Exploiting Macro and Micro Relations toward Web Intelligence

Mitsuru Ishizuka

School of Information Science and Technology
Univerecity of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@i.u-tokyo.ac.jp

Keywords: Relation extraction/mining, relational similarity, relational search, semantic computing, Web intelligence.

Relations are basic elements for representing knowledge, such as in semantic network, logic and others. In Web intelligence research, the extraction or mining of meaningful knowledge and the utilization of the knowledge for intelligent services are key issues. In this talk, I will present some of our researches related to these issues, ranging from macro relations to micro ones. Here we mostly use Web texts, and the use of their huge data though a search engine becomes a key function together with text analysis.

The first topic concerns with the extraction of human-human and company-company relations from the Web [1-14]. Relation types between two entities are also extracted here. An open Web service based on this function has been operated in Japan by a company. One technology related to this one is namesake disambiguation [15-17].

Wikipedia is a good reliable source for wide knowledge, unlike other Web information. In order to extract the knowledge or data from Wikipedia in the form that computers can understand and manipulate, several attempts including ours [18-23] have been carried out, typically to extract triplets such as (entity, attribute, value).

After we worked on computing similarity between two words based on the distributional hypothesis [24, 25], we have been interested in computing similarity between two word pairs (or two entity pairs) [26-28]. Like in the previous case, we are mainly utilizing distributional hypothesis, and have invented an efficient clustering method for dealing with several tens of thousands of lexical patterns. Based on this mechanism, we have implemented a latent relational search engine, which accepts two entity pairs with one missing component such as {(Tokyo, Japan), (?, France)} as a query, and produces an answer such as (? = Paris) with its evidence. As an extension of this mechanism, we recently invented an efficient co-clustering method, which works well to find arbitrary existing relations between two nouns in sentences [29]. This problem setting is called open information extraction (open IE).

B.-T. Zhang and M.A. Orgun (Eds.): PRICAI 2010, LNAI 6230, pp. 4–7, 2010.

The final topic of the talk is Concept Description Language (CDL), which has been designed to serve as a common language for representing concept meaning expressed in natural language texts [30-32]. Unlike Semantic Web which provides machine-readable meta-data in the form of RDF, CDL aims to encode the meaning of the whole texts in a machine-understandable form. The basic representation element in CDL is micro relations existing between entities in the text; 44 relation types are defined. CDL has been discussed in a W3C incubator group for international standardization since 2007. It is intended to be a basis of semantic computing in next generation, and also become a medium for overcoming language barrier in the world. Current issues of CDL are, among others, an easy semi-automatic way of converting natural language texts into the CDL description, and an effective mechanism of semantic retrieval on the CDL database.

References

1. Matsuo, Y., Tomobe, H., Hasida, K., Ishizuka, M.: Mining Social Network of Conference Participants from the Web. In: Proc. 2003 IEEE/WIC Int'l Conf. on Web Intelligence (WI 2003), Halifax, Canada (2003)
2. Mori, J., Matsuo, Y., Ishizuka, M., Faltings, B.: Keyword Extraction from the Web for FOAF Metadata. In: Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland, pp. 1–8 (2004)
3. Mori, J., Sugiyama, T., Matsuo, Y., Tomobe, H., Ishizuka, M.: Real-world Oriented Information Sharing using Social Network. In: Proc. 25th Int'l Sunbelt Social Network Conf, SUNBELT XXV (2005)
4. Mori, J., Matsuo, Y., Hashida, K., Ishizuka, M.: Web Mining Approach for a User-centered Semantic Web. In: Proc. Int'l Workshop on User Aspects on the Semantic Web in 2nd European Semantic Web Conf. (ESWC 2005), Heraklion, Greek, pp. 177–187 (2005)
5. Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M.: POLYHONET: An Advanced Social Network Extraction System from the Web. In: Proc. 15th World Wide Web Conf. (WWW 2006), Edinburgh, UK (2006) (CD-ROM)
6. Matsuo, Y., Hamasaki, M., Nakamura, Y., Nishimura, T., Hasida, K., Takeda, H., Mori, J., Bollegala, D., Ishizuka, M.: Spinning Multiple Social Networks for Semantic Web. In: Proc. 21st National Conf. on Artificial Intelligence (AAAI 2006), Boston, MA, USA, pp. 1381–1387 (2006)
7. Mori, J., Tsujishita, T., Matsuo, Y., Ishizuka, M.: Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 487–500. Springer, Heidelberg (2006)
8. Mori, J., Matsuo, Y., Ishizuka, M.: Extracting Keyphrases to Represent Relations in Social Networks from Web. In: Proc. 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007), Hyderabad, India, pp. 2820–2825 (2007)
9. Matsuo, Y., Mori, J., Ishizuka, M.: Social Network Mining from the Web. In: Poncelet, P., Teisseire, M., Masseglia, F. (eds.) Data Mining Patterns – New Methods and Applications, ch. VII, pp. 149–175. Information Science Reference (2007)
10. Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M.: POLYPHONET: An Advanced Social Network Extraction System from the Web. Journal of Web Semantics 5(4), 262–278 (2007)

