RESOLVING AMBIGUITY IN ENTITY AND FACT EXTRACTION THROUGH A HYBRID APPROACH

HEJAB MA'AZER KHALED AL FAWAREH

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
JULY 2010

RESOLVING AMBIGUITY IN ENTITY AND FACT EXTRACTION THROUGH A HYBRID APPROACH

A Thesis submitted to the College of Arts and Sciences in full fulfillment of the requirements for the degree of Doctor of Philosophy
Universiti Utara Malaysia

By
Hejab Ma'azer Khaled Al Fawareh

Kolej Sastera dan Sains
(UUM College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

HEJAB MA'AZER KHALED AL FAWAREH

calon untuk Ijazah **PhD**
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"RESOLVING AMBIQUITY IN ENTITY AND FACT EXTRACTION THROUGH A HYBRID APPROACH"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***25 Ogos 2009.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
August 25, 2009.

Pengerusi Viva: *(Chairman for Viva)*	**Prof. Dr. Mohd. Zaini Abdul Karim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Normaziah Abdul Aziz**	Tandatangan *(Signature)*
Pemeriksa *Dalam*: *(Internal Examiner)*	**Dr. Faudziah Ahmad**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof Dr. Wan Rozaini Sheik Osman**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof Dr. Norita Mohd Norwawi**	Tandatangan *(Signature)*

Tarikh:
(Date) ***August 25, 2009***

PERMISSION TO USE

In presenting this thesis in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of Research and Graduate Studies. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean of Research and Graduate Studies

College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman.

ABSTRAK

Tesis ini memperkenalkan dua teknik baru yang dikenali sebagai *Unambiguous Entity Extraction (UEE)* dan *Unambiguous Fact Extraction (UFE)* untuk menyelesaikan kesamaran di dalam cabutan entiti dan fakta. Kedua-dua teknik diperolehi dengan menggabungkan 4 teori dan kaedah yang utama yang dikenali sebagai teori kemungkinan, set fuzi, kaedah pengkalan pengetahuan dan teknik-teknik pemprosesan bahasa tabii (pemprosesan sintak dan semantik). Di dalam tesis ini hanya perkataan yang tergolong di dalam kata nama sahaja yang disifatkan sebagai entiti. Kesamaran entiti berlaku apabila sesuatu perkataan mempunyai lebih dari satu semantik. Teknik UEE direka dan dibangunkan untuk menentukan semantik yang paling berkemungkinan bagi sesuatu perkataan. Teknik UEE telah diuji dengan menggunakan 12 kes ujian yang mengandungi 111 ayat. Hasil kajian yang diperolehi menunjukkan teknik UEE boleh memberi kadar *precision* sebanyak 85.7% dan kadar *recall* sebanyak 80.3%. Manakala teknik UFE pula menfokuskan kepada cabutan fakta dari sesuatu ayat. Fakta di sini bermaksud pernyataan yang dapat dipersembahkan secara formal dan ditentukan kebenarannya. Teknik UFE direka dan dibangunkan untuk memilih fakta yang paling berkemungkinan dengan memilih fakta yang paling mungkin bagi sesuatu ayat. Di dalam menilai teknik UFE, kes-kes ujian telah dijana dan diuji. Setiap kes ujian mengandungi ayat di dalam ujuran dari 5 ke 8. Hasil kajian yang diperolehi di dalam bentuk kalkulus predikat di nilai secara manual. Hasil keputusan menunjukkan teknik UFE berjaya mencabut fakta yang paling berkemungkinan.

