New Strategies for Automated Random Testing

MIAN ASBAT AHMAD

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE
THE UNIVERSITY of York
UNITED KINGDOM

12TH MARCH 2013

Abstract

This is the abstract text with the formula $v_L(t) = \int_{-\infty}^{t} \frac{di_L}{dt}$. This paper describes the computation of feature point correspondences using the spectra of a Hermitian property matrix. Firstly, a complex Laplacian (Hermitian) matrix is constructed from the Gaussian-weighted distances and the difference of SIFT angles between each pair of points in the two images to be matched. Matches are computed by comparing the complex eigenvectors of the Hermitian property matrices for the two point sets acquired from the two images. Secondly, we embed the complex modal structure within Carcassoni's iterative alignment method to render it more robust to rotation. Our method has been evaluated on both synthetic and real-world data.

Contents

List of Tables

List of Figures

Acknowledgements

Several people have contributed to the completion of my PhD dissertation. However, the most prominent personality deserving due recognition is my worthy supervisor, Dr. Manuel Oriol. Thank you Manuel for your endless help, valuable guidance, constant encouragement, precious advice, sincere and affectionate attitude.

I thank my assessor Prof. John Clark for his constructive feedback on my various reports and presentations. I am also thankful and highly indebted to Prof. Richard Paige for his generous help, cooperation and guidance during my research at the University of York.

Special thanks to my father Prof. Mushtaq A. Mian who provided a conducive environment, valuable guidance and crucial support at all levels of my educational career and my very beloved mother whose love, affection and prayers have been my most precious assets. Also I am thankful to my elder brothers Dr. Ashfaq, Dr. Aftab, Dr. Ishaq, Dr. Afaq and my sister Dr. Haleema who have been the source of inspiration for me to pursue higher studies. My immediate

younger brother Dr. Ilyas and my younger sister Ayesha studying in the UK, deserve recognition for their help, well wishes and moral support. Last but not the least I am very thankful to my dear wife Dr. Munazza for her company, help and cooperation throughout my stay at York.

I was funded by Departmental Overseas Research Scholarship (DORS), a financial support awarded to overseas students on the basis of outstanding academic ability and research potential. I am truly grateful to the Department of Computer Science for financial support that allowed me to concentrate on my research.

Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree other than Doctor of Philosophy of the University of York. This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references.

I hereby give consent for my thesis, if accepted, to be made available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Publications:

Some of the material contained in this thesis has appeared in the following published conference and workshop papers:

Kazakov, D. and Shahid, A. (2008). Extracting Multilingual Dictionaries for the teaching of CS and AI. In *4th UK Workshop on AI in Education* as part of the annual SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK.

Kazakov, D. and Shahid, A. (2009). Unsupervised Construction of a Multilingual WordNet from Parallel Corpora. In *Workshop on Natural Language Processing methods and Corpora in Translation, Lexicography, and Language Learning (RANLP '09)*, Borovets, Bulgaria.

Shahid, A. and Kazakov, D. (2009). Automatic Multlingual Lexicon Generation using Wikipedia as a resource. In *Proceedings of the International Conference on Agents and Artificial Intelligence, (ICCART '09)*, Porto, Portugal.

Shahid, A. and Kazakov, D. (2010). Retrieving Lexical Semantics from Parallel Corpora. *Polibits, 5*, 25-28.

Shahid, A. and Kazakov, D. (2011). Using Multilingual Corpora to Extract Semantic Information. In *Proceedings of the Symposium on Learning Language Models from Multilingual Corpora, AISB'11 Convention*, York, UK.

I feel it a great honour to dedicate my PhD thesis to my beloved parents for their significant contribution in achieving the goal of academic excellence reflected in my dissertation.

CHAPTER 1

Introduction and Motivation

1.1 MainSection

1.1.1 subsection

1.1.1.1 subsubsection

CHAPTER 2

Literature Review

CHAPTER 3

Extraction of Multilingual Lexicons from Wikipedia

CHAPTER 4

Extraction of Multilingual Synsets from Aligned Corpora

CHAPTER 5

Morphology and Lexical Distances

CHAPTER 6

Conclusion

References

Adafre, S. & de Rijke, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of the EACL Workshop on New Text*.

