

Text Mining for Biomedicine an Overview: selected bibliography

Sophia Ananiadou^a & Yoshimasa Tsuruoka^b
University of Manchester^a, University of Tokyo^b
National Centre for Text Mining^{a,b}
<http://www.nactem.ac.uk/>

(i) Overviews on Text Mining for Biomedicine

Ananiadou, S. & J. McNaught (eds) (2006) *Text Mining for Biology and Biomedicine*, Artech House.

MacMullen, W.J, and S.O. Denn, "Information Problems in Molecular Biology and Bioinformatics," *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 5, 2005, pp. 447--456.

Lars Juhl Jensen, J. Saric and P. Bork (2006) "Literature mining for the biologist: from information retrieval to biological discovery", In *Nature Reviews Genetics*, Vol. 7, Feb. 2006, pp 119-129

Blaschke, C., L. Hirschman, and A. Valencia, "Information Extraction in Molecular Biology," *Briefings in Bioinformatics*, Vol. 3, No. 2, 2002, pp. 1--12.

Cohen, A. M., and W. R. Hersh, "A Survey of Current Work in Biomedical Text Mining," *Briefings in Bioinformatics*, Vol. 6, 2005, pp. 57--71.

Nédellec, C., "Machine Learning for Information Extraction in Genomics---State of the Art and Perspectives." In *Text Mining and its Applications*, pp. 99--118, S. Sirmakessis (ed.), Berlin: Springer-Verlag, Studies in Fuzziness and Soft Computing 138, 2004.

Rebholz-Schuhmann, D., H. Kirsch, and F. Couto, "Facts from Text—Is Text Mining Ready to Deliver?" *PLoS Biology*, Vol. 3, No. 2, 2005, pp. 0188--0191, <http://www.plosbiology.org>, June 2005.

Shatkay, H., and R. Feldman, "Mining the Biomedical Literature in the Genomic Era: An Overview," *Journal of Computational Biology*, Vol. 10, No. 6, 2004, pp. 821--855.

Yandell, M. D., and W. H. Majoros, "Genomics and Natural Language Processing," *Nature Reviews/Genetics*, Vol. 3, 2002, pp. 601--610.

Mack, R., and M. Hehenburger, "Text-based Knowledge Discovery: Search and Mining of Life-sciences Documents," *Drug Discovery Today*, Vol. 7, No. 11, 2002, pp. S89--S98.

Hirschman, L., et al., "Accomplishments and Challenges in Literature Data Mining for Biology," *Bioinformatics*, Vol. 18, No. 12, 2002, pp. 1553--1561.

(ii) Forums for biomedical text mining on the Web

BLIMP <http://blimp.cs.queensu.ca/>

"BLIMP covers all publications related to the fast-growing field of biomedical literature and text mining. It is a one-stop resource, letting researchers find out who-does-what in the area and where it is published, bridging across the many discipline-specific venues in which biomedical text-mining papers are published."

BIONLP <http://www.ccs.neu.edu/home/futrelle/bionlp/>
Bob Futrelle's NLP for Biotext Mining

<http://www.text-mining.org/>

A comprehensive collection of text mining resources, including links to publications, commercial suppliers, news items, research groups, events, etc

(iii) Named Entity Recognition and Terminology Management

Ananiadou, S., Friedman, C. & Tsujii, J (eds) (2004) Named Entity Recognition in Biomedicine, Special Issue, *Journal of Biomedical Informatics*, vol. 37 (6).

Morgan, A., et al., "Gene Name Extraction Using FlyBase Resources", *Proceedings of ACL Workshop, NLP in Biomedicine*, Sapporo, Japan, 2003, pp.1--8.

Harkema, H., et al., "A Large-Scale Terminology Resource for Biomedical Text Processing", *Proceedings of BioLINK*, 2004, pp.53--60.

Liu, H., Y. Lussier, and C. Friedman, "A Study of Abbreviations in UMLS" *Proc. AMIA*, 2001, pp.393--397.