ABSTRACT

This thesis presents two new techniques namely Unambiguous Entity Extraction (UEE) and Unambiguous Fact Extraction (UFE) to resolve ambiguity in entity and fact extraction. Both techniques are obtained by hybriding 4 major theories and approaches namely, possibility theory, fuzzy sets, a knowledge-based approach, NLP techniques (syntactic and semantic processing). In this thesis, a word that is classified into to Noun part-of-speech is considered as an entity. An entity is ambiguous if it has more than one semantic. The UEE technique is designed and developed to assign the most possible semantic to the word. The technique was tested using 12 test cases with 111 sentences. The obtained results indicate that UEE technique is able to give precision rate of 85.7% and recall rate of 80.3%. On the other hands, UFE focuses on extracting an ambiguous fact from a sentence. A fact is a meaning that can be formally represented as a statement and determined its truthfulness. A sentence contains an ambiguous fact if it can be interpreted into more than one meaning. The UFE technique is designed and developed to select the most possible fact by selecting the most possible meaning from a sentence. In evaluating UFE technique, test cases have been created and tested. Each test case consists of sentences in the range of 5 to 8. The obtained results in the form of predicate calculus are evaluated manually. The results suggest that UFE technique is successful.

ACKNOWLEDGMENTS

First of all, I would like to thank to Allah Subhanahhu Wataalaa, for guiding me in my PhD research work. Without HIM this work is impossible. There are many people I would like to thank. First, to my both supervisors Associate Prof. Dr. Wan Rozaini S. Osman and Associate Prof. Dr. Norita Md Norwawi for their undivided support that they gave to me during the PhD. My special thanks to my mother, father and brothers for supporting my studies and praying for my success. To my sister in law thank you so much for always giving positive encouragement during my PhD studies. To my brothers in laws, thank you for making my life in Malaysia happy. To my parent in law, I owe them for their sincere love. The last but not the least, I would like to give millions of thanks to my wife Dr. Shaidah Jusoh, for giving me two beautiful children (Seeta and Abdulrahman), and absorbing all the stress during 4 years of my PhD study. She is the person who make me stand on my own feet and walk through in the midst of uncertainty, and give me undivided love and attention to make this dream comes true.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter consists of six sections; research motivation, problem statement, research objectives, scope of the research, contribution of the research, and thesis organization.

1.1 Research Motivation

Texts have become the major medium for communication in English language as well as other languages such as Malay, Arabic, French and etc. It is a medium for communication, knowledge accumulation, and information distribution (McCallum, 2005). About 80% of the valuable information is stored in text documents. These text documents can be found on personal desktop computers, intranets and in the World Wide Web (Web). The valuable information is normally embedded inside unstructured texts. The world's largest text repository is World Wide Web (WWW). The Web is being constantly augmented and maintained by millions of people around the world. Extracting the valuable information requires reading. However, this task is very time consuming. Having an automated system that can extract facts from the texts, stored them in a database, and then use machine learning algorithms to manipulate those facts is indeed necessary.

The contents of the thesis is for internal user only

BIBLIOGRAPHY

Aasa, J. (2004). *Unsupervised resolution of PP attachment ambiguities in Swedish.* Master's thesis, Stockholm University.

Abney, S., Schapire, R. E., & Singer, Y. (1999). Boosting applied to tagging and PP attachment. In *Proceedings of the EMNLP/VLC-99*, (pp. 38–45).

Agirre, E., & Edmonds, P. (2007). Introduction. In E. Agirre, & P. Edmonds (Eds.) *Word Sense Disambiguation: Algorithms and Applications*, (pp. 1–28). New York : Springer Verlag.

Agirre, E., & Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the Semantic Annotation and Intelligent Annotation workshop*, (pp. 11–19).

Agirre, E., & Stevenson, M. (2007). *Knowledge sources for WSD*, (pp. 217–251). New York: Springer Verlag.

Alam, Y. S. (2004). Decision trees for sense disambiguation of prepositions: Case of over. In *Proceedings of the Workshop on Computational Lexical Semantics*, (pp. 52–59).

Alegre, M. A., Sopena, J. M., & Lloberas, A. (1999). PP-attachment: A committee machine approach. In *Proceedings of the Joint SIGDAT Conference on*

Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC- 99). (pp. 231–238).

Alfawareh, H. M., & Jusoh, S. (2009). Extracting semantic for information extraction. In *Proceedings of the Second International Conference on Computing and Informatics*. Kuala Lumpur, Malaysia: Universiti Utara Malaysia.