Adler, M. & Elhadad, M. (2006). An Unsupervised Morpheme-based HMM for Hebrew Morphological Disambiguation. In *Proceedings of the ACL/CONLL*, (pp. 665–672).

Agirre, E. & Rigau, G. (1995). A Proposal for Word Sense Disambiguation using Conceptual Distance. In *Proceedings of the First International Conference on Recent Advances in Natural Language Processing*, Bulgaria.

Ahn, D., Jijkoun, V., Mishne, G., M̀uller, K., de Rijke, M., & Schlobach, S. (2004). Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*.

Alfred, R., Kazakov, D., Bartlett, M., & Paskaleva, E. (2007). Hierarchical Agglomerative Clustering for Cross-Language Information Retrieval. *International Journal of Translation*, *19*(1), 139–162.

Banerjee, S. & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense

Disambiguation Using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, (pp. 136–145).

Berton, A., Fetter, P., & Regel-Brietzmann, P. (1996). Compund Words in Large-Vocabulary German Speech Recognition Systems. In *Proceedings of The Fourth International Conference on Spoken Language Processing (ICSLP '96)*, (pp. 1165–1168)., Philadelphia, PA, USA.

Black, P. (2006). Dictionary of Algorithms and Data Structures. `http://www.nist.gov/dads/HTML/manhattanDistance.html`.

Braschler, M. & Schäuble, P. (1998). Multilingual Information Retrieval based on Document Alignment Techniques. In Nikolaou, C. & Stephanidis, C. (Eds.), *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98)*, (pp. 183–197)., London.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. California: Wadsworth International.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference*, (pp. 110–117)., Brisbane, Australia. Elsevier Science.

Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., & Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79–85.

Brown, P., Pietra, S., Pietra, V., & Mercer, R. Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, (pp. 264–270).

Brown, P., Pietra, S., Pietra, V., & Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), 263–311.

Brown, P., Pietra, V., deSouza, P., Lai, J., & Mercer, R. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467–479.

Buckley, C. (1985). Implementation of the SMART Information Retrieval System. Technical report 85-686. Cornell University.

Buckley, C. & Voorhees, E. (2000). Evaluating Evaluation Measure Stability. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, (pp. 33–40)., Trento, Italy.

Bunescu, R. & Pasca, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 9–16)., Trento, Italy.

Burnard, L. & Aston, G. (1998). The BNC Handook: Exploring the British National Corpus. Edinburgh: Edinburgh University Press.

Can, B. & Manandhar, S. (2009). Unsupervised Learning of Morphology by using Syntactic Categories. In *Working Notes CLEF 2009 Workshop*.

Charitakis, K. (2007). Using Parallel Corpora to Create a Greek-English Dictionary with Uplug. In *Proceedings of the 16th Nordic Conference on Computational Linguistics - NODALIDA'07*.

Chew, P., Bader, B., Kolda, T., & Abdelali, A. (2007). Cross-Language Information Retrieval Using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, volume 4592, chapter Using Statistics in Lexical Analysis. Lawrence Erlbaum.

Clark, A. (2000). Inducing Syntactic Categories by Context Distribution Clustering. In *The Fourth Conference on Natural Language Learning (CoNLL)*, (pp. 91–94).

Collins, M. & Brooks, J. (1995). Prepositional Phrase Attachment through a Backed-Off Model. In *Third Workshop on Very Large Corpora, Association for Computational Linguistics, ACL.*

Creutz, M. & Lagus, K. (2007). Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Transactions on Speech and Language Processing*, *4(1)*.

Dagan, I. & Itai, A. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, *20*, 563–596.

Dagan, I., Itai, A., & Schwall, U. (1991). Two Languages are More Informative than One. *ACL*, *29*, 130–137.

Dasgupta, S. & Ng, V. (2007). Unsupervised Part-of-Speech Acquisition for Resource-Scarce Languages. In *Proceedings of the EMNLP-CoNLL*, (pp. 218–227).

Davies, D. & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions and Pattern Analysis and Machine Intelligence*, *1/2*, 224–227.

de Saussure, F. (1959). *Course in General Linguistics*. Philosophical Library, New York.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*.

Diab, M. (2000). An Unsupervised Method for Multilingual Word Sense Tagging using Parallel Corpora: A Preliminary Investigation. In *ACL-2000 Workshop on Word Senses and Multilinguality*, (pp. 1–9)., Hong Kong.