Bodenreider, O., J.A. Mitchell, and A.T. McCray, "Evaluation of the UMLS as a Terminology and Knowledge Resource for Biomedical Informatics", *Proc. AMIA*, 2002, pp.61--65.

Krauthammer, M., and G. Nenadic, "Term Identification in the Biomedical Literature", *Journal of Biomedical Informatics*, Special Issue on *Named Entity Recognition in Biomedicine*, Vol.37, No. 6, 2004, pp.512--526.

Hirschman, L., A. Morgan, and A.S. Yeh, "Rutabaga by Any Other Name: Extracting Biological Names". *Journal of Biomedical Informatics*, Vol.35, No. 4, 2002, pp.247--259.

Tsuruoka, Y., and J. Tsujii, "Probabilistic Term Variant Generator for Biomedical Terms," *Proceedings of 26th Annual ACM SIGIR Conference*, 2003, pp.167--173.

- Tsuruoka, Y. and J. Tsujii, "Improving the Performance of Dictionary-based Approaches in Protein Name Recognition" *Journal of Biomedical Informatics*, Special Issue on *Named Entity Recognition in Biomedicine*, Vol.37, No. 6, 2004, pp. 461--470.
- Krauthammer, M., A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for identifying gene and protein names in journal articles" *Gene*, Vol. 259, No. (1-2), 2001, pp.245--252.
- Gaizauskas, R., G. Demetriou, and K. Humphreys, "Term Recognition and Classification in Biological Science Journal Articles", *Proceedings of Workshop on Computational Terminology for Medical and Biological Applications*, Patras, Greece, 2000, pp.37--44.
- Fukuda, K., et al., "Towards Information Extraction: Identifying Protein Names from Biological Papers", *Proceedings of PSB*, Hawaii, USA, 1998, pp.707--718.
- Narayanaswamy, M., K.E. Ravikumar, and K. Vijay-Shanker, "A Biological Named Entity Recognizer", *Proceedings of PSB*, 2003, pp.427--438.
- Frantzen, K., et al., "Protein Names and How to Find Them". *Int J Med Inf*, Vol.67, No. (1-3), 2002, pp. 49--61.
- Collier, N., C. Nobata, and J. Tsujii, "Extracting the Names of Genes and Gene Products with a Hidden Markov Model" *Proceedings of COLING*, Saarbrücken, Germany, 2000, pp. 201--207.
- Kazama, J., T. Makino, Y. Ohta, and J. Tsujii, "Tuning Support Vector Machines for Biomedical Named Entity Recognition." *Proc. ACL Workshop NLP in the Biomedical Domain*, Philadelphia, USA, 2002, pp.1--8.
- Yamamoto, K., et al. "Protein Name Tagging for Biomedical Annotation in Text", *Proc. of ACL Workshop NLP in Biomedicine*, Sapporo, Japan, 2003, pp.65--72.
- Tanabe, L., and W.J. Wilbur, "Tagging Gene and Protein Names in Biomedical Text", *Bioinformatics*, 18(8), 2002, pp.1124--1132.
- Cohen, K.B., G.K. Acquaah-Mensah, A.E. Dolbey, and L. Hunter, "Contrast and Variability in Gene Names," *Proceedings of ACL Workshop on NLP in the Biomedical Domain*, Philadelphia, USA, 2002, pp.14--20.
- Tuason, O., L. Chen, H. Liu, J.A. Blake, and C. Friedman, "Biological Nomenclature: A Source of Lexical Knowledge and Ambiguity", *Proceedings of Pac Symp Biocomputing*, Hawaii, 2004, pp. 238--249.
- Nenadic, G., I. Spasic, and S. Ananiadou, "Mining Biomedical Abstracts: What's in a Term?" In *Natural Language Processing – IJCNLP 2004*, Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, et al (Eds.), LNCS vol. 3248, 2005, pp.797--806.

