and weak views: A case study on wrapper induction. In *International Joint Conference on Artificial Intelligence (IJCAI 2003)*. 2003. URL
http://www.ics.uci.edu/~muslea/PS/ijcai-03.pdf.

[Nah00] Un Yong Nahm and Raymond J. Mooney. Using information extraction to aid the discovery of prediction rules from text. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*. Boston, MA, 2000. URL
http://citeseer.ist.psu.edu/nahm00using.html.

[Neu02] Günter Neumann and Jakob Piskorski. A shallow text processing core engine. *Journal of Computational Intelligence*, 2002. URL
http://www.dfki.de/%7Eneumann/publications/new-ps/comp-intell.pdf.

[nor] normalizemime v2004-02-04. http://hyvatti.iki.fi/~jaakko/spam/.

[ODF] *Open Document Format for Office Applications (OpenDocument) v1.0*. URL
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office.
OASIS Standard, 1 May 2005.

[Per04] Janet Perna and Alfred Spector. Unstructured information management. *IBM Systems Journal*, 43(3), 2004. URL
http://www.research.ibm.com/journal/sj43-3.html.

[Pes03] Leonid Peshkin and Avi Pfeffer. Bayesian information extraction network. In *International Joint Conference on Artificial Intelligence (IJCAI 2003)*. 2003. URL http://citeseer.ist.psu.edu/565989.html.

[Qui95] J. Ross Quinlan and R. Mike Cameron-Jones. Induction of logic programs: FOIL and related systems. *New Generation Computing*, 13(3,4):287–312, 1995. URL
http://citeseer.ist.psu.edu/quinlan95induction.html.

[Reu00] Reuters corpus, volume 1 (English language, 1996-08-20 to 1997-08-19), 2000.

[RISa] RISE Repository. http://www.isi.edu/info-agents/RISE/.

[RISb] RISE seminar announcements corpus. URL
http://www-2.cs.cmu.edu/~dayne/SeminarAnnouncements/__Source__.html.

[Rot01] Dan Roth and Wen-tau Yih. Relational learning via propositional algorithms:

An information extraction case study. In *International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pp. 1257–1263. 2001. URL http://citeseer.ist.psu.edu/roth01relational.html.

[Rot02] Dan Roth and Wen-tau Yih. Probabilistic reasoning for entity & relation recognition. In *COLING'02*. 2002. URL http://l2r.cs.uiuc.edu/~danr/Papers/er-coling02.pdf.

[Sch01] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden Markov models for information extraction. In *Proceedings of the International Symposium on Intelligent Data Analysis*. 2001. URL http://citeseer.ist.psu.edu/sche01active.html.

[Sch02] Tobias Scheffer, Stefan Wrobel, Borislav Popov, Damyan Ognianov, Christian Decomain, and Susanne Hoche. Learning hidden Markov models for information extraction actively from partially labeled text. *Künstliche Intelligenz*, (2), 2002. URL http://kd.cs.uni-magdeburg.de/~scheffer/papers/kisi.ps.gz.

[Sie02] Christian Siefkes. *A Toolkit for Caching and Prefetching in the Context of Web Application Platforms*. Diplomarbeit, TU Berlin, 2002.

[Sie03] Christian Siefkes. Learning to extract information for the Semantic Web. In Robert Tolksdorf and Rainer Eckstein, eds., *Tagungsband Berliner XML Tage 2003*, pp. 452–459. 2003. URL http://www.siefkes.net/papers/ie-semantic-web.pdf.

[Sie04a] Christian Siefkes. A shallow algorithm for correcting nesting errors and other well-formedness violations in XML-like input. In *Extreme Markup Languages (EML) 2004*. 2004. URL http://www.siefkes.net/papers/eml/EML2004.pdf.

[Sie04b] Christian Siefkes, Fidelis Assis, Shalendra Chhabra, and William S. Yerazunis. Combining Winnow and orthogonal sparse bigrams for incremental spam filtering. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, eds., *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004)*, vol. 3202 of *Lecture Notes in Artificial Intelligence*, pp. 410–421. Springer, 2004. URL http://www.siefkes.net/papers/winnow-spam.pdf.

[Sie05a] Christian Siefkes. Incremental information extraction using tree-based context representations. In Alexander Gelbukh, ed., *Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2005)*, vol. 3406 of *Lecture Notes in Computer Science*, pp. 510–521. Springer, 2005. URL http://www.siefkes.net/papers/incremental-ie.pdf.

[Sie05b] Christian Siefkes and Peter Siniakov. An overview and classification of adaptive approaches to information extraction. *Journal on Data Semantics*, IV:172–212, 2005. URL http://www.siefkes.net/papers/overview-ie.pdf. LNCS 3730.