Diab, M. & Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging

using Parallel Corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H., & Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1 (COLING '98)*, Stroudsburg, PA, USA.

Duda, R. & Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc.

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferrández, S., Toral, A., Ferrández, O., Ferrández, A., & Muñoz, R. (2007). *Lecture Notes in Computer Science*, volume 4592, chapter Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering. Springer.

Fišer, D. (2007). Leveraging Parallel Corpora and Existing WordNets for Automatic Construction of the Slovene Wordnet. In *Proceedings of (L&TC 2007)*, Poznań, Poland.

Francis, W. (1964). A Standard Sample of Present-Day English for use with Digital Computers. Report to the U.S. Office of Education on Cooperative Research Project No. E-007.

Fung, P. & Wu, D. (1995). Coerced Markov Models for Cross-Lingual Lexical-Tag Relations. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, (pp. 240 – 255)., Leuven, Belgium.

Gabrilovich, E. & Markovitch, S. (2006). Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Association for the Advancement of Artificial Intelligence, AAAI'06*.

Gale, W., Church, K., & Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, *26*(5-6), 415–439.

Giles, J. (2005). Internet Encyclopaedias go Head to Head. *Nature*, *438(7070):900-901*.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, *27(2):153-198*.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences*. Cambridge: The Press Syndicate of the University of Cambridge.

Harris, Z. (1955). From Phoneme to Morpheme. *Language*, *31(2)*.

Hartigan, J. (1975). *Clustering Algorithms (Probability & Mathematical Statistics)*. Cambridge: John Wiley & Sons Inc.

Hull, D. & Grefenstette, G. (1996). Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jain, A. & Dubes, R. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Jardino, M. & Adda, G. (1993). Automatic Word Classification Using Simulated Annealing. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume 2, (pp. 41 – 44)., Minneapolis.

Jones, G., Fantino, F., Newman, E., & Zhang, Y. (2008). Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented With Dictionaries Mined from Wikipedia. In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies*, Hydrabad, India.

Kaplan, A. (1950). An Experimental Study of Ambiguity in Context. *Mechanical Translation*, *1*(1-3).

Katz, S. (1987). Estimation of Probabilities for Sparse Data for the Language Model Component of a Speech Recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35(3)*, 400–401.

Kawaba, M., Nakasaki, H., Utsuro, T., & Fukuhara, T. (2008). Cross-Lingual Blog Analysis based on Multilingual Blog Distillation from Multilingual Wikipedia Entries. In *Proceedings of International Conference on Weblogs and Social Media, ICWSM'08*.

Kazakov, D. (2000). Achievements and Prospects of Learning Word Morphology with Inductive Logic Programming. In Cussens, J. & Dzeroski, S. (Eds.), *Learning Language in Logic*, (pp. 89–109). Springer.

Kazakov, D., Cussens, J., & Manandhar, S. (2006). On The Duality of Semantics and Syntax: The PP Attachment Case. Technical report YCS 409. Department of Computer Science, University of York, UK.

Kazakov, D. & Manandhar, S. (2001). Unsupervised Learning of Word Segmentation Rules with Genetic Algorithms and Inductive Logic Programming. *Machine Learning*, *43*(1-2), 121–162.

Kazakov, D. & Shahid, A. (2008). Extracting Multilingual Dictionaries for the Teaching of CS and AI. In *4th UK Workshop on AI in Education*.

Kilgarriff, A. & Rosenzweig, J. (1999). Framework and Results for English Senseval. *Computers and the Humanities*, *34(1)*, 15–48.

Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. `http://www.isi.edu/~koehn/publications/europarl/`.

Kvålseth, T. (1987). Entropy and Correlation: Some Comments. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-17*, 517–519.

Lancaster, F. (1968). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Wiley.

Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T. K. & Littman, M. L. (1990). Fully Automatic Cross-Language Document Retrieval using Latent Semantic Indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, (pp. 31–38).

Lee, L. (1999). Measures of Distributional Similarity. In *Proceedings of the 37th ACL*.

Lefever, E. & Hoste, V. (2009). SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, (pp. 82–87)., Boulder, Colorado.