- Nenadic, G., S. Ananiadou, and J. McNaught, "Enhancing Automatic Term Recognition through Recognition of Variation", *Proceedings of COLING 2004*, Geneva, Switzerland, 2004, pp. 604--610.
- Tsuruoka, Y., S. Ananiadou, and J. Tsujii, "A Machine Learning Approach to Automatic Acronym Generation", *Proc. of Bio-LINK, ISMB*, 2005.
- Aronson, A.R., "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program," *Proceedings of AMIA*, 2001, pp.17--21.
- Yu, H., and E. Agichtein, "Extracting Synonymous Gene and Protein Terms from Biological Literature," *Bioinformatics*, 19, Suppl 1, 2003, pp.1340--349.
- Hatzivassiloglou, V., P.A. Duboue, and A. Rzhetsky, "Disambiguating Proteins, Genes, and RNA in Text: A Machine Language Approach" *Bioinformatics*, 17, Suppl 1, 2001, pp.97--106.
- Pakhomov, S., "Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts", *Proceedings of 40th ACL Conference*, 2002, pp.160--167.
- Liu, H., S.B. Johnson, and C. Friedman, "Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS", *J Am Med Inform Assoc*, Vol.9, No.6, 2002, pp. 621--636.
- Nenadic, G., I. Spasic, and S. Ananiadou, "Mining Term Similarities from Corpora", *Terminology*, Vol. 10, No.1, 2004, pp.55--80.
- Ogren, P., et al., "The Compositional Structure of Gene Ontology Terms", *Proc. PSB*, 2004, pp.214--225.
- Nobata, C., N. Collier, and J. Tsujii, "Automatic Term Identification and Classification in Biological Texts," *Proceedings of Natural Language Pacific Rim Symposium*, 1999, pp.369--374.
- Nenadic, G., H. Mima, I. Spasic, I. S. Ananiadou, and J. Tsujii, "Terminology-based Literature Mining and Knowledge Acquisition in Biomedicine", *International Journal of Medical Informatics*, Vol. 67, No.(1-3), 2002, pp.33--48.
- Nenadic, G., I. Spasic, and S. Ananiadou, "Terminology-Driven Mining of Biomedical Literature", *Bioinformatics*, Vol. 19, No.8, 2003, pp.938--943.
- Thelen, M., and E. Riloff, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts", *Proceedings of EMNLP*, 2002.
- Lee, K.J., et al., "Biomedical Named Entity Recognition using Two-Phase Model based on SVMs", *Journal of Biomedical Informatics*, Special Issue, *Named Entity Recognition in Biomedicine*, Vol.37, No. 6, 2004, pp. 436--447.

Takeuchi, K. and N. Collier, "Bio-medical Entity Extraction using Support Vector Machines," *Proceedings of ACL Workshop NLP in Biomedicine*, Sapporo, Japan, 2003, pp. 57--64.

Kim, J., et al., "Introduction to the Bio-Entity Recognition Task at JNLPBA," *Proc. Int. Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004, pp. 70--75.

Spasic, I., and S. Ananiadou, "A Flexible Measure of Contextual Similarity for Biomedical Terms" *Proceedings of Pacific Symposium on Biocomputing (PSB, 2005)*, Hawaii, USA, 2005.

Torii, M., S. Kamboj, and K. Vijay-Shanker, "An Investigation of Various Information Sources for Classifying Biological Names," *Proceedings of ACL Workshop NLP in Biomedicine*, Sapporo, Japan, 2003, pp.113--120.

Spasic, I., S. Ananiadou, and J. Tsujii, "MaSTerClass: a Case-based Reasoning System for the Classification of Biomedical Terms", *Bioinformatics*, Vol.21, No.11, 2005, pp.2748--2758.

Raychaudhuri, S., J.T. Chang, P.D. Sutphin, and R.B. Altman, "Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature," *Genome Res*, Vol.12, No.1, 2002, pp.203--214.

Koike, A., Y. Niwa, and T. Takagi, "Automatic Extraction of Gene/Protein Biological Functions from Biomedical Text", *Bioinformatics*, Vol. 21, No.7, 2005, pp.1227--1236.