Alfawareh, H. M., Jusoh, S., & Norwawi, N. M. (2008a). Resolving ambiguity/ uncertainty in fact extraction. In *Proceedings of the 21st International CODATA Conference on Scientific Information for Society from Today to the Future* (pp. 343–349). Kyiv, Ukraine.

Alfawareh, H. M., Jusoh, S., & Osman, R. S. (2008b). Ambiguity in text mining. In *Proceedings of the 2nd International Conference on Computer and Communication Engineering (ICCCE'08)*, (pp. 1172–1176).

Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st International Conference on General WordNet*, (pp. 466–471).

Allen, J. (1988). *Natural Language Understanding*. United States of America:

Appelt, D., Jerry, E., Hobbs, R., Bear, J., Israel, D., Kameyama, M., & Tyson, M. (1993). Fastus: A finite-state processor for information extraction from realworld text. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, (pp. 1172–1181).

Asahara, M., & Matsumoto, Y. (2003). Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the North American Chapter of the Association for Computational Linguistics- Human Language Technology Conference*.

Baker, K. L., Franz, A. M., Franz, E. M., & Jordan, P. W. (1994). Coping with ambiguity in knowledge-based natural language analysis. In *Proceedings of FLAIRS*.

Baldwin, T., Kordoni, V., & Villavicencio, A. (2009). Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistic*, *35* (2), 119–149.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial intelligence*, (pp. 805–810).

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the Web. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-07)*, (pp. 2670–2676).

Bastian, A. (1994). How to handle the flexibility of linguistic variables with applications. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, *2* (4), 463–484.

Behrang, M., & Srini, N. (2003). Semantic extraction with wide-coverage lexical resources. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, (pp. 64–66). Morristown, NJ, USA: Association for Computational Linguistics.

Benamara, F. (2005). Reasoning with prepositions within a cooperative questionanswering framework. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, (pp. 145–152).*

Bick, E. (2004). A named entity recognizer for Danish. In *Proceedings of the Conferenceon Language Resources and Evaluation.*

Bien, Z., & Song, W. (2003). Blend of soft computing techniques for effective human machine interaction in service robotic system. *Fuzzy Sets and Systems*, *134*(1).

Bird, D. L., & Munoz, C. (1983). Automatic generation of random self-checking test cases. *IBM System Journal*, *22* (3), 229–245.

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science*, (4), 427–455.

Bodenreider, O., & Zweigenbaum, P. (2000). Identifying proper names in parallel medical terminologies. *Stud Health Technol Inform*, *77* , 443–447.

Boonthum, C., Toida, S., & Levinstein, I. (2005). Sense disambiguation for preposition 'with'. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, (pp. 153–172).

Boonthum, C., Toida, S., & Levinstein, I. (2006). Preposition senses: Generalized disambiguation model. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006)*, (pp. 196–207).

Botterill, G., & Carruthers, P. (1999). *The philosophy of psychology*. Cambridge, UK: Cambridge University Press.

Brown, P., Stephen, E., Pietra, D., Vincent, J., Pietra, D., & Mercer, R. L. (1991). Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting for Computational Linguistics*, (pp. 264–270).

Bruce, R., & Weibe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 139–145).

Burnard, L. (2000). *Reference Guide for the British National Corpus*. Oxford, UK: Oxford University Computing Services.

Calliff, M. E., & Mooney, R. J. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference onArtificial Intelligence*, (pp. 328–334).

Calvo, H., Gelbukh, A., & Kilgarriff, A. (2005). Distributional thesaurus versus wordnet: A comparison of backoff techniques for unsupervised pp attachment. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 177–188).

Cayrac, D., Dubois, D., & Prade, H. (1996). Handling uncertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application. *IEEE Transactions on Fuzzy Systems*, *4* (3), 251–269.

Chang, C.-H., Kayed, M., Girgis, M. R., & Shaalan, K. (2006). A survey of web information extraction systems. *IEEE Transactions of Knowledge and Data Engineering*, *18* (10), 1411–1428.

Chang, C. H., & Kup, S.-C. (2004). A semi-supervised approach of web data extraction with visual support. *Intelligent System*, *19* (6), 56–64.

