Distributional approaches to semantic analysis

Diarmuid Ó Séaghdha

Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge
do242@cam.ac.uk

HIT-MSRA Summer Workshop on Human Language Technology
August 2011
http://www.cl.cam.ac.uk/~do242/Teaching/HIT-MSRA-2011/


UNIVERSITY OF CAMBRIDGE

Bibliography I

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries (DL-00)*, San Antonio, TX.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee-Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, Montreal, Canada.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the Web. In *Proceedings of IJCAI-07*, Hyderabad, India.

Lawrence W. Barsalou. 1983. Ad hoc categories. *Memory and Cognition*, 11(3):211–227.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Bibliography II

David Blei and Jon McAuliffe. 2007. Supervised topic models. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS-07)*, Vancouver, Canada.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Sergei Brin. 1998. Extracting patterns and relations from the World Wide Web. In *Proceedings of the International Workshop on the Web and Databases (WebDB-98)*, Valencia, Spain.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the ACL-06 Interactive Presentation Sessions*, Sydney, Australia.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the SemEval-2 Workshop*, Uppsala, Sweden.

Bibliography III

Jonathan Chang and David Blei. 2009. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1):124–150.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of NIPS-09*, Vancouver, BC.

Stanley F. Chen and Joshua Goodman. 1996. An empirical studyof smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA.

James Curran. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Hal Daumé III. 2009. Markov random topic fields. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP-09)*, Suntec, Singapore.

Bibliography IV

Dmitry Davidov, Ari Rappoport, and Moshe Koppel. 2007. Fully unsupervised discovery of concept-specific relationships by web mining. In *Proceedings of ACL-07*, Prague, Czech Republic.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, Genoa, Italy.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, Prague, Czech Republic.

Bibliography V

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. R. Frith. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Oxford Philological Society, Oxford.

Roxana Girju, Adriana Badelescu, and Dan Moldovan. 2003. Learning semantic constraings for the automatic discovery of part-whole relations. In *Proceedings of HLT-NAACL-03*, Edmonton, Canada.

Bibliography VI

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, , and Deniz Yuret. 2007. Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-07)*, Prague, Czech Republic.

Amit Goyal and Hal Daumé III. 2011. Approximate scalable bounded space sketch for large data nlp. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Tom L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Bibliography VII

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*, Nantes, France.

Gregor Heinrich. 2009. Parameter estimation for text analysis. Technical report, Fraunhofer IGD.
`http://www.arbylon.net/publications/text-est2.pdf`.

Eduard Hovy, Zornitsa Kozareva, and Ellen Riloff. 2009. Toward completeness in concept extraction and classification. In *Proceedings of EMNLP-09*, Singapore.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of CogSci-00*, Philadelphia, PA.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Bibliography VIII

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the Web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, Columbus, OH.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, Boston, MA.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, MD.

Bibliography IX

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC-86)*, Toronto, ON.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Jinhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING-98)*.

Bibliography X

Dekang Lin. 1998b. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on Evaluation of Parsing Systems*, Granada, Spain.

sca Marius Pa and Benjamin Van Durme. 2008. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of ACL-08:HLT*, Columbus, OH.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th International World Wide Web Conference (WWW-08)*, Beijing, China.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA.

Bibliography XI

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI-08)*, Helsinki, Finland.

Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, Columbus, OH.

Preslav Nakov and Zornitsa Kozareva. 2011. Combining relational and attributional similarity for semantic relation classification. In *Proceedings of the 8th Conference on Recent Advances in Natural Language Processing (RANLP-11)*, Hissar, Bulgaria.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*.

Bibliography XII

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK.

Diarmuid Ó Séaghdha and Ann Copestake. 2009. Using lexical and relational similarity to classify semantic relations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, Athens,Greece.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, University of Cambridge.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*, Suntec, Singapore.

Bibliography XIII

J. Pustejovsky, J. Castan o, J. Zhang, M. Kotecki, and B. Cochran. 2002. Robust rela- tional parsing over biomedical literature: Extracting *inhibit* relations. In *Proceedings of the 7th Pacific Symposium on Biocomputing (PSB-02)*, Lihue, HI.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Suntec, Singapore.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, MI.

Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical Report CS-TR-4956, University of Maryland.

Bibliography XIV

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Luis Sarmento, Valentin Jijkoun, Maarten de Rijke, and Eugenio Olveira. 2007. "more like these": Growing entity classes from seeds. In *Proceedings of CIKM-07*, Lisbon, Portugal.

H. Sebastian Seung and Daniel D. Lee. 2000. The manifold ways of perception. *Science*, 290(5500):2268–2269.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS-05*, Vancouver, BC.

Mark Stevenson and Mark A. Greenwood. 2005. A semantic approach to IE pattern induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, MI.

Bibliography XV

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Stastical Association*, 101(476):1566–1581.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML-01)*, Freiburg, Germany.

Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Bibliography XVI

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of COLING-02*, Taipei, Taiwan.

Dominic Widdows. 2004. *Geometry and Meaning*. CSLI Publications, Stanford, CA.