Lefever, E. & Hoste, V. (2010a). Construction of a Benchmark Data Set for Cross-lingual Word Sense Disambiguation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Malta.

Lefever, E. & Hoste, V. (2010b). SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, (pp. 15–20)., Uppsala, Sweden.

Lesk, M. (1986). Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a Pine Cone from a Ice Cream Cone. In *Proceedings of SIGDOC'86*.

Levenstein, V. (1966). Binary Codes Capable of Correcting Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Li, X., Szpakowicz, S., & Matwin, S. (1995). A WordNet-based Algorithm for Word Sense Disambiguation. In *Proceedings of IJCAI-95*, (pp. 1368–1374)., Montreal, Canada.

Luhn, H. (1959). The Automatic Creation of Literature Abstracts. *IBM Journal*

of Research and Development, *2*, 159–165.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, *19(2)*.

Martin, S., Liermann, J., & Ney, H. (1998). Algorithms for Bigram and Trigram Word Clustering. *Speech Communication*, *24(1)*, 19 – 37.

McEnery, A. (2003). *Lecture Notes in Computer Science*, chapter Corpus Linguistics, (pp. 448–463). Oxford University Press.

Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York.

Mihalcea, R. & Moldovan, D. (1999). A Method for Word Sense Disambiguation of Unrestricted Text. In *Proceedings of the 37th Meeting of ACL*, College Park, MD.

Miller, D., Leek, T., & Schwartz, R. (1999). A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 214–221)., Berkeley, California, United States.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *Journal of Lexicography*, *3(4):235-244*.

Mitchell, T. (1997). *Machine Learning*. MIT Press and McGraw-Hill.

Najork, M. & Wiener, J. (2001). Breadthfirst Crawling Yields High-quality Pages. In *Proceedings of the 10th International Conference on World Wide*

Web.

Ng, H., Wang, B., & Chan, Y. (2007). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, (pp. 455–462)., Sapporo, Japan.

Oard, D. & Dorr, B. (1996). A Survey of Multilingual Text Retrieval. Technical report. University of Maryland at College Park College Park, MD, USA.

Och, F. (1999). An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics (EACL '99)*, Bergen, Norway.

Och, F. & Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, Hong Kong.

Och, F. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29(1):19-51*.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Pirelli, V. (1993). Morphology, Analogy and Machine Translation. PhD Thesis. Salford University, UK.

Pirkola, A. (1998). The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, New York, USA.

Potthast, M., Stein, B., & Anderka, M. (2008). Wikipedia-based Multilingual Retrieval Model. In *Proceedings of the 30th European Conference on IR Research, ECIR'08*, Glasgow.

Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning*, *1(1)*, 81–106.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Richman, A. & Schone, P. (2008). Mining Wiki Resources for Multilingual Named Entity Recognition. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL'08*, (pp. 1–9)., Columbus, Ohio.

Robertson, S. (2006). On GMAP and Other Transformations. In *CIKM '06 Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, New York, USA.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Proceedings of Advances in Web Intelligence*, Lodz, Poland.

Russell, S. & Norvig, P. (1995). *Artificial Intelligence*. Prentice Hall.

Sagot, B. & Fišer, D. (2008). Building a Free French WordNet from Multilingual Resources. In *Proceedings of OntoLex 2008*, Marrackech.

Salton, G. (1970). Automatic Processing of Foreign Language Documents. *Journal of the American Society for Information Science*, *21*, 187–194.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

Salton, G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18(11)*, 613–620.

Sato, S. (2009). Crawling English-Japanese Person-Name Transliterations from the Web. In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain.

Schäuble, P. (1997). *Multimedia Information Retrieval*. Kluwer Academic Publishers.

Schone, P. & Jurafsky, D. (2000). Knowledge-free Induction of Morphology

using Latent Semantic Analysis. In *Proceedings of the CoNLL*, (pp. 67–72).

Sedding, J. & Kazakov, D. (2004). WordNet-Based Text Document Clustering. In *3rd Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND)*, Geneva, Switzerland.

Shahid, A. & Kazakov, D. (2009). Automatic Multilingual Lexicon Generation using Wikipedia as a Resource. In *Proceedings of the International Conference on Agents and Artificial Intelligence, ICAART*, Porto, Portugal.

