Pasini, Tommaso; **Navigli, Roberto**

Train-o-matic: supervised word sense disambiguation with no (manual) effort. (English)

Zbl 07153716
Artif. Intell. 279, Article ID 103215, 22 p. (2020).

Summary: Word Sense Disambiguation (WSD) is the task of associating the correct meaning with a word in a given context. WSD provides explicit semantic information that is beneficial to several downstream applications, such as question answering, semantic parsing and hypernym extraction. Unfortunately, WSD suffers from the well-known knowledge acquisition bottleneck problem: it is very expensive, in terms of both time and money, to acquire semantic annotations for a large number of sentences. To address this blocking issue we present Train-O-Matic, a knowledge-based and language-independent approach that is able to provide millions of training instances annotated automatically with word meanings. The approach is fully automatic, i.e., no human intervention is required, and the only type of human knowledge used is a task-independent WordNet-like resource. Moreover, as the sense distribution in the training set is pivotal to boosting the performance of WSD systems, we also present two unsupervised and language-independent methods that automatically induce a sense distribution when given a simple corpus of sentences. We show that, when the learned distributions are taken into account for generating the training sets, the performance of supervised methods is further enhanced. Experiments have proven that Train-O-Matic on its own, and also coupled with word sense distribution learning methods, lead a supervised system to achieve state-of-the-art performance consistently across gold standard datasets and languages. Importantly, we show how our sense distribution learning techniques aid Train-O-Matic to scale well over domains, without any extra human effort. To encourage future research, we release all the training sets in 5 different languages and the sense distributions for each domain of SemEval-13 and SemEval-15 at http://trainomatic.org.

MSC:

68T Artificial intelligence

Keywords:

word sense disambiguation; corpus generation; word sense distribution learning; multilinguality

Software:

BabelNet; fast-ppr; LexSemTM; Nasari; Penn Treebank; Train-o-matic; word2vec

Full Text: DOI

References:

[1] Navigli, R., Word sense disambiguation: a survey, ACM Comput. Surv., 41, 2, 1-69 (2009)

[2] Navigli, R., Natural language understanding: instructions for (present and future) use, (Proceedings of IJCAI (2018)), 5697-5702

[3] (Fellbaum, C., WordNet: An Electronic Database (1998), MIT Press: MIT Press Cambridge, MA) · Zbl 0913.68054

[4] Agirre, E.; de Lacalle, O. L.; Soroa, A., Random walks for knowledge-based word sense disambiguation, Comput. Linguist., 40, 1, 57-84 (2014)

[5] Moro, A.; Raganato, A.; Navigli, R., Entity linking meets word sense disambiguation: a unified approach, Transactions of the Association for Computational Linguistics (TACL), 2, 231-244 (2014)

[6] Maru, M.; Scozzafava, F.; Martelli, F.; Navigli, R., SyntagNet: challenging supervised word sense disambiguation with lexical-semantic combinations, (Proceedings of EMNLP-IJCNLP (2019))

[7] Tripodi, R.; Pelillo, M., WSD-games: a game-theoretic algorithm for unsupervised word sense disambiguation, (Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (2015)), 329-334

[8] Chaplot, D. S.; Salakhutdinov, R., Knowledge-based word sense disambiguation using topic models, (Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18) (2018)), 5062-5069

[9] Zhong, Z.; Ng, H. T., It makes sense: a wide-coverage word sense disambiguation system for free text, (Proceedings of the ACL

2010 System Demonstrations (2010), Association for Computational Linguistics: Association for Computational Linguistics Uppsala, Sweden), 78-83

[10] Kågebäck, M.; Salomonsson, H., Word sense disambiguation using a bidirectional LSTM, (Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon. Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon, CogALex@COLING 2016, Osaka, Japan, December 12, 2016 (2016)), 51-56

[11] Yuan, D.; Richardson, J.; Doherty, R.; Evans, C.; Altendorf, E., Semi-supervised word sense disambiguation with neural models, (Proceedings of COLING (2016)), 1374-1385

[12] Raganato, A.; Delli Bovi, C.; Navigli, R., Neural sequence learning models for word sense disambiguation, (Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 (2017)), 1156-1167

[13] Pilehvar, M. T.; Navigli, R., A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation, Comput. Linguist., 40, 4, 837-881 (2014)

[14] Postma, M.; Bevia, R. I.; Vossen, P., More is not always better: balancing sense distributions for all-words word sense disambiguation, (Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (2016)), 3496-3506

[15] Pasini, T.; Navigli, R., Train-O-Matic: large-scale supervised word sense disambiguation in multiple languages without manual training data, (Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017)), 78-88

[16] Pasini, T.; Elia, F.; Navigli, R., Huge automatically extracted training-sets for multilingual word sense disambiguation, (Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018 (2018)), 1694-1698