[Sie06] Christian Siefkes. A comparison of tagging strategies for statistical information extraction. In *HLT-NAACL 2006*. 2006. URL http://www.siefkes.net/papers/tagging-strategies-ie.pdf.

[Sko03] Marios Skounakis, Mark Craven, and Soumya Ray. Hierarchical hidden Markov models for information extraction. In *International Joint Conference on Artificial Intelligence (IJCAI 2003)*. 2003. URL http://www.biostat.wisc.edu/~craven/papers/ijcai03.pdf.

[SM04] C. M. Sperberg-McQueen and Lou Burnard. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, 2004. URL http://www.tei-c.org/P4X/.

[Sod95] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. CRYSTAL: Inducing a conceptual dictionary. In Chris Mellish, ed., *International Joint Conference on Artificial Intelligence (IJCAI 1995)*, pp. 1314–1319. San Francisco, 1995. URL `http://citeseer.ist.psu.edu/soderland95crystal.html`.

[Sod97a] Stephen Soderland. *Learning Text Analysis Rules for Domain-specific Natural Language Processing*. Ph.D. thesis, University of Massachusetts, Amherst, 1997. URL `http://citeseer.ist.psu.edu/279256.html`.

[Sod97b] Stephen Soderland. Learning to extract text-based information from the World Wide Web. In *Proc. Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 251–254. 1997. URL `http://citeseer.ist.psu.edu/soderland97learning.html`.

[Sod99] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1–3):233–272, 1999. URL `http://citeseer.ist.psu.edu/soderland99learning.html`.

[Sod01] Stephen Soderland. Building a machine learning based text understanding system. In *Proc. IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. 2001. URL `http://www.smi.ucd.ie/ATEM2001/proceedings/soderland-atem2001.pdf`.

[SpA] SpamAssassin. `http://www.spamassassin.org/`.

[SpB] SpamBayes. `http://spambayes.sourceforge.net/`.

[Sut05] Charles Sutton and Andrew McCallum. Composition of conditional random fields for transfer learning. In *HLT/EMNLP 2005*. 2005. URL `http://www.cs.umass.edu/~mccallum/papers/transfer-emnlp05.pdf`.

[Tho99] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *Proc. 16th International Conf. on Machine Learning*, pp. 406–414. 1999. URL `http://citeseer.ist.psu.edu/thompson99active.html`.

[Tid] HTML Tidy. `http://tidy.sourceforge.net/`.

[TKS03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, eds., *Proceedings of CoNLL-2003*, pp. 142–147. Edmonton, Canada, 2003. URL `http://cnts.uia.ac.be/conll2003/pdf/14247tjo.pdf`.

[Tre] TreeTagger. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

[txt] txt2html. `http://txt2html.sourceforge.net/`.

[Wal99] Norman Walsh and Leonard Muellner. *DocBook: The Definitive Reference*. O'Reilly, Sebastopol, CA, 1999.

[Wit91] Ian H. Witten and Timothy C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4), 1991.

[Wit99] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999. URL `http://www.cs.waikato.ac.nz/ml/weka/`.

[XMLa] *Extensible Markup Language (XML) 1.0 (Third Edition)*. URL `http://www.w3.org/TR/REC-xml/`. W3C Recommendation, 04 February 2004.

[XMLb] *Extensible Markup Language (XML) 1.1*. URL `http://www.w3.org/TR/xml11/`. W3C Recommendation, 04 February 2004, edited in place 15 April 2004.

[XPa] *XML Path Language (XPath) 2.0*. URL `http://www.w3.org/TR/xpath20/`. W3C Candidate Recommendation, 3 November 2005.

[Yer03] William S. Yerazunis. Sparse binary polynomial hashing and the CRM114 discriminator. In *2003 Spam Conference*. MIT, Cambridge, MA, 2003. URL `http://crm114.sourceforge.net/CRM114_paper.html`.

[Yer04] William S. Yerazunis. The spam-filtering accuracy plateau at 99.9% accuracy and how to get past it. In *2004 Spam Conference*. MIT, Cambridge, MA, 2004. URL `http://crm114.sourceforge.net/Plateau_Paper.pdf`.

[Zav03] Jakub Zavrel and Walter Daelemans. Feature-rich memory-based classification for shallow NLP and information extraction. In Jürgen Franke, Gholamreza Nakhaeizadeh, and Ingrid Renz, eds., *Text Mining, Theoretical Aspects and Applications*, pp. 33–54. Springer Physica, 2003. URL `http://cnts.uia.ac.be/cnts/ps/20040106.3653.zd03.pdf`.

[Zha03] Le Zhang and Tian shun Yao. Filtering junk mail with a maximum entropy model. In *Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03)*. 2003. URL `http://www.nlplab.cn/zhangle/paper/junk.pdf`.