Semantic Similarity Frontiers: From Concepts to Documents: Extended Bibliography

David Jurgens
jurgens@stanford.edu
Stanford University

Mohammad Taher Pilehvar
pilehvar@di.uniroma1.it
Sapienza University of Rome

This document is the extended bibliography for the EMNLP 2015 tutorial "Semantic Similarity Frontiers: From Concepts to Documents: Extended Bibliography" and can be referenced as (Jurgens and Pilehvar 2015b). This bibliograpy is intended as a living document and we welcome suggestions via the GitHub page hosting this document: https://github.com/davidjurgens/similarity-tutorial.

1. Foundations

Resources and Fundamental Techniques

1. (Jurafsky and Martin 2000)
2. (Manning, Raghavan, and Schütze 2008)
3. (Turney, Pantel, and others 2010)
4. (Tversky and Gati 1982)
5. (Miller 1995)
6. (Fellbaum 1998)
7. (Hovy et al. 2006)
8. (Bond and Paik 2012)
9. (Petrolito and Bond 2014)
10. (Gawron 2014)
11. (Webber, Moffat, and Zobel 2010)

2. Sense Similarity

2.1 Approaches tied to sense inventory
2.1.1 Graph-based techniques.

1. (Budanitsky and Hirst 2006)
2. (Hirst and St-Onge 1998)
3. (Sussna 1993)
4. (Wu and Palmer 1994)
5. (Leacock and Chodorow 1998)
6. (Jiang and Conrath 1997)
7. (Resnik 1995)
8. (Lin 1998)
9. (Morris and Hirst 1991)
10. (Jarmasz and Szpakowicz 2003)
11. (Kozima and Furugori 1993)

12. (Kozima and Ito 1997)
13. (Morris and Hirst 1991)
14. (Banerjee and Pedersen 2003)
15. (Pilehvar, Jurgens, and Navigli 2013)

2.1.2 Explicit semantic representation.

1. (Kashyap et al. 2014a)
2. (Iacobacci, Pilehvar, and Navigli 2015a)
3. (Camacho-Collados, Pilehvar, and Navigli 2015a)
4. (Camacho-Collados, Pilehvar, and Navigli 2015b)
5. (Chen, Liu, and Sun 2014)
6. (Rothe and Schütze 2015)

2.2 Approaches not tied to sense inventory

1. (Reisinger and Mooney 2010)
2. (Huang et al. 2012)
3. (Wu and Giles 2015)
4. (Liu et al. 2015)

3. Word Similarity

3.1 Approaches

Corpus-based Approaches

1. (Landauer and Dumais 1997)
2. (Turney 2001)
3. (Turney et al. 2003)
4. (Bullinaria and Levy 2007)
5. (Zesch, Müller, and Gurevych 2008)
6. (Bullinaria and Levy 2012)
7. (Mikolov et al. 2013a, 2013b; Mikolov, Yih, and Zweig 2013)
8. (Baroni, Dinu, and Kruszewski 2014)
9. (Levy and Goldberg 2014a)
10. (Levy and Goldberg 2014b)
11. (Pennington, Socher, and Manning 2014)
12. (Levy, Goldberg, and Dagan 2015)
13. (Vilnis and McCallum 2015)
14. (Liu et al. 2015)
15. (Li and Jurafsky 2015)

Knowledge Base-based Approaches

1. (Strube and Ponzetto 2006)
2. (Gabrilovich and Markovitch 2007)
3. (Hughes and Ramage 2007)
4. (Yeh et al. 2009)
5. (Agirre et al. 2009)

6. (Pilehvar, Jurgens, and Navigli 2013)
7. (Faruqui and Dyer 2015)

Retrofitting Approaches

1. (Goikoetxea et al. 2015)
2. (Faruqui et al. 2015)
3. (Iacobacci, Pilehvar, and Navigli 2015b)

3.2 Benchmarks

1. (Radinsky et al. 2011)
2. (Rubenstein and Goodenough 1965)
3. (Landauer and Dumais 1997) (TOEFL)
4. (Finkelstein et al. 2001)
5. (Turney 2001) (ESL)
6. (Jarmasz and Szpakowicz 2012)
7. (Bruni et al. 2012)
8. (Luong, Socher, and Manning 2013)
9. (Hill, Reichart, and Korhonen 2014)

3.3 Comparison Techniques

1. (Gawron 2014)

4. Phrase Similarity

4.1 Approaches

1. (Mitchell and Lapata 2009)
2. (Zanzotto et al. 2010)
3. (Erk, Padó, and Padó 2010)
4. (Baroni and Zamparelli 2010)
5. (Guevara 2010)
6. (Socher et al. 2012)
7. (Socher et al. 2013)
8. (Hashimoto et al. 2014)
9. (Socher et al. 2014)
10. (Wieting et al. 2015)
11. (He, Gimpel, and Lin 2015)