[17] Pasini, T.; Navigli, R., Two knowledge-based methods for high-performance sense distribution learning, (Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 (2018)), 5374-5381

[18] Brin, S.; Page, L., The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst., 30, 1-7, 107-117 (1998)

[19] Pilehvar, M. T.; Jurgens, D.; Navigli, R., Align, disambiguate and walk: a unified approach for measuring semantic similarity, (Proceedings of ACL (2013)), 1341-1351

[20] Pilehvar, M. T.; Collier, N., De-conflated semantic representations, (Proceedings of the Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX (2016)), 1680-1690

[21] Lofgren, P. A.; Banerjee, S.; Goel, A.; Seshadhri, C., Fast-ppr: scaling personalized pagerank estimation for large graphs, (Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014), ACM), 1436-1445

[22] Navigli, R.; Ponzetto BabelNet, S. P., The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artif. Intell., 193, 217-250 (2012) · Zbl 1270.68299

[23] Camacho-Collados, J.; Pilehvar, M. T.; Navigli, R., NASARI: a novel approach to a semantically-aware representation of items, (Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2015), Association for Computational Linguistics: Association for Computational Linguistics , DenverColorado), 567-577

[24] Camacho-Collados, J.; Pilehvar, M. T.; Navigli, R., Nasari: integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, Artif. Intell., 240, 36-64 (2016) · Zbl 1386.68184

[25] Bennett, A.; Baldwin, T.; Lau, J. H.; McCarthy, D.; Bond, F., LexSemTm: a semantic dataset based on all-words unsupervised sense distribution learning, (Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany (2016)), 1513-1524

[26] Camacho-Collados, J.; Navigli, R., BabelDomains: large-scale domain labeling of lexical resources, (Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, vol. 2 (2017)), 223-228

[27] Ziemski, M.; Junczys-Dowmunt, M.; Pouliquen, B., The United Nations parallel corpus v1.0, (Chair, N. C.C.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S., Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (2016), European Language Resources Association (ELRA): European Language Resources Association (ELRA) Portoroz, Slovenia), 3530-3534

[28] Raganato, A.; Camacho-Collados, J.; Navigli, R., Word sense disambiguation: a unified evaluation framework and empirical comparison, (Proceedings of EACL. Proceedings of EACL, Valencia, Spain (2017)), 99-110

[29] Edmonds, P.; Cotton, S., SENSEVAL-2: overview, (The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, Association for Computational Linguistics (2001)), 1-5

[30] Snyder, B.; Palmer, M., The English all-words task, (Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain (2004)), 41-43

[31] Pradhan, S. S.; Loper, E.; Dligach, D.; Palmer, M., SemEval-2007 task 17: English lexical sample, SRL and all words, (Proceedings of the 4th International Workshop on Semantic Evaluations (2007)), 87-92

[32] Navigli, R.; Jurgens, D.; Vannella, D., SemEval-2013 task 12: multilingual word sense disambiguation, (Proceedings of the

7Th International Workshop on Semantic Evaluation (SemEval 2013), in Conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013). Proceedings of the 7Th International Workshop on Semantic Evaluation (SemEval 2013), in Conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), Atlanta, USA (2013)), 222-231

[33] Moro, A.; Navigli, R., SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking, (Proceedings of the 9th International Workshop on Semantic Evaluation. Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015 (2015)), 288-297

[34] Miller, G. A.; Leacock, C.; Tengi, R.; Bunker, R., A semantic concordance, (Proceedings of the 3rd DARPA Workshop on Human Language Technology. Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, N.J. (1993)), 303-308

[35] Taghipour, K.; Ng, H. T., One million sense-tagged instances for word sense disambiguation and induction, (Proceedings of the 19th Conference on Computational Natural Language Learning. Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015 (2015)), 338-344

[36] Camacho-Collados, J.; Delli Bovi, C.; Raganato, A.; Navigli, R., A large-scale multilingual disambiguation of glosses, (Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016. Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož Slovenia, May 23-28, 2016 (2016)), 1701-1708

[37] Delli Bovi, C.; Camacho-Collados, J.; Raganato, A.; Navigli, R., EuroSense: automatic harvesting of multilingual sense annotations from parallel text, (Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2 (2017)), 594-600

[38] Koehn, P., Europarl: a parallel corpus for statistical machine translation, (MT Summit, vol. 5 (2005)), 79-86

[39] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J., Efficient estimation of word representations in vector space, arXiv preprint

[40] Marcus, M. P.; Santorini, B.; Marcinkiewicz, M. A., Building a large annotated corpus of English: the penn treebank, Comput. Linguist., 19, 2, 313-330 (1993)

[41] Raganato, A.; Delli Bovi, C.; Navigli, R., Automatic construction and evaluation of a large semantically enriched Wikipedia, (Proceedings of IJCAI. Proceedings of IJCAI, New York City, NY, USA (2016)), 2894-2900

