2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2016, 2017, pages 285-290

Corpus management system: Semantic aspects of representation and processing of search queries

Nevzorova O., Mukhamedshin D., Galieva A., Gataullin R.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2016 IEEE. There are several well-known corpus management systems (Sketch Engine, Manatee, EXMARaLDA, etc.). The system presented in this article has search functionalities comparable to those. However, it also takes into account certain specifics of Turkic languages. The Tatar corpus management system (http://corpus.antat.ru) is specifically designed to work with Turkic linguistic corpora. Functionality offered by the corpus management system includes search of lexical units, morphological and lexical search, search of syntactic units, search of the n-gram based on grammar and others. The semantic model of the Tatar language data representation is the core of the system. The search is performed using open source tools (database management system MariaDB, Redis data store). The Tatar language has a complicated agglutinative morphology; and we consider the system of grammatical categories represented in grammatical annotation of the Tatar corpus as a key to semantics of the language. Selecting and combining grammatical, lexical and other parameters of a query, we may get certain sets of semantic samples from semantically unstructured corpus data. The main task of our research is detecting and describing a class of grammatically conditioned semantic phenomena and developing a system of queries to the corpus for extraction of these semantic phenomena. Experiments with queries to the Tatar corpus show that semantically relevant combinations of query parameters may differ by level of complexity. The results of the work may be used for document clustering and classification, as well as for Tatar grammar building and other purposes.

http://dx.doi.org/10.1109/SETIT.2016.7939881

Keywords

Corpus manager, grammar, morphological formulas, query, semantic information, the Tatar language

References

[1] Cormen, T. H. Introduction to algorithms. MIT press. 253-280 (2009).

[2] Davies, M. The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, linguistic insights. International journal of corpus linguistics, 14(2), 159-190 (2009).

[3] Han, J., Haihong, E., Le, G., Du, J. Survey on NoSQL database. In Pervasive computing and applications (ICPCA), 2011 6th international conference on 26-28 Oct. 2011 (pp. 363-366). IEEE (2011).

[4] MySQL 5. 6 Reference Manual: 19 Partitioning, Oracle Corporation. http://dev. mysql. com/doc/refman/5. 6/en/partitioning. html.

[5] Rychly, P. Corpus managers and their effective implementation (Doctoral dissertation, PhD thesis, Faculty of Informatics, Masaryk University, Brno) (2000).

[6] Rychlý, P. Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (pp. 65-70). within MU: Faculty of Informatics Further information (2007).

[7] Kilgarriff, A., Baisa, V., Bušta, J., Jakubiék, M., Ková, V., Michelfeit, J., & Suchomel, V. The Sketch Engine: Ten years on. Lexicography, 1(1), 7-36 (2014).

[8] Schmidt, T., Wörner, K. EXMARaLDA-Creating, analyzing and sharing spoken lan-guage corpora for pragmatics research. Pragmatics-Quarterly Publication of the International Pragmatics Association, 19(4), 565 (2009).

[9] Brik, M., Touahria, M. EduBank: A bank of educational resources based on ontologies. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, Sousse (pp. 92-96). IEEE (2012).

[10] Haboubi, S., Maddouri, S., Amiri, H. Word classification in bilingual printed documents. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, Sousse (pp. 502-506). IEEE (2012).

[11] Jaffali, S., Jamoussi, S. Principal component analysis neural network for textual document categorization and dimension reduction. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, Sousse (pp. 835-839). IEEE (2012).

[12] Ali, S. H. Miner for OACCR: Case of medical data analysis in knowledge discovery. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, Sousse (pp. 962-975). IEEE (2012).

[13] Kilgarriff, A., Baisa, V., Bušta, J., Jakubiék, M., Ková, V., Michelfeit, J.,. & Suchomel, V. The Sketch Engine: Ten years on. Lexicography, 1(1) (pp. 7-36) (2014).

[14] Kilgarriff, A., Jakubcek, M., Kovár, V., Rychlý, P., Suchomel, V. Finding terms in corpora for many languages with the Sketch Engine. Proceedings of the Demonstrations at the 14th Conferencethe European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, The Association for Computational Linguistics (pp. 53-56) (2014).

[15] Ken, M. Recent Developments in the Czech National Corpus. Challenges in the Management of Large Corpora (CMLC-3), 1 (2015).

[16] Zakharov, V. Corpora of the Russian Language. In International Conference on Text, Speech and Dialogue (pp. 1-13). Springer Berlin Heidelberg (2013, September).