Finding semantic similarity in Vietnamese

Nguyen D.T., Pham S.B.
Faculty of Information Technology, Information Technology Institute, Vietnam National University, Hanoi, Viet Nam

Abstract: Finding semantic similarity is an important task in many natural language processing applications. Despite numerous works for popular languages, there is still limited research done for Vietnamese. In this paper, we tackle the problem of finding semantic similarity for Vietnamese using Random Indexing and Hyperspace Analogue to Language to represent the semantics of words and documents. We build a system to find synonyms in Vietnamese. Experimental results show that our system achieves accuracies of 75% for finding synonyms for verbs and 65% for synonyms for nouns. © 2010 IEEE.
Author Keywords: Hyperspace Analogue to Language; Random projection; Semantic vector; Word space model
Index Keywords: Hyperspace analogue to languages; NAtural language processing; Random indexing; Random projection; Semantic similarity; Word space model; Computational linguistics; Semantics; Vector spaces; Natural language processing systems

Year: 2010
Source title: Proceedings - 2010 International Conference on Asian Language Processing, IALP 2010
Art. No.: 5681551
Page : 91-94
Link: Scorpus Link
Correspondence Address: Nguyen, D. T.; Faculty of Information Technology, Information Technology Institute, Vietnam National University, Hanoi, Viet Nam; email: datnt88@gmail.com
Sponsors: Proj. Int. Coop. Exch. Natl. Nat. Sci. Found. China (NSFC);Key Proj.Next Gener. Inf. Retr." (No. 60736044);Natl. Nat. Sci. Found. China (NSFC);Heilongjiang Institute of Technology (HIT)"
Conference name: 2010 International Conference on Asian Language Processing, IALP 2010
Conference date: 28 December 2010 through 30 December 2010
Conference location: Harbin
Conference code: 83700
ISBN: 9.78077E+12
DOI: 10.1109/IALP.2010.78
Language of Original Document: English
Abbreviated Source Title: Proceedings - 2010 International Conference on Asian Language Processing, IALP 2010
Document Type: Conference Paper
Source: Scopus
Authors with affiliations:
• Nguyen, D.T., Faculty of Information Technology, Information Technology Institute, Vietnam National University, Hanoi, Viet

Nam
- Pham, S.B., Faculty of Information Technology, Information Technology Institute, Vietnam National University, Hanoi, Viet Nam

References:

- Salton, G., McGill, M.J., (1986) Introduction to Modern Information Retrieval, , McGraw-Hill, Inc. New York, NY, USA
- Livesay, K., Burgress, C., Producing high-dimensional semantic spaces from lexical co-occurrence (1997) Behavior Research Methods Instruments Computers, 28, pp. 203-208
- Sahlgren, M., An introduction to random indexing (2005) Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, , Copenhagen, Denmark
- Banerjee, S., Pedersen, T., Extended gloss overlaps as a measure of semantic relatedness.in (2003) IJCAI, pp. 805-810
- Jiang, J.J., Conrath, D.W., Semantic similarity based on corpus statistics and lexical taxonomy (1997) ROCLING'97
- Resnik, P., Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language (1999) JAIR 11, pp. 95-130
- Roark, B., Charniak, E., Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction (1998) Proceedings of COLING-ACL'98
- Thelen, M., Riloff, E., A bootstrapping method for learning semantic lexicons using extracting pattern contexts (2002) Proceedings of EMNLP'02
- Phillips, W., Riloff, E., Exploiting strong syntactic heuristics and co-training to learn semantic lexicons (2002) Proceedings of EMNLP'02
- Yarowsky, D., Unsupervisedword sense disambiguation rivaling supervisedmethods (1995) Proceedings of ACL'95, pp. 189-196
- Van Der Plas, L., Tiedemann, J., Finding synonyms using automatic word alignment and measures of distributional similarity (2006) Association for Computational Linguistics Morristown, , NJ, USA
- Curran, J., Moens, M., Improvements in automatic thesaurus extraction (2002) Proceedings of the Workshop on Unsupervised Lexical Acquisition, pp. 59-67
- Van Der Plas, L., Bouma, G., Syntactic contexts for finding semantically similar words (2005) Proceedings of the Meeting of Computational Linguistics in the Netherlands (CLIN)
- Landauer, T.K., Dumais, S.T., A solution to plato's problem (1997) Psychological Review, 104, pp. 211-240
- Schuetze, H., Dimensions of meaning (1992) Proceedings of Supercomputing'92, pp. 787-796
- Ando, R.K., Semantic lexicon construction: Learning from unlabeled data via spectral analysis (2004) HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL- 2004)
- Landauer, T.K., Foltz, P.W., Laham, D., (1998) An Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259-284
- Pham, D.D., Tran, G.B., Pham, S.B., A hybrid approach to Vietnamese word segmentation using part of speech tags (2009) International Conference on Knowledge and Systems Engineering
- Kanerva, P., Kristoferson, J., Holst, A., Random indexing of text samples for latent semantic analysis (2000) Proc. of the 22nd Annual Conference of the Cognitive Science Society, , Erlbaum (editor), New Jersey, USA