4.2 Benchmarks

1. (Mitchell and Lapata 2010)

5. Sentence Similarity

1. (Allison and Dix 1986)
2. (Wise 1996)
3. (Kešelj et al. 2003)

4. (Croce, Moschitti, and Basili 2011)
5. (Agirre et al. 2012)
6. (Agirre et al. 2013a)
7. (Bär et al. 2012)
8. (Chávez and Lonardi 2010)
9. (Bär, Zesch, and Gurevych 2013)
10. (Sultan, Bethard, and Sumner 2014)
11. (Sultan, Bethard, and Sumner 2015)
12. (Kiros et al. 2015)
13. (Kenter and de Rijke 2015)

5.1 Benchmarks

1. (Dolan, Quirk, and Brockett 2004)
2. (Li et al. 2006)
3. (Agirre et al. 2012)
4. (Agirre et al. 2013b)
5. (Agirre et al. 2014)
6. (Agirre, Banea, and others 2015)

6. Paragraph Similarity

6.1 Approaches

1. (Galitsky, Kuznetsov, and Usikov 2013)
2. (Le and Mikolov 2014)
3. (Higgins et al. 2014)

7. Document Similarity

7.1 Approaches

Vector Space or Decomposition-based Approaches

1. (Salton, Wong, and Yang 1975)
2. (Deerwester et al. 1990)
3. (Hofmann 1999)
4. (Xu, Liu, and Gong 2003)
5. (Xu and Gong 2004)
6. (Cai, He, and Han 2011)

BM25 Approaches

1. (Robertson, Zaragoza, and Taylor 2004)

Topic-model Approaches

1. (Blei, Ng, and Jordan 2003)
2. (Blei and Lafferty 2009)
3. (Teh et al. 2006)

4. (Kim et al. 2012)
5. (Griffiths and Tenenbaum 2004)
6. (Li and McCallum 2006)
7. (Steyvers and Griffiths 2007)
8. (Ghahramani, Jordan, and Adams 2010)
9. (Mao et al. 2012)
10. (Wang et al. 2013)
11. (McCallum, Wang, and Corrada-Emmanuel 2007)
12. (El-Kishky et al. 2014)

Structured Document Similarity Approaches

1. (Wilkinson 1994)
2. (Tekli, Chbeir, and Yetongnon 2009)
3. (Huang et al. 2013)
4. (Xiong and Callan 2015)

Knowledge Base-based Approaches

1. (Lakkaraju, Gauch, and Speretta 2008)
2. (Yazdani and Popescu-Belis 2013)
3. (Schuhmacher and Ponzetto 2014)
4. (Franco-Salvador, Rosso, and Navigli 2014)

7.2 Benchmarks

1. (Lee, Pincombe, and Welsh 2005)

8. Cross-Level Semantic Similarity

8.1 Approaches

- (Biçici and Way 2014)
- (Kashyap et al. 2014b)
- (Jimenez et al. 2014)
- (Chávez et al. 2014)
- (Pedersen 2014)
- (Pilehvar and Navigli 2015a)

8.2 Benchmarks

- (Jurgens, Pilehvar, and Navigli 2014)

9. Resources

This section includes the references for works presented in the Tools and Resources section, but see the slides for full details and URLs.

- WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi 2004)
- ADW (Pilehvar and Navigli 2015b)
- NLTK (Bird 2006)
- DKPro Similarity (Bär, Zesch, and Gurevych 2013)
- S-Space Package (Jurgens and Stevens 2010)
- DISSECT (Dinu and others 2013)
- Gensim (Řehůřek and Sojka 2010)
- CROWN (Jurgens and Pilehvar 2015a)
- BabelNet (Navigli and Ponzetto 2012)
- WordVectors.org (Faruqui and Dyer 2014)

References

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland.

Agirre, Eneko, Carmen Banea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, s-panish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June*.

Agirre, Eneko, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*, pages 385–393, Montréal, Canada.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013a. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, USA, June. Association for Computational Linguistics.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013b. *SEM 2013 Shared Task: Semantic textual similarity, including a pilot on typed-similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 32–43, Atlanta, Georgia.

Allison, Lloyd and Trevor I Dix. 1986. A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23(5):305–310.

Banerjee, Satanjeev and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 3, pages 805–810.

Bär, Daniel, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of SemEval-2012*, pages 435–440, Montreal, Canada.

Bär, Daniel, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. In *ACL (Conference System Demonstrations)*, pages 121–126.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010*

Conference on Empirical Methods in Natural Language Processing, pages 1183–1193. Association for Computational Linguistics.