Cantor M.N., et al., "An Evaluation of Hybrid Methods for Matching Biomedical Terminologies: Mapping the Gene Ontology to the UMLS", *Stud. Health Technol. Inform.* (95), 2003, pp.62--67

Mima, H., S. Ananiadou, and G. Nenadic, "The ATTRACT Workbench: Automatic Term Recognition and Clustering for Terms". In *Text, Speech and Dialogue* Matousek, V., Mautner, P., Moucek, R. and Tauser, K. (eds.), Lecture Notes in Artificial Intelligence 2166, Springer Verlag, Heidelberg, pp.126--133, 2001.

Nenadic, G., I. Spasic, and S. Ananiadou, "Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts". *Proc. of LREC*, 2002, pp. 2155--2162.

Chang, J.T., H. Schütze, and R.B. Altman, "Creating an Online Dictionary of Abbreviations from Medline," *Journal of the American Medical Informatics Association*, Vol. 9, No. 6, 2004, pp. 612--620.

Pustejovsky, J., et al., "Automatic Extraction of Acronym--Meaning Pairs from Medline Databases," *Medinfo*, Vol. 10, 2001, pp. 371--375.

Adar, E., "Sarad: a Simple and Robust Abbreviation Dictionary," *Bioinformatics*, Vol. 20, No. 4, 2004, pp. 527--533.

Schwartz, A.S. and M.A. Hearst, "A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text," *Pacific Symposium on Biocomputing*, 2003, pp. 451-462.

Yoshida, M., K. Fukuda, and T. Takagi, "Pnad-css: a Workbench for Constructing a Protein Name Abbreviation Dictionary," *Bioinformatics*, Vol. 16, 2000, pp. 169--75.

Liu, H., Y.A. Lussier, and C. Friedman, "A Study of Abbreviations in the UMLS," *Proceedings of the American Medical Informatics Association Annual Symposium*, 2001, pp. 393--397.

Pustejovsky, J., et al., "Medstrat: Creating Large-Scale Information Servers for Biomedical Libraries," *ACL Workshop on Natural Language Processing in the Biomedical Domain*, 2002, pp. 85—92.

Liu, H., S.B. Johnson, and C. Friedman, "Automatic Resolution of Ambiguous Terms based on Machine Learning and Conceptual Relations in the UMLS," *Journal of the American Medical Informatics Association*, Vol. 9, No. 4, 2002, pp. 621--636.

Pakhomov, S., "Supervised Maximum Entropy based Approach to Acronym and Abbreviation Normalization in Medical Texts," *Proceedings of the 40th annual meeting of the association for computational linguistics*, 2002, pp. 160—167.

Yu, Z., Y. Tsuruoka, and J. Tsujii, "Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense per Discourse Hypothesis," *SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics*, 2003, pp. 57--62.

Franzen, K., et al., "Protein Names and How to Find Them," *International Journal of Medical Informatics*, Vol. 67, No. 1--3, 2002, pp. 49--61.

Zhou, G., et al., "Recognizing Names in Biomedical Texts: A Machine Learning Approach," *Bioinformatics*, Vol. 20, No. 4, 2004, pp. 1178--1190.

Proux, D., et al., "Detecting Gene Symbols and Names in Biomedical Texts: A First Step toward Pertinent Information," *Proc. 9th Workshop on Genome Informatics*, 1998, pp. 72--80.

Morgan, A., et al., "Gene Name Identification and Normalization using a Model Organism Database," *Journal of Biomedical Informatics*, Vol. 37, 2004, pp. 396--410.

Mika, S., and B. Rost, "Protein Names Precisely Peeled off Free Text," *Bioinformatics*, Vol. 20, Suppl. 1, 2004, pp. i241--i247.

Hanisch, D., et al., "Playing Biology's Name Game: Identifying Protein Names in Scientific Text," *Proc. Pacific Symp. on Biocomputing*, 2003, pp. 403--414.