[42] Manion, S. L.; Sainudiin, R., An iterative "sudoku style" approach to subgraph-based word sense disambiguation, (Proceedings of the Third Joint Conference on Lexical and Computational Se- Mantics (*SEM 2014). Proceedings of the Third Joint Conference on Lexical and Computational Se- Mantics (*SEM 2014), Dublin, Ireland (2014)), 40-50

[43] Melamud, O.; Goldberger, J.; Dagan, I., Context2vec: learning generic context embedding with bidirectional LSTM, (Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016 (2016)), 51-61

[44] Luo, F.; Liu, T.; Xia, Q.; Chang, B.; Sui, Z., Incorporating glosses into neural word sense disambiguation, (Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1 (2018)), 2473-2482

[45] Kumar, S.; Jat, S.; Saxena, K.; Talukdar, P., Zero-shot word sense disambiguation using sense definition embeddings, (Proceedings of ACL (2019)), 5670-5681

[46] Otegi, A.; Aranberri, N.; Branco, A.; Hajic, J.; Neale, S.; Osenova, P.; Pereira, R.; Popel, M.; Silva, J.; Simov, K., QTLeap WSD/NED corpora: semantic annotation of parallel corpora in six languages, (Proceedings of the 10th Language Resources and Evaluation Conference, LREC (2016)), 3023-3030

[47] Koehn, P., Europarl: a parallel corpus for statistical machine translation, (MT Summit, vol. 5 (2005)), 79-86

[48] Agirre, E.; Branco, A.; Popel, M.; Simov, K., Europarl QTLeap WSD/NED Corpus (2015), LINDAT/CLARIN Digital Library at the Institute of Formal and Applied Linguistics, Charles University in Prague

[49] Scarlini, B.; Pasini, T.; Navigli, R., Just "OneSeC" for producing multilingual sense-annotated data, (Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1 (2019)), 699-709

[50] Agirre, E.; de Lacalle, O. L., Publicly available topic signatures for all WordNet nominal senses, (Proceedings of the Fourth International Conference on Language Resources and Evaluation. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 26-28, 2004 (2004)), 1123-1126

[51] Fernández, J.; Castillo Valdés, M.; Rigau Claramunt, G.; Atserias Batalla, J.; Tormo, J., Automatic acquisition of sense examples using exretriever, (IBERAMIA Workshop on Lexical Resources and the Web for Word Sense Disambiguation (2004)), 97-104

[52] Leacock, C.; Miller, G. A.; Chodorow, M., Using corpus statistics and WordNet relations for sense identification, Comput. Linguist., 24, 1, 147-165 (1998)

[53] Lesk, M., Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, (Proceedings of the 5th Annual Conference on Systems Documentation. Proceedings of the 5th Annual Conference on Systems Documentation, Toronto, Ontario, Canada (1986)), 24-26

[54] Banerjee, S.; Pedersen, T., An adapted Lesk algorithm for word sense disambiguation using WordNet, (International Conference on Intelligent Text Processing and Computational Linguistics (2002), Springer), 136-145 · Zbl 1044.68819

[55] Kilgarriff, A.; Rosenzweig, J., Framework and results for English SENSEVAL, Comput. Humanit., 34, 1-2, 15-48 (2000)

[56] Vasilescu, F.; Langlais, P.; Lapalme, G., Evaluating variants of the lesk approach for disambiguating words, (Proceedings of the Fourth International Conference on Language Resources and Evaluation. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 26-28, 2004 (2004)), 633-636

[57] Kilgarriff, A., How dominant is the commonest sense of a word?, (Text, Speech and Dialogue, 7th International Conference.

Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings (2004)), 103-112

[58] Agirre, E.; Martinez, D., Unsupervised WSD based on automatically retrieved examples: the importance of bias, (Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)), 25-32

[59] Mohammad, S.; Hirst, G., Determining word sense dominance using a thesaurus, (Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3-7 April 2006 (2006)), 121-128

[60] McCarthy, D.; Koeling, R.; Weeds, J.; Carroll, J., Finding predominant senses in untagged text, (Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21-26 July 2004 (2004)), 280-287

[61] Chan, Y. S.; Ng, H. T., Word sense disambiguation with distribution estimation, (IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence. IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5 (2005)), 1010-1015

[62] Chan, Y. S.; Ng, H. T., Estimating class priors in domain adaptation for word sense disambiguation, (ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006 (2006)), 89-96

[63] Lau, J. H.; Cook, P.; McCarthy, D.; Gella, S.; Baldwin, T., Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models, (Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1 (2014)), 259-270

[64] Di Fabio, A.; Conia, S.; Navigli, R., VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling, (Proceedings of EMNLP-IJCNLP (2019))