Biçici, Ergun and Andy Way. 2014. Rtm-dcu: Referential translation machines for semantic similarity. In *Proceedings of SemEval*.

Bird, Steven. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Blei, David M and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bond, Francis and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th International Global WordNet Conference*, pages 64–71.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Bullinaria, J.A. and J.P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, (3):510.

Bullinaria, John A and Joseph P Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907.

Cai, Deng, Xiaofei He, and Jiawei Han. 2011. Locally consistent concept factorization for document clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):902–913.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015a. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015b. A unified multilingual semantic representation of concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751, Beijing, China, July. Association for Computational Linguistics.

Chávez, Alexander, Héctor Dávila, Yoan Gutiérrez, Antonio Fernández-Orquín, Andrés Montoyo, and Rafael Muñoz. 2014. Umcc_dlsi_semsim: Multilingual system for measuring semantic textual similarity. *SemEval 2014*, page 716.

Chávez, Edgar and Stefano Lonardi, editors. 2010. *String Processing and Information Retrieval - 17th International Symposium, SPIRE 2010, Los Cabos, Mexico, October 11-13, 2010. Proceedings*, volume 6393 of *Lecture Notes in Computer Science*. Springer.

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035, Doha, Qatar.

Croce, Danilo, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046. Association for Computational Linguistics.

Deerwester, Scott C., Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Dinu, Georgiana et al. 2013. Dissect-distributional semantics composition toolkit. In *ACL*.

Dolan, Bill, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland.

El-Kishky, Ahmed, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316.

Erk, Katrin, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Faruqui, Manaal, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL*.

Faruqui, Manaal and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA, June. Association for Computational Linguistics.

Faruqui, Manaal and Chris Dyer. 2015. Non-distributional word vector representations. In *ACL*.

Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Franco-Salvador, Marc, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of EACL*, pages 414–423.

Gabrilovich, Evgeniy and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Galitsky, Boris A, Sergei O Kuznetsov, and Daniel Usikov. 2013. Parse thicket representation for multi-sentence search. In *Conceptual Structures for STEM Research and Education*. Springer, pages 153–172.

Gawron, Jean Mark. 2014. Improving sparse word similarity models with asymmetric measures. In *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics*.

Ghahramani, Zoubin, Michael I Jordan, and Ryan P Adams. 2010. Tree-structured stick breaking for hierarchical data. In *Advances in neural information processing systems*, pages 19–27.

Goikoetxea, Josu, Aitor Soroa, Eneko Agirre, and Basque Country Donostia. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1434–1439.

Griffiths, DMBTL and MIJJB Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17.

Guevara, Emiliano. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Hashimoto, Kazuma, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2014. Jointly learning word representations and composition functions using predicate-argument structures. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1544–1555.

He, Hua, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*.

Higgins, Derrick, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel Tetreault, Daniel Blanchard, et al. 2014. Is getting the right answer just about choosing the right words? the role of syntactically-informed features in short answer scoring. In *ACL*.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Hirst, Graeme and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*. MIT Press, pages 305–332.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882, Jeju, South Korea. Association for Computational Linguistics.

Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

Hughes, Thad and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL*, pages 581–589.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015a. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on*

Natural Language Processing (Volume 1: Long Papers), pages 95–105, Beijing, China, July. Association for Computational Linguistics.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015b. Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Jarmasz, Mario and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of RANLP*, pages 212–219.

Jarmasz, Mario and Stan Szpakowicz. 2012. Roget's thesaurus and semantic similarity. *arXiv preprint arXiv:1204.0245*.

Jiang, Jay J. and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, pages 19–30, Taiwan.

Jimenez, Sergio, George Duenas, Julia Baquero, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2014. Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. *SemEval 2014*, page 732.

Jurafsky, Dan and James H Martin. 2000. *Speech & language processing*. Pearson Education India.

Jurgens, David and Mohammad Taher Pilehvar. 2015a. Reserating the awesometastic: An automatic extension of the wordnet taxonomy for novel terms. In *NAACL*.

Jurgens, David and Mohammad Taher Pilehvar. 2015b. Semantic similarity frontiers: From concepts to documents. In *EMNLP Tutorial*.

Jurgens, David, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. *SemEval 2014*, page 17.

Jurgens, David and Keith Stevens. 2010. The s-space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 30–35. Association for Computational Linguistics.

Kashyap, Abhay, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. 2014a. Meerkat mafia: Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 416–423, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Kashyap, Abhay, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. 2014b. Meerkat mafia: Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 416–423. Association for Computational Linguistics.

Kenter, Tom and Maarten de Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international conference on information and knowledge management (CIKM)*.

Kešelj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceeding of Pacific Association for Computational Linguistics*.

Kim, Joon Hee, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792. ACM.