Chang, J. & Schutze, H. (2006) Abbreviations in Biomedical Text, in Text Mining for Biology and Biomedicine, Artech house, pp.99-119

Ananiadou, S. & Nenadic, G. (2006) Automatic Terminology Management in Biomedicine, in Text Mining for Biology and Biomedicine, Artech house, pp.67-98

Park, J.C. & Kim, J.J. (2006) Named Entity Recognition, in Text Mining for Biology and Biomedicine, Artech house, pp. 121- 142

(iv) Information Extraction

McNaught, J. & Black, W. (2006) Information Extraction, in Text Mining for Biology and Biomedicine, Ananiadou & McNaught (eds), Artech House, pp.143-178

Park, J.C., "Using Combinatory Categorical Grammar to Extract Biomedical Information," *IEEE Intelligent Systems*, Vol. 16, No. 1, 2001, pp.62--67.

Yakushiji, A., et al., "Event Extraction from Biomedical Papers Using a Full Parser," *Proc. Pacific Symp. on Biocomputing (PSB 2001)*, Kauai, Hawaii, Jan. 3--7, 2001, pp. 408--419.

Yakushiji, A., et al., "Biomedical Information Extraction with Predicate-Argument Structure Patterns," *Proc. Int. Symp. on Semantic Mining in Biomedicine*, 2005, pp. 60--69.

McDonald, D.M., et al., "Extracting Gene Pathway Relations Using a Hybrid Grammar: The Arizona Relation Parser," *Bioinformatics*, Vol. 20, No. 18, 2004, pp. 3370--3378.

Leroy, G., H. Chen, and J.D. Martinez, "A Shallow Parser based on Closed-Class Words to Capture Relations in Biomedical Text," *Journal of Biomedical Informatics*, Vol. 36, No. 3, 2003, pp.145--158.

Huang, M., et al., "Discovering Patterns to Extract Protein-Protein Interactions from Full Texts," *Bioinformatics*, Vol. 20, No. 18, pp. 3604--3612.

Hahn, U., and J. Wermter, "High-Performance Tagging on Medical Texts," *Proc. 20th Int.l Conf. on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, Aug. 23-27, 2004, pp. 973--979.

Hahn, U., M. Romacker, and S. Schulz, "Discourse Structures in Medical Reports--- Watch Out! The Generation of Referentially Coherent and Valid Text Knowledge Bases in the MEDSYNDIKATE System," *International Journal of Medical Informatics*, Vol. 53, No. 1, 1999, pp. 1--28.

Ono, T., et al., "Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics*, Vol. 17, No. 2, 2001, pp.155--161.

Ding, J., et al., "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser," *Proc. 15th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI 2003)*, Sacramento, CA, Nov. 3--5, 2003, pp. 467--473.

Alphonse, E., et al., "Event-based Information Extraction for the Biomedical Domain:The Caderige Project," *Proc. Int. Workshop on Natural language Processing in Biomedicine and its Applications (JNLPBA)*, 2004, pp. 43--49.

Corney, D. P., et al., "BioRAT: Extracting Biological Information from Full-length Papers," *Bioinformatics*, Vol. 20, No. 17, 2004, pp. 3206--3213.

Daraselia, N., et al., "Extracting Human Protein Interactions from MEDLINE using a Full-sentence Parser," *Bioinformatics*, Vol. 20, No. 5, 2004, pp. 604--611.

De Bruijn, B., and J. Martin, "Getting to the (C)ore of Knowledge: Mining Biomedical Literature," *International Journal of Medical Informatics*, Vol. 67, 2002, pp. 7--18.

Friedman, C., et al., "GENIES: A Natural-language Processing System for the Extraction of Molecular Pathways from Journal Articles," *Bioinformatics*, Vol. 17, 2001, pp. S74--S82.

Gaizauskas, R., et al., "Protein Structures and Information Extraction from Biological Texts: The PASTA System," *Bioinformatics*, Vol. 19, No. 1, 2003, pp. 135--143.

Hu, Z., et al., "Literature Mining and Database Annotation of Protein Phosphorylation Using a Rule-based System," *Bioinformatics*, Vol. 21, 2005, 2759--2765.

