УДК 81:316.77

М. А. Попова, И. В. Ермолина
Уральский федеральный университет
им. первого Президента России Б. Н. Ельцина
г. Екатеринбург, Россия

Лингвистика и выявление плагиата

Статья посвящена проблеме автоматизации процесса выявления плагиата в интернет источниках, как одной из множества путей взаимодействия между исследователями в области информационных технологий, лингвистики, компьютерной семантики и других прикладных дисциплин.

Linguistics and plagiarism detection

The article aims to outline the trends of plagiarism detection in automatic systems for detecting plagiarism from web sources, thereby indicating one of the most perspective ways of collaboration between information studies and linguistics in academic research.

The problem of plagiarism detection has become a part of educational process in our information society where everyone has an unlimited access to the sources of the Internet overflowed with data of the unknown origin. That's why very often the procedure of students' theses verification on the subject of originality may result in revelations, such as passages of information cited from unreliable sites, the so-called "third party – a setting which has not been studied so far" [1], is completely unknown or resigned in the academic world (as distinct from credible domains: *.org, *.com, *.net, *.edu, *.gov, *.il.us, *.uk, for example) [2].

The amount of trickier cases, when the source of information is omitted and the text presented in a paper is copy-pasted and transformed makes it difficult for a human to recognize plagiarism (assuming the volume of each given paper and human factors that obstruct us from reading the whole text at once). In terms of this the notion of "intelligent" plagiarism occurs, meaning that this type of plagiarism is less exposed to detection [3]. The procedure of experimental automatic detection of plagiarism is named "artificial" plagiarism due to the algorithmic principle of programming and generating suspicious documents (by means of text-

summarizing and other techniques), as opposed to "simulated" plagiarism – human reformulation of passages from a document [4]. Any text composed by a human being in terms of information retrieval studies (another corresponding field) is "unstructured". "The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure" [5]. However, within the last few years computer science has achieved several successes in the development of computer-readable formats (PDF, XML for text-documents) [6], in natural language processing (NLP), in the form of IBM Watson, a supercomputer capable of human speech recognition and confidence-rated responsing, and in devising smarter plagiarism detectors [7].

Thus during the 5th International Competition on Plagiarism Detection there were 18 plagiarism detectors tested on the efficiency of text retrieval and text alignment of external plagiarism detection, as reported in the overall of the event [8]. Aiming to find solutions for applied and narrow tasks, namely the collection of a wholesome database of plagiarised documents, the researchers were interested in the improved characteristics of latest software versions, as well as in the possibilities to learn more about a real plagiarist's behaviour. The accomplishment of this task was seen by researchers in the introduction of cyclic translation and text-summarizing techniques extending the mentioned goals to the field of machine translation and document understanding, studies largely based on principles of computational semantics [9].

The applied character of computational linguistics allows scientists to deal with a variety of tasks successfully so far, however when dealing with software renovation one should bear in mind that the tasks will occur eternally, so that collaboration between information studies and linguistics promises a boundless set of targets in a variety of branches, such as machine learning, document understanding, natural language processing, automated text summarizing, information retrieval and plagiarism detection.

ЛИТЕРАТУРА

1. Crowdsourcing interaction logs to understand text-reuse from the web / M. Potthast, M. Hagen, M. Völske, B. Stein // Proceedings of the 51st Annual Meeting for Computational Linguistics. 2013. URL: http://webis.de/ (дата обращения: 15.01.2014).

2. Evaluating internet sources. Tips and tricks for evaluating web sites / University library. University of Illinois at Urbana-Champaign.

URL: http://www.library.illinois.edu/ugl/howdoi/webeval.html (дата обращения: 18.01.2014).

3. Advances towards semantic plagiarism detection / H. Issa, K. Hose, S. Metzger, R. Schenkel // Working notes of the LWA 2011 – Learning, knowledge, adaptation. URL: http://www.mpi-inf.mpg.de/~khose/publications/WIR2011-plagiarism.pdf (дата обращения: 15.01.2014).

4. Völske M. Analyzing a large corpus of crowdsourced plagiarism // Master's thesis, 2013. URL: http://www.uni-weimar.de/medien/webis/teaching/theses/voelske_2013.pdf (дата обращения: 10.01.2014).

5. Manning Christopher D. Raghavan Prabhakar, Schütze Hinrich Introduction to information retrieval. Cambrigde University Press, 2008. URL: http://nlp.stanford.edu/IR-book/ (дата обращения: 8.01.2014).

6. A primer on machine-readability for online documents and data (2012). URL: https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data (дата обращения: 04.01.2014).

7. IBM Watson Ecosystem Program (2013). URL: http: //www-03.ibm.com/innovation/us/watson/pdf/IBM_Watson_Ecosystem_program_11_4_2013.pdf (дата обращения: 19.01.2014).

8. Overview of the 5th International Competition on Plagiarism Detection (2013). URL: http: //www.uni-weimar.de/medien/webis/publications/papers/stein_2013h.pdf (дата обращения: 05.01.2014).

УДК 81:316.77

В. А. Райскина, Т. В. Сластникова
Московский городской педагогический университет
г. Москва, Россия

Аффиксация как экспрессивный способ словообразования в молодежной речи

В статье рассматриваются различные способы формирования арготической лексики. Подробно представлены ее аффиксальные возможности. Аффиксация, являясь одним из наиболее продуктивных способов словообразования во французском арго, помимо функции создания новых арготизмов осуществляет функцию эмоционально-