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of NIPS*.

Kozima, Hideki and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*, EACL '93, pages 232–239, Utrecht, The Netherlands.

Kozima, Hideki and Akira Ito. 1997. Context-sensitive word distance by adaptive scaling of a semantic space. In *Recent Advances in Natural Language Processing: Selected Papers from RANLP 1995, Volume 136 of Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory, Chapter 2*, pages 111–124, Tzigov Chark, Bulgaria. John Benjamins Publishing Company.

Lakkaraju, Praveen, Susan Gauch, and Mirco Speretta. 2008. Document similarity based on concept tree distance. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 127–132. ACM.

Landauer, Thomas K 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; Psychological Review*, 104(2):211.

Le, Quoc V and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31 st International Conference on Machine Learning*.

Leacock, Claudia and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*. MIT Press, pages 265–283.

Lee, M, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. *Cognitive Science*.

Levy, Omer and Yoav Goldberg. 2014a. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Levy, Omer and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, Jiwei and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *EMNLP*.

Li, Wei and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Li, Yuhua, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA.

Liu, Yang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *AAAI Conference on Artificial Intelligence*.

Luong, Minh-Thang, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNL*, volume 104.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Mao, Xian-Ling, Zhao-Yan Ming, Tat-Seng Chua, Si Li, Hongfei Yan, and Xiaoming Li. 2012. Sshlda: a semi-supervised hierarchical topic model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 800–809. Association for Computational Linguistics.

McCallum, Andrew, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, pages 249–272.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mitchell, Jeff and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 430–439. Association for Computational Linguistics.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1).

Navigli, Roberto and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pedersen, Ted. 2014. Duluth: Measuring cross–level semantic similarity with first and second–order dictionary overlaps. *SemEval 2014*, page 247.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004*, pages 38–41. Association for Computational Linguistics.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

Petrolito, Tommaso and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*.

Pilehvar, Mohammad Taher, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1341–1351, Sofia, Bulgaria.

Pilehvar, Mohammad Taher and Roberto Navigli. 2015a. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Pilehvar, Mohammad Taher and Roberto Navigli. 2015b. An open-source framework for multi-level semantic similarity measurement. In *Proceedings of NAACL-HLT*, pages 76–80.

Radinsky, Kira, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Řehůřek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of ACL*, pages 109–117.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453.

Robertson, Stephen, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM.

Rothe, Sascha and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China, July. Association for Computational Linguistics.

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

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schuhmacher, Michael and Simone Paolo Ponzetto. 2014. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 543–552. ACM.

Socher, Richard, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer.

Socher, Richard, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Socher, Richard, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Strube, Michael and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Sultan, Md Arafat, Steven Bethard, and Tamara Sumner. 2014. Dls@cu: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Sultan, Md Arafat, Steven Bethard, and Tamara Sumner. 2015. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International*

Workshop on Semantic Evaluation (SemEval 2015), pages 148–153, Denver, Colorado, June. Association for Computational Linguistics.

Sussna, Michael. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of Second International Conference on Information and Knowledge Base Management*, pages 67–74, Washington D.C., USA.

Teh, Yee Whye, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Tekli, Joe, Richard Chbeir, and Kokou Yetongnon. 2009. An overview on xml similarity: background, current trends and future directions. *Computer science review*, 3(3):151–173.

Turney, Peter. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*.

Turney, Peter, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of Recent Advances in Natural Language Processing*.

Turney, Peter D, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Tversky, Amos and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123.

Vilnis, Luke and Andrew McCallum. 2015. Word representations via gaussian embedding. In *ICLR*.

Wang, Chi, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 437–445. ACM.

Webber, William, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, November.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*, 3.

Wilkinson, Ross. 1994. Effective retrieval of structured documents. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–317. Springer-Verlag New York, Inc.

Wise, Michael J. 1996. Yap3: improved detection of similarities in computer program and other texts. *ACM SIGCSE Bulletin*, 28(1):130–134.

Wu, Zhaohui and C. Giles. 2015. Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia. In *AAAI Conference on Artificial Intelligence*.

Wu, Zhibiao and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32^{nd} Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Las Cruces, New Mexico.

Xiong, Chenyan and Jamie Callan. 2015. Esdrank: Connecting query and documents through external semi-structured data. In *International Conference on Information and Knowledge Management*, volume 6, pages 3–1.

Xu, Wei and Yihong Gong. 2004. Document clustering by concept factorization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209. ACM.

Xu, Wei, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM.

Yazdani, Majid and Andrei Popescu-Belis. 2013. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3185–3189. AAAI Press.

Yeh, Eric, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. 2009. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics.

Zanzotto, Fabio Massimo, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271. Association for Computational Linguistics.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. Using Wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, pages 861–866.