Charniak, U. (2001). Unsupervised learning of name structure from coreference data. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, (pp. 48–54).

Chiara, C. M., Fernando, F., & Patrizia, G. (2008). Ambiguity detection in multimodal systems. In *Proceedings of the Working Conference on Advanced visual interfaces*, (pp. 331–334). New York, NY, USA: ACM.

Chodorow, M., Tetreault, J., & N.Han (2007). Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, (pp. 25–30).

Chu, C.-T., Sung, Y.-H., Yuan, Z., & Jurafsky, D. (2006). Detection of word fragments in mandarin telephone conversation. In *International Conference on Spoken Language Processing*. URL pubs/fragment-icslp-06.pdf

Church, K., & Ramesh, R. (1982). Coping with syntactic ambiguity or how to put the block in the box on the table. *Computational Linguistics*, *8* (3-4), 139–149.

Collins, M., & Brooks, J. (1995). Prepositional phrase attachment through a backedoff model. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, (pp. 27–38).

Collobert, R., & Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (pp. 560–567).

Cowie, J., & Wilks, Y. (2000). *Information extraction*. New York.

Cucchiarelli, A., & Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. In *Proceedings of the Computational Linguistics*, (pp. 123–131).

Cucerzan, S., & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In

Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Dedek, J. (2008). Extraction of semantic information from Web resources. In *Proceedings of Contributed Papers, Part I,* (pp. 224–229).

Etizioni, O., Cafarella, M. M., Downey, S., Kok, A. M., Popescu, T., Shaked, S., Soderl, D., Weld, S., & Yates, A. (2004). Web-scale information extraction in know it all. In *Proceedings of the Thirteenth International World Wide Web Conference.*

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named-entity extraction from the Web: an experimental study. *Artificial Intelligence, 165* (1), 91–134.

Evsukoff, A., S.Gentil, & Montmain, J. (2000). Fuzzy reasoning in coperative supervision systems. *Control Engineering, 8* (4), 389–407.

Feldman, R., & Sanger, J. (2007). *The text mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* United State of America: Cambridge University Press.

Firebaugh, M. W. (1998). *Artificial Intelligence: A Knowledge-based Approach.* United States of America: Boyd & Fraser.

Fleischman, M. (2001). Automated subcategorization of named entities. In *39th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop.*

Fleischman, M., & Hovy, E. (2002). Fine grained classification of named entities. In *Proceeding of the 19th International Conference on Computational Linguistics (COLING).*

Franz, A. (1996). Learning PP attachment from corpus statistics. In S. Wermter, E. Riloff, & G. Scheler (Eds.) *Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, (pp. 188–202). Berlin, Germany: Springer.

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. thesis, University of Connecticut.

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, (pp. 236–243).

Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence*, (pp. 1048–1053).

Gaizauskas, R., Demetriou, G., Artymiuk, P. J., & Willett, P. (2003). Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, *19* (1), 135–143.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., & Wilks, Y. (1996). Description of the LASIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, (pp. 207–220).

Gala, N., & Lafourcade, M. (2006). PP attachment ambiguity resolution with corpusbased pattern distributions and lexical signatures. *ECTI Transactions on Computerand Information Technology*, *2* (2), 116–120.

Gardent, C., Guillaume, B., Perrier, G., & Falk, I. (2005). Maurice gross' grammar lexicon and natural language processing. In *Proceedings of the 2nd Language and Technology Conference*. Pozan, Poland.

Gaussier, E., & Cancedda, N. (2001). Probabilistic models for PP-attachment resolution and NP analysis. In *Proceedings of the ACL/EACL-2001 Workshop on Computational Natural Language Learning (CoNLL-2001)*, (pp. 1–8).

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*, 245–288.

Grenager, T., Klein, D., & Manning, C. D. (2005). Unsupervised learning of field segmentation models for information extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (pp. 371–378).

Grishman, R. (1997). Information extraction: Techniques and challenges. In *Proceedings of the SCIE*, (pp. 207–220).