Huang, M., et al., "Discovering Patterns to Extract Protein-Protein Interactions from Full Biomedical Texts," *Bioinformatics*, Vol. 20, 2004, pp. 3604--3612.

Kim, J., and J. C. Park, "BioIE: Retargetable Information Extraction and Ontological Annotation of Biological Interactions from the Literature," *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 3, 2004, pp. 551--568.

Leroy, G., and H. Chen, "Filling Preposition-based Templates to Capture Information from Medical Abstracts," *Proc. Pacific Symp. on Biocomputing 7 (PSB)*, 2002, pp. 362--373.

Leroy, G., and H. Chen, "Genescene: An Ontology-enhanced Integration of Linguistic and Co-occurrence Based Relations in Biomedical Texts," *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 5, 2005, pp. 457--468.

Leroy, G., H. Chen, and J. D. Martinez, "A Shallow Parser based on Closed-class Words to Capture Relations in Biomedical text," *Journal of Biomedical Informatics*, Vol. 36, 2003, pp. 145--158.

- Novichkova, S., S. Egorov, and N. Daroselia, "MedScan, a Natural Language Processing Engine for MEDLINE Abstracts," *Bioinformatics*, Vol. 19, 2003, pp. 1699--1706.
- Ono, T., et al., "Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics*, Vol. 17, No. 2, 2001, pp. 155--161.
- Pyysalo, S., et al., "Analysis of Link Grammar on Biomedical Dependency Corpus Targeted at Protein-Protein Interactions," *Proc. Int. Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, 2004, pp. 15--21.
- Rinaldi, F., et al., "Mining Relations in the GENIA Corpus," *Proc. Second European Workshop on Data Mining and Text Mining for Bioinformatics*, 2004, pp. 61--68.
- Rzhetsky, A., et al., "GeneWays: A System for Extracting, Analyzing, Visualizing and Integrating Molecular Pathway Data," *Journal of Biomedical Informatics*, Vol. 37, No. 1, 2004, pp. 43--53.
- Šarić, J., L. J. Jensen, and I. Rojas, "Large-scale Extraction of Gene Regulation for Model Organisms in an Ontological Context," *In Silico Biology*, Vol. 5, No. 0004, 2004, <http://www.bioinfo.de.isb/2004/05/0004/>, June 2005
- Chun, H., et al., "Extraction of Gene-Disease Relations from MedLine using Domain Dictionaries and Machine Learning," *Proc. Pacific Symp. on Biocomputing (PSB)*, 2006, pp. 4--15.
- Ramani, Al, et al, "Using Biomedical Literature Mining to Consolidate the Set of Known Human Protein-Protein Interactions," *Proc. ACL-ISMB Workshop on Linking Biological Literature*, 2005, pp. 46-53.
- Bunescu, R. et al., "Comparative Experiments on Learning Information Extractors for Proteins and their Interactions," *Journal Artificial Intelligence in Medicine (Special Issue on Summarization and Information Extraction from Medical Documents)*, 2004.

(v) Annotation of Biomedical Corpora

- Kim, J.D. & Tsujii, J. (2006) Corpora and their Annotation, in *Text Mining for Biology and Biomedicine*, Ananiadou, S. & McNaught, J. (eds), Artech House, pp. 179-211.
- Cohen, K., et al., "Corpus Design for Biomedical Natural Language Processing," *Proc. ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005, pp. 38--45.
- Kim, J., et al., "GENIA Corpus---a Semantically Annotated Corpus for Bio-Textmining," *Bioinformatics*, Vol. 19, Suppl. 1, 2003, pp. i180--i182.

Tanabe, L., et al., "GENETAG: a Tagged Corpus for Gene/Protein Named Entity Recognition," *BMC Bioinformatics*, Vol. 6, Suppl. 1, S3.

Smith, L.H., et al., "MedTag: a Collection of Biomedical Annotations," *Proc. ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005, pp. 32--37.