Shahid, A. & Kazakov, D. (2010). Retrieving Lexical Semantics from Multilingual Corpora. *Polibits*, *5*, 25–28.

Shahid, A. & Kazakov, D. (2011). Using Multilingual Corpora to Extract Semantic Information. In *Proceedings of the Symposium on Learning Language Models from Multilingual Corpora, AISB'11 Convention*, York, UK.

Sheridan, P. & Ballerini, J.-P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 58–65).

Snyder, B. & Barzilay, R. (2008). Unsupervised Multilingual Learning for Morphological Segmentation. In *The Annual Conference of the Association for Computational Linguistics*.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Applications in Retrieval. *Journal of Documentation*, *28(1)*, 11–21.

Specia, L., Nunes, M., & Stevenson, M. (2005). Eploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP-2005)*, Borovets, Bulgaria.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. In *6th ACM SIGKDD, World Text Mining Conference*,

Boston, MA, USA.

Strehl, A. (2002). Relationbship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. PhD Thesis. The University of Texas at Austin.

Talvensaari, T., Juhola, M., Laurikkala, J., & Järvelin, K. (2007). Corpus-based Cross-language Information Retrieval in Retrieval of Highly Relevant Documents. *Journal of the American Society for Information Science and Technology*, *58(3)*, 322–334.

Tiedemann, J. (1999). Uplug - a Modular Corpus Tool for Parallel Corpora. In *In the Parallel Corpus Symposium (PKS99)*, Uppsala University, Sweden.

Tiedemann, J. (2004). The OPUS Corpus - Parallel & Free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Tufis, D. (2000). Design and Development of a Multilingual Balkan WordNet. *Romanian Journal of Information Science and Technology Special Issue*, *7:1-2*.

Tufis, D., Ion, R., & Ide, N. (2004). Fine-Grained Word Sense Disambiguation based on Parallel Corpora, Word Alignment, Word Clustering, and Aligned WordNets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.

Tyers, F. & Pienaar, J. (2008). Extracting Bilingual Word Pairs from Wikipedia. In *Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference, LREC'08*, (pp. 19–22).

van den Bosch, A., Daelemans, W., & Weijters, T. (1996). Morphological Analysis as Classification: an Inductive-Learning Approach. In *Proceedings of the Second International Conference on New Methods in Language Processing*

(NeMLap-2), (pp. 79–89)., Bilkent University, Ankara, Turkey.

van der Plas, L. & Tiedemann, J. (2006). Finding Synonyms using Automatic Word Alignment and Measures of Distributional Similarity. In *Proceedings of ACL/COLING 2006*, Sydney, Australia.

van Rijsbergen, C. (1979). *Information Retrieval*. Butterworth-Heinemann.

Voorhees, E. & Harman, D. (1999). Overview of the Seventh Text REtrieval Conference (TREC-7). In Voorhees, E. & Harman, D. (Eds.), *In NIST Special Publication 500-242*, (pp. 1–23).

Vossen, P. (1996). Right or Wrong: Combining Lexical Resources in the Euro-WordNet Proejct. In *Proceedings of Euralex-96 International Congress*.

Vossen, P. (1998). *EuroWordNet: a Multilingual Database with Lexical Semantic Networks for European Languages*. Kluwer.

Wagner, R. & Fischer, M. (1974). The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, *21:1*, 168–173.

Wong, W. & Fu, A. (2000). Incremental Document Clustering for Web Page Classification. In *IEEE 2000 International Conference on Information Society in 21st Century: Emerging Technologies and New Challenges*.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, New York, USA.

Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, (pp. 454–460)., Nantes, France.

Yarowsky, D. (1994). Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd*

Annual Meeting of the Association for Computational Linguistics, Las Cruces.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivalling Supervised Methods. *ACL*, *33*, 189–196.

Young, P. (1994). Cross Language Information Retrieval Using Latent Semantic Indexing. Master's Thesis. University of Knoxwille, Tennessee: Konxwille.

Yvon, F. (1996). Prononcer par analogies: motivations, formalisations et évaluations. PhD Thesis. ENST Paris, France.

Zesch, T. Muller, C. & Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

Zesch, T., Gurevych, I., & Muhlhauser, M. (2007). Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT'07*, (pp. 205–208).