Grishman, R. (2005). NLP: An information extraction perspective. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2005)*.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, (pp. 466–471).

Guang, Q., Kangmiao, L., Jiajun, B., Chun, C., & Zhiming, K. (2007). Quantify query ambiguity using ODP metadata. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 697–698). New York, NY, USA: ACM.

Hale, R. (2005). Text mining: Getting more value from literature resources. *Drug Discovery Today, 10* (6), 377–379.

Hansen, S. L. (2005). Concept-based representation of prepositions. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of*

Prepositions and their Use in Computational Linguistics Formalisms and Applications, (pp. 138–144).

Harabagiu, S., Bejan, C. A., & Morrescu, P. (2005). Shallow semantics for relation extraction. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI '05)*, (pp. 1061–1066).

Hartrumpf, S. (1999). Hybrid disambiguation of prepositional phrase attachment and interpretation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC- 99)*, (pp. 111–120).

Hartrumpf, S. (2002). *Hybrid Disambiguation in Natural Language Analysis*. Ph.D. thesis, FernUniversit¨at, Hagen, Fachbereich Informatik, Hagen, Germany.

Hauptmann, A. G., Chen, M. Y., Christel, M., Lin, W., & Yang, J. (2007). A hybrid approach to improving semantic extraction of news video. In *Proceedings of the International Conference on Semantic Computing (ICSC)*, (pp. 79–86).

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, (pp. 3–10).

Herwijnen, V., Olga, J., Terken, A., Bosch, V., & E.Marsi (2003). Learning PP attachment for filtering prosodic phrasing. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL 2003)*, (pp. 139-146).

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19* (1), 103–120.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. (1997). FASTUS: A cascaded finite-state transducer for extraction

information from natural language text. In E. Roche, & Y. Schabes (Eds.) *Finite States Devices for Natural Language Processing*, (pp. 383–406).

Huang, C.-Y., Chang, J.-R., & Chang, Y.-H. (2010). Design and analysis of GUI testcase prioritization using weight-based methods. *Journal System Software*, *83* (4), 646–659.

Huang, F. (2005). *Multilingual Named Entity Extraction and Translation from Text and Speech*. Ph.D. thesis, Carnegie Mellon University.

Humphreys, K., Demetriou, G., & Gaizauskas, R. (2000). Bioinformatics applications of information extraction from scientific journal articles. *Journal of Information Science*, *26* (2), 75–85.

James, A., & Hema, R. (2002). Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 307–314). New York, NY, USA: ACM.

Jeffrey, D., & Gupta, N. (2008). Experiments with test case prioritization using relevant slices. *Journal System Software*, *81* (2), 196–221.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*.

Jones, R., Ghani, R., Mitchell, T., & Riloff, E. (2003). Active learning for information extraction with multiple view feature sets. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing*. United States of America: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural language Processing, Computational Linguistics and Speech Recognition*. United States of America: Prentice Hall.

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, (pp. 638–646). Morristown, NJ, USA: Association for Computational Linguistics.

Jusoh, S., & Alfawareh, H. M. (2007). Natural language interface for online sales. In *Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007)*, (pp. 224–228). Malaysia: IEEE.

Jusoh, S., & Alfawareh, H. M. (2009). Semantic extraction from texts. In *Proceedings of the 2009 International Conference on Computer Engineering and Applications*, (pp. 595–601). IACSIT, Manila, Phillipphines: World Academic Union.

Kamsties, E. (2001). *Surfacing Ambiguity in Natural Language Requirements*. Phd thesis, Fraunhofer- Institue fr Experimentelles Software Engineering.

Karanikas, H., Tjortjis, C., & Theodoulidis, B. (2000). An approach to text mining using information extraction. In *Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference*.

Karat, C., Vergo, J., & Nahamoo, D. (2003). Conversational interface technologies. In J. A. Jacko, & A. Sears (Eds.) *The Human-Computer Interaction Handbook*, (pp. 169–186). Lawrence Erlbaum Associates.

Kies, D. (2009 (accessed Jan 23, 2009)). *Modern English Grammar*. http://papyr.com/hypertextbooks/grammar/.