11. Jin, Y., Matsuo, Y., Ishizuka, M.: Extracting a Social Network among Entities by Web Mining. In: Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies, Athens, GA, USA, 10 p. (2006)
12. Jin, Y., Matsuo, Y., Ishizuka, M.: Extracting Social Networks among Various Entities on the Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 251–266. Springer, Heidelberg (2007)
13. Jin, Y., Ishizuka, M., Matsuo, Y.: Extracting Inter-firm Networks from the World Wide Web using a General-purpose Search Engine. Information Review 32(2), 196–210 (2008)
14. Jin, Y., Matsuo, Y., Ishizuka, M.: Ranking Companies Based on Multiple Social Networks Mined from the Web. In: Kang, K. (ed.) E-commerce, INTECH, ch. 6, pp. 75–98 (2010)
15. Bollegala, D., Matsuo, Y., Ishizuka, M.: Extracting Key Phrases to Disambiguate Personal Names on the Web. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 223–234. Springer, Heidelberg (2006)
16. Bollegala, D., Matsuo, Y., Ishizuka, M.: Extracting Key Phrases to Disambiguate Personal Name Queries in Web Search. In: Proc. of the Workshop "How can Computational Linguistics Improve Information Retrieval?" at the Joint 21st Int'l Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), Sydney, Australia, pp. 17–24 (2006)
17. Bollegala, D., Matsuo, Y., Ishizuka, M.: Disambiguating Personal Names on the Web using Automatically Extracted Key Phrases. In: Proc. European Conf. on Artificial Intelligence (ECAI 2006), Trento, Italy, pp. 553–557 (2006)
18. Nguyen, D.P.T., Matsuo, Y., Ishizuka, M.: Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia. In: Proc. IJCAI 2007 Workshop on Text-Mining and Link-Analysis (TextLink 2007), Hyderabad, India, 11 p. (2007) (CD-ROM)
19. Nguyen, D.P.T., Matsuo, Y., Ishizuka, M.: Subtree Mining for Relation Extraction from Wikipedia. In: Companion Volume of Proc. of the Main Conf. of Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007), Rochester, New York, pp. 125–128 (2007)
20. Nguyen, D.P.T., Matsuo, Y., Ishizuka, M.: Relation Extraction from Wikipedia Using Subtree Mining. In: Proc. 22nd Conf. on Artificial Intelligence (AAAI 2007), pp. 1414–1420 (2007)
21. Watanabe, K., Bollegala, D., Matsuo, Y., Ishizuka, M.: A Two-Step Approach to Extracting Attributes for People on the Web. In: Proc. WWW 2009 2nd Web People Search Evaluation Workshop (WEPS 2009), Madrid, Spain, 6 p. (2009)
22. Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., Ishizuka, M.: Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In: Proc. of Joint Conf. of 47th Annual Meeting of the Association for Computational Linguistics and 4th Int'l Joint Conf. on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), Singapore, pp. 1021–1029 (2009)
23. Yan, Y., Li, H., Matsuo, Y., Ishizuka, M.: Multi-view Bootstrapping for Relation Extraction by Exploiting Web Features and Linguistic Features. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 525–536. Springer, Heidelberg (2010)
24. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engines. In: Proc. 16th Int'l World Wide Web Conf. (WWW 2007), Banff, Canada, pp. 757–766 (2007)
25. Bollegala, D., Matsuo, Y., Ishizuka, M.: WWW sits the SAT: Measuring Relational Similarity from the Web. In: Proc. 18th European Conf. on Artificial Intelligence (ECAI 2008), Patras, Greece, pp. 333–337 (2008)

26. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring the Similarity between Implicit Semantic Relations using Web Search Engines. In: Proc. 2009 Second ACM Int'l Conf. on Web Search and Data Mining (WSDM 2009), Barcelona, Spain, pp. 104–113 (2009) (CD-ROM)
27. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring the Similarity between Implicit Semantic Relations from the Web. In: Proc. 18th Int'l World Wide Web Conf. (WWW 2009), Madrid, Spain, pp. 651–660 (2009)
28. Bollegala, D., Matsuo, Y., Ishizuka, M.: A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web. In: Proc. 2009 Conf. on Empirical Methods in Natural Language Processing (EMNLP 2009), Singapore, pp. 803–812 (2009)
29. Bollegala, D., Matsuo, Y., Ishizuka, M.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. In: Proc. 19th Int'l World Wide Web Conf. (WWW 2010), Raleigh, North Carolina, USA, pp. 151–160 (2010)
30. Report of W3C Incubator Group on Common Web Language (2008), http://www.w3.org/2005/Incubator/cwl/XGR-cwl-20080331/
31. Yan, Y., Matsuo, Y., Ishizuka, M., Yokoi, T.: Annotating an Extension Layer of Semantic Structure for Natural Language Text. In: Proc. 2nd IEEE Int'l Conf. on Semantic Computing, Santa Clara, CA, USA, pp. 174–181 (2008) (CD-ROM)
32. Yan, Y., Matsuo, Y., Ishizuka, M., Yokoi, T.: Relation Classification for Semantic Structure Annotation of Text. In: Proc. 2008 IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2008), Sydney, Australia, pp. 377–380 (2008) (CD-ROM)