Tateisi, Y., and J. Tsujii, "Part-of-Speech Annotation of Biology Research Abstracts," *Proc. Int. Conf. on Language Resource and Evaluation (LREC)*, 2004, pp. 1267--1270.

Tateisi, Y., et al., "Syntax Annotation for the GENIA corpus," *Proc. IJCNLP Companion volume*, 2005, pp. 222--227.

Pakhomov, S., A. Coden, and C. Chute, "Creating a Test Corpus of Clinical Notes Manually Tagged for Part-of-Speech Information," *Proc. COLING Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004, pp. 62--65.

Erjavec, T., et al., "Encoding Biomedical Resources in TEI: The Case of the GENIA Corpus," *Proc. ACL Workshop on Natural Language Processing in Biomedicine*, 2003, pp. 97--104.

(vi) Tagging and Parsing

Benett, N., et al., "Extracting noun phrases for all of MEDLINE," *Proc. AMIA Symp.*, 1999, pp. 671-675

Bod, R., "An Efficient Implementation of a New DOP Model," *Proc. EACL*, 2003.

Brants, T., "TnT - A Statistical Part-of-Speech Tagger," *Proc. Applied Natural Language Processing Conf.*, 2000.

Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," *Computational Linguistics*, 1995

Carreras, X., and L. Marquez, "Phrase recognition by filtering and ranking with perceptrons," *Proc. RANLP*, 2003.

Charniak, E., "A Maximum-Entropy-Inspired Parser," *Proc. NAACL*, 2000.

Charniak, E., and M. Johnson, "Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking," *Proc. ACL*, 2005.

Clegg, A., and A. Shepherd, "Evaluating and integrating treebank parsers on a biomedical corpus," *Proc. ACL Workshop on Software*, 2005.

Collins, M., "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," *Proc. EMNLP*, 2002, pp. 1-8.

- Collins, M., "Head-Driven Statistical Models for Natural Language Parsing," Ph.D. thesis, University of Pennsylvania, 1999.
- Giménez, J., and L. Màrquez, "Fast and accurate part-of-speech tagging: The SVM approach revisited," *Proc. RANLP*, 2003, pp. 153-163.
- Hara, T., Y. Miyao and J. Tsujii, "Adapting a probabilistic disambiguation model of an HPSG parser to a new domain," *Proc. Int. Joint Conf. on Natural Language Processing*, 2005
- Henderson, J., "Discriminative training of a neural network statistical parser," *Proc. ACL*, 2004.
- Kudoh, T., and Y. Matsumoto, "Chunking with support vector machines," *Proc. NAACL*, 2001
- Kudoh, T., and Y. Matsumoto, "Use of support vector learning for chunk identification," *Proc. CoNLL*, 2000, pp. 142-144
- Lease, M., and E. Charniak, "Parsing Biomedical Literature," *Proc. IJCNLP*, 2005.
- Miyao, Y., et al., "Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank," *Proc. IJCNLP*, 2004.
- Miyao, Y., and J. Tsujii, "Probabilistic disambiguation models for wide-coverage HPSG parsing," *Proc. ACL*, 2005, pp. 83--90.
- Reynar, J., and A. Ratnaparkhi, "A Maximum Entropy Approach to Identifying Sentence Boundaries," *Proc. ANLP*, 1997, pp. 16-19.
- Smith, L., et al., "MedPost: a part-of-speech tagger for bioMedical text," *Bioinformatics*, Vol. 20, No. 14, 2004, pp. 2320-2321
- Toutanova K., et al., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," *Proc. HLT-NAACL*, 2003.
- Tsuruoka, Y., and J. Tsujii, "Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data," *Proc. HLT/EMNLP*, 2005. pp. 467--474.
- Tsuruoka, Y., et al., "Developing a Robust Part-of-Speech Tagger for Biomedical Text," *Proc. 10th Panhellenic Conference on Informatics*, 2005, pp. 382-392.
- Zhang, T., et al., "Text chunking based on a generalization of winnow," *Journal of Machine Learning Research*, Vol. 2, 2002, pp. 615-638