Kirsten, B., & Jeffrey, H. (2006). Advancing ambiguity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, (pp. 103– 106). New York, NY, USA: ACM.

Klir, G. J., & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. United States of America: Prentice Hall.

Kokkinakis, D. (2000). Supervised PP-attachment for Swedish: Combining unsupervised and supervised training data. *Nordic Journal of Linguistics*, *3* (2), 191–213.

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the EACL 2006 Student Research Workshop*, (pp. 15–21).

Lassen, T. (2006). An ontology-based view of prepositional senses. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, (pp. 45–50). Trento.

Leacock, C., Towell, G., & Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of the Workshop on Human Language Technology*, (pp. 260–265).

Lee, S., & Lee, G. (2005). Heuristic methods for reducing errors of geographic named entities learned by bootstrapping. In *Proceeding of the International Joint Conference on Natural Language Processing*.

Leroy, G., & Chen, H. (2002). Distributional thesaurus versus WordNet: A comparison of backoff techniques for unsupervised PP attachment. In *Proceedings of the Pacific Symposium on Biocomputing*, (pp. 350–361).

Li, H.-F., Heo, N., Moon, K., Lee, J.-H., & Lee, G. (2002). Lexical transfer ambiguity resolution using automatically-extracted concept co-occurrence information. *International Journal of Computer Processing of Languages (IJCPOL), 13* (1), 53–68.

Lin, D., & Pantel, P. (2001). Induction of semantic classes from natural language text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (pp. 317 – 322).

Lindstromberg, S. (2001). Preposition entries in UK monolingual learners dictionaries: Problems and possible solutions. *Applied Linguistics, 22* (1), 79–103.

Lisa, B., & Bruce, C. W. (1998). Resolving ambiguity for cross-language retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 64–71). New York, NY, USA: ACM.

Liu, H., Hu, Z., Torii, M., Wu, C., & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *Journal of the American Medical Informatics Associations (JAMIA), 13*, 497–507.

Liu, H., Johnson, S. B., & Friedman, C. (2002). Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *Journal of the American Medical Informatics Associations (JAMIA), 9*, 621–636.

Luger, G. F. (2002). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. England: Addison Wesley.

Malik, R. (2006). *CONAN: Text Mining in Biomedical domain*. Phd thesis, Utrecht University, Austria.

May, J., Brunstein, A., Natarajan, P., & Weischedel, R. (2003). Surprise! what's in a cebuano or Hindi name? *ACM Transactions on Asian Language Information Processing (TALIP)*, *2* (3), 169–180.

Maynard, D., Bontcheva, K., & Hamish, C. (2003). Towards a semantic extraction of named entities. In *Proceedings of Recent Advances in Natural Language*. Bulgaria.

McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. *Queue*, *3*, 48–57.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. In *Proceedings of the Conference on Computational Natural Language Learning*.

McCarthy, D., Carroll, J., & Preiss, J. (2001). Disambiguating noun and verb senses using automatically acquired selectional preferences. In *Proceedings of the SENSEVAL-2 Workshop at the European Chapter ACL*, (pp. 119–122). Toulouse, France.

Merlo, P., Crocker, M.W., & Berthouzoz, C. (1997). Attaching multiple prepositional phrases: Generalized backed-off estimation. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, (pp. 149– 155).

Mihalcea, R., & Moldovan, D. (2001). A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal of Artificial Intelligence Tools*, *10* (1-2), 5–21.

Milne, R. (1980). Parsing against lexical ambiguity. In *Proceedings of the 8th conferenceon Computational linguistics*, (pp. 350–353). Morristown, NJ, USA.

Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter*, *7* (1), 3–10.

Mutlum, B. (2005). *Word Sense Disambiguation Based on Sense Similarity and Syntactic Context*. Master's thesis, School of Engineering, Koc University, Turkey.

Mutton, P., & Golbeck, J. (2003). Visualization of semantic metadata and ontologies. In *Proceedings of the Seventh International Conference on Information Visualization*, (pp. 300–305).

