Extracting parallel texts from the web

Hung L.Q., Cuong L.A.
Faculty of Information Technology, Quynhon University, Viet Nam; University of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam

Abstract: Parallel corpus is the valuable resource for some important applications of natural language processing such as statistical machine translation, dictionary construction, cross-language information retrieval. The Web is a huge resource of knowledge, which partly contains bilingual information in various kinds of web pages. It currently attracts many studies on building parallel corpora based on the Internet resource. However, obtaining a parallel corpus with high accuracy is still a challenge. This paper focuses on extracting parallel texts from bilingual web-sites of the English and Vietnamese language pair. We first propose a new way of designing content-based features, and then combining them with structural features under a framework of machine learning. In the experiment we obtain 88.2% of precision for the extracted parallel texts. © 2010 IEEE.

Index Keywords: Bi-lingual information; Content-based features; Cross language information retrieval; Dictionary constructions; Internet resources; Language pairs; Machine-learning; NAtural language processing; Parallel corpora; Parallel text; Statistical machine translation; Structural feature; Web page; Computational linguistics; Information retrieval; Learning algorithms; Software agents; Systems engineering; Natural language processing systems

Year: 2010
Source title: Proceedings - 2nd International Conference on Knowledge and Systems Engineering, KSE 2010
Art. No.: 5632135
Page : 147-151
Link: Scorpus Link
Correspondence Address: Hung, L. Q.; Faculty of Information Technology, Quynhon UniversityViet Nam; email: hungqnu@gmail.com
Conference name: 2nd International Conference on Knowledge and Systems Engineering, KSE 2010
Conference date: 7 October 2010 through 9 October 2010
Conference location: Hanoi
Conference code: 83923
ISBN: 9.78077E+12
DOI: 10.1109/KSE.2010.14
Language of Original Document: English
Abbreviated Source Title: Proceedings - 2nd International Conference on Knowledge and Systems Engineering, KSE 2010
Document Type: Conference Paper
Source: Scopus

Authors with affiliations:

- Hung, L.Q., Faculty of Information Technology, Quynhon University, Viet Nam
- Cuong, L.A., University of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam

References:

- Kumano, A., Hirakawa, H., Building an MT dictionary from parallel texts based on linguisitic and statistical information (1994) Proc. 15th COLING, pp. 76-81
- Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roosin, P., A statistical approach to machine translation (1990) Computational Linguistics, 16 (2), pp. 79-85
- Chen, J., Nie., J.Y., Automatic construction of parallel English-Chinese corpus for cross-language information retrieval (2000) Proc. ANLP, pp. 21-28. , Seattle
- Chen, J., Chau, R., Yeh, C.-H., Discovering parallel text from the world wide web (2004) Proc. Australasian Workshop on Data Mining and Web Intelligence (DMWI2004)
- Davis, M., Dunning, T., A TREC evaluation of query translation methods for multi-lingual text retrieval (1995) Fourth Text Retrieval Conference (TREC- 4), , NIST
- Volk, M., Vintar, S., Buitelaar, P., Ontologies in cross-language information retrieval (2003) Wissensmanagement, pp. 43-50
- Melamed, I.D., Word-to-word models of translation equivalence (1998) IRCS Technical Report 98-08, , University of Pennsylvania
- Simard, M., Foster, G.F., Isabelle, P., Rfeti Using Cognates to Align Sentences in Bilingual Corpora
- Oard, D.W., Cross-language text retrieval research in the USA (1997) Third DELOS Workshop European Research Consortium for Informatics and Mathematics
- Resnik, P., Smith, N.A., The Web as a Parallel Corpus (2003) Computational Linguistics, 29 (3), pp. 349-380. , DOI 10.1162/089120103322711578
- Resnik, P., Parallel strands: A preliminary investigation into mining the Web for bilingual text (1998) Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA-98), pp. 28-31. , Langhorne, PA, October
- Resnik, P., Mining the Web for bilingual text (1999) Proceedings of the 37th Annual Meeting of the ACL, pp. 527-534. , College Park, MD, June
- Utsuro, T., Yamane, H.I.M., Matsumoto, Y., Nagao, M., Bilingual text matching using bilingual dictionary and statistics (1994) Proc. 15th COLING, pp. 1076-1082
- Van Dang, B., Ho, B.-Q., Automatic construction of english-Vietnamese parallel corpus through web mining (2007) Proceedings of 5th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future (RIVF'2007), , Hanoi, Vietnam
- Ma, X., Mark, L., BITS: A method for bilingual text search over the Web (1999) Machine Translation Summit VII, , September, 1999