Abstract

With the problem of increased web resources and the huge amount of information available, the necessity of having automatic summarization systems appeared. Since summarization is needed the most in the process of searching for information on the web, where the user aims at a certain domain of interest according to his query, domain-based summaries would serve the best. Despite the existence of plenty of research work in the domain-based summarization in English, there is lack of them in Arabic due to the shortage of existing knowledge bases. In this paper an Ontology-based Summarization System for Arabic Documents, OSSAD, is introduced. Domain knowledge is extracted from an Arabic corpus and represented by topic related concepts/keywords and the lexical relations among them. The user’s query is first expanded by using the Arabic WordNet and then by adding the domain-specific knowledge
base to the expansion. For summarization, decision tree algorithm (C4.5) is used, which was trained by a set of features extracted from the original documents. For the testing dataset, Essex Arabic Summaries Corpus (EASC) was used. Recall Oriented Understudy for Gisting Evaluation (ROUGE) was used to compare OSSAD summaries with the human summaries along with other automatic summarization systems, showing that the proposed approach demonstrated promising results.

References

- William BLACK, Sabri ELKATEB, "Introducing the Arabic WordNet Project,"
Third International WordNet Conference (GWC-06), Korea, 2006.
- Xing Jiang, Ah-Hwee Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents", Data Mining, Fifth IEEE International Conference, Singapore, 2005.
- Philipp Cimiano, Johanna Völker, "Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery", 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Spain, pp. 227-238, 2005.
- Kavita Ganesan, ChengXiang Zhai, Jiawei Han, "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions", the 23rd International Conference on Computational Linguistics (COLING &apos;10), China, 2010.
- Jonas Sjobergh, "Older versions of the ROUGE evaluation system were easier to fool", the International Journal of Information Processing and Management, Vol. 43, No. 6, pp. 1500-1505, 2007.

**Index Terms**

Computer Science

Pattern Recognition
Keywords

Arabic text summarization  Knowledge-based summarization  Query expansion
Ontology extraction from text

Arabic WordNet