Nadas, A., Nahamoo, D., Picheny, M., & Powell, J. (1991). An iterative approximation of the most informative split in the construction of decision trees. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 565–568). Toronto.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition:Learning to Recognize 100 Entity Types with Little Supervision*. Ph.D. thesis, University of Ottawa.

Nadh, K., & Christian, H. (2009). Prepositional phrase attachment ambiguity resolution using semantic hierarchies. In *Proceedings of the Ninth IASTED*

International Conference on Artificial Intelligence and Applications, (pp. 73–80).

Nakayama, K., Hara, T., & Nishio, S. (2008). Wikipedia link structure and text mining for semantic relation extraction. In S. Bloehdorn, M. Grobelnik, P. Mika, & D. T. Tran (Eds.) *Proceedings of CEUR Workshop*, vol. 334.

Navigli, R. (2008). A structural approach to the automatic adjudication of word sense disagreements. *Journal of Natural Language Engineering*, *14* (4), 293–310.

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, *41* (2), 1–69.

Navigli, R., & Velardi, P. (2002). Automatic adaptation of WordNet to domains. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: a knowledgebased approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27* (7), 1075–1088.

Nedellec, C., & Nazarenko, A. (2005). Ontologies and information extraction: A necessary symbiosis. In P. Buitelaar, P. Comiano, & B. Magnin (Eds.) *Ontology Learning from Text: Methods, Evaluation and Applications*.

Negnevitsky, M. (2004). *Artificial Intelligence: A Guide to Intellligent System*. England: Addison Wesley.

Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, *17* (2), 155–161.

Pan, J. (1999). Software testing. URL http://www.ece.cmu.edu/ koopman/

Pedrycz, W. (1989). *Fuzzy Control and Fuzzy Systems*. John Wiley and Son Inc.

Piskorski, J. (2004). Named-entity recognition for Polish with SProUT. In L. Bolc, Z. Michalewicz, & T. Nishida (Eds.) *Lecture Notes in Computer Science*, vol. 3490, (pp. 122–133).

Portet, F., Reiter, E., Hunter, J., & Sripada, S. (2007). Automatic generation of textual summaries from neonatal intensive care data riccardo bellazzi. In A. Abu-Hanna, & J. Hunter (Eds.) *11th Conference on Artificial Intelligence in Medicine (AIME 07)*, (pp. 227–236). Springer-Verlag.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). Topic modeling for the social sciences. In *Workshop on Applications for Topic Models: Text and Beyond (NIPS 2009)*. Canada.

Rao, R. (2003). From unstructured data to actionable intelligence. In *Proceedings of the IEEE Computer Society*.

Ratnaparkhi, A., Reynar, J., & Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the workshop on Human Language Technology*, (pp. 250–255).

Rau, L. F. (1991). Extracting company names from text. In *Proceedings of the Conference on Artificial Intelligence Applications of IEEE*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference Artificial Intelligence (IJCAI)*, (pp. 448–453).

Ritchie, G., & Thompson, H. (1984). *Natural language processing. In: Artificial Intelligence: Tools, Techniques, and Applications*. Harper and Row.

Samuelsson, Y., Ackstr, O. T., Velupillai, S., Eklund, J., Fishel, M., & Saers, M. (2008). Mixing and blending syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, (pp. 248– 252). Manchester, England.

Sassano, M. (2002). An empirical study of active learning with support vector machines for Japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (pp. 505–512).

Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, N.J: Lawrence Erlbaum.

Schutze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing'92*, (pp. 787–796).

Sekimizu, T., Park, H., & Tsuji, J. (1998). *Identifying the interactions between genes and gene products based on frequently seen verbs in MEDLINE abstract.* Tokyo Japan: Universal Academy Press.

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity. In *Proceedings of the Conference on Language Resources and Evaluation.*

Singh, N. (2004). *The use of syntactic structure in relationship extraction.* Master's thesis, MIT.

Smith, D. (2002). Detecting and browsing events in unstructured text. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.*

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning, 34*, 233–272.

Sopena, J. M., Lloberas, A., & Moliner, J. L. (1998). A connectionist approach to prepositional phrase attachment for real world texts. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, (pp. 1233–1237).

Srihari, R., & Li, W. (1999). Information extraction supported question answering. In *Proceedings of the Eighth Text REtrieval Conference*, (pp. 185–196).

Srinivas, M., & Bhattacharyya, P. (2006). Prepositional phrase attachment through semantic association using connectionist approach. In *Proceedings of the Third International WordNet Conference (GWC2006)*, (pp. 273–277).

Stavrianou, A., Andritsos, P., & Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *SIGMOD Rec.*, *36* (3), 23–34.

Stevenson, M., & Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistic*, *27* (3), 321–349.

Stokoe, C., Oakes, M. P., & Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th ACM SIGIR conference on research and development in information retrieval*, (pp. 159–166). ACM.

Sun, J., Karray, F., Basir, O., & Kamel, M. (2002). Natural language understanding through fuzzy logic inference and its application to speech recognition. In *Proceedings of the IEEE International on Fuzzy Systems*, vol. 2, (pp. 1120–1125).

Surdeanu, M., Johansson, R., Meyers, A., Arquez, L., & Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic

dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning.*

Toral, A. (2005). DRAMNERI: A free knowledge based tool to named entity recognition. In *Proceedings of the First Free Software Technologies Conference*, (pp. 27–32).

Torsun, I. S. (1995). *Foundations of Intelligent Knowledge-Based Systems.* London: Academic Press Limited.

Vailaya, A., Bluvas, P., Kincaid, R., Kuchinsky, A., Creech, M., & Adler, A. (2005). An architecture for biological information extraction and representation. *Bioinformatics, 21* (4), 430–438.

Vanschoenwinkel, B., & Manderick, B. (2003). A weighted polynomial information gain kernel for resolving prepositional phrase attachment ambiguities with support vector machines. In *Proceedings of the 18th international joint conference on Artificial intelligence*, (pp. 133–138). San Francisco.

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics.*

Volk, M. (2003). German prepositions and their kin. A survey with respect to the resolution of PP attachment ambiguities. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, (pp. 77–85).

Wang, D., Li, T., Zhu, S., & Ding, S. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval* , (pp. 307–314). New York, USA: ACM.

Wang, L. X. (1994). *Adaptive Fuzzy Systems and Control: Design and stability analysis*. Prentice Hall.

Wilks, Y. (1978). A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, *6*, 53–74.

Wu, H., & Furugori, T. (1996). Prepositional phrase attachment through a hybrid disambiguation model. In *Proceedings of the 16th conference on Computational linguistics*, (pp. 1070–1073). Morristown, NJ, USA: Association for Computational Linguistics.

Yangarber, R., Grishman, R., Tapanainen, P., & Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000*, (pp. 940–946).

Yangarber, R., Jokipi, L., Rauramo, A., & Huttunen, S. (2005). Information extraction from epidemiological reports. In *Proceedings of Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing HLT/EMNLP-2005*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, (pp. 189–196).

Yarowsky, D. (1996). Homograph disambiguation in text-to-speech synthesi. In J. Hirschberg, R. Sproat, & J. van Santen (Eds.) *Progress in Speech Synthesis*, (pp. 159–175).

Youjin, C., & Jong-Hyeok, L. (2005). Practical word-sense disambiguation using co-occurring concept codes. *Machine Translation*, *19* (1), 59–82.

Zadeh, L. (2006). In E. Sanchez (Ed.) *Fuzzy Logic and the Semantic Web*, (pp.163–210).

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control* , *8* (3), 338–353.

Zadeh, L. A. (1978). Pruf- a meaning representation language for natural languages. *International Journal of Man-Machine Studies*, *10*, 395–459.

Zavrel, J., Daelemans, D., & Veenstra, J. (1997). Resolving PP attachment ambiguities with memory-based learning. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-97)*, (pp. 136–144).

Zhao, S., & Lin, D. (2004). A nearest-neighbor method for resolving PP-attachment ambiguity. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, (pp. 545– 54).