Information Retrieval and Machine Translation in Solving the Task of Automatic Recognition of Adopted Fragments of Text Documents

Yury Krapivin
Brest State Technical University
Brest, Belarus, 224000
ybox@list.ru

***Abstract*—The solution for the task of automatic recognition of adopted fragments in multilingual information environment with the cross-language functionality and focus on the detection of both explicit and implicit adoptions of text fragments by means of the well-developed linguistic text analysis, that is based on the knowledge of natural language, together with the existed effective information retrieval and machine translation tools is proposed in the article.**

I. Introduction

The term "information retrieval" was proposed by Calvin Mooers in the late forties of the previous century and denoted a package of measures with the aim to automate the process of searching for information in unstructured text documents (data searching) and searching for documents (document retrieval). Information retrieval (IR) is usually treated as a process of searching and providing the user with the information according to request that represents his information need [1].

The main tasks of IR are:

- Classical retrieval, comprising automatic indexing of documents and users' requests (according to Moore's definition).
- Retrieval of documents matched to the input one, when the request is represented by some document-pattern used to search the texts similar to it in content. This type of retrieval can lead up to more precise results in comparison with classical retrieval [2].
- Topical retrieval, which covers the topical filtration of documents, taking in to account the tonality and stylistic nuance, and is applied to detect the texts with special or peculiar vernacular vocabulary. The systems, applicative to solve these tasks, are focused on the specific universe of discourse that permits to carry out deep retrieval on a certain theme.
- Clustering and classifying of documents are auxiliary technologies of informational retrieval used for more effective representation of its results via automatic classes identification of input request (the task of clustering) or via adding every detected document to one of predefined categories (the task of classifying).

A qualitative solution of all tasks, mentioned above, needs to involve quite effortful approaches of textual information analysis, based on natural language knowledge, usually enclosed in dictionaries, grammars, organizing rules of syntactic and semantic structures from words and phrases, etc., which constitute linguistic knowledge bases. The latter, by-turn, are the foundations of linguistic processors functioning – the most powerful automatic text processing tools, without which, at the moment, it cannot do a single automatic text processing task [3].

A. The task of automatic recognition of adopted text fragments

An actual task of automatic recognition of adopted text fragments (plagiarism) immediately refers to the task of retrieval of documents matched to the input one.

At present, there are some systems of adopted text fragments recognition (e.g. AntiPlagiat [4], Ephorus [5], WCopyFind [6], JPlag [7], Copyscape [8] etc.), which are based on the algorithms, implemented according to widely-spread approaches: strings coincident, feature-quantification, information retrieval [9].

The approaches mentioned above are mostly focused on solving the task of automatic recognition of adopted text fragments either from the point of view that takes into consideration a relevance of proper text documents according to a certain specified similarity measure, or that takes into consideration explicitly undefined lexical adoptions of fragments, which means the same text fragment that belongs to different text documents, as well as minimal discrepancies may be allowed, for example, due to the usage of parenthesis, synonyms etc. Those solutions are focused on the recognition of lexical adoptions of text fragments, which take into account simple morphological transformations and synonymy relations, and don't apply a well-developed linguistic analysis of text documents, as well as don't propose to take into consideration more complicated text modifications (for

example, the usage of voice synonyms and different synonymous constructions at a noun-phrases level, object-parametrical relations; paraphrase etc.,). They don't provide any task solution in the case of implicit (semantic) adoption that refers to the fragments of different text documents of the same meaning, expressed via various chains of symbols, in relation to a specified system of knowledge.

A huge number of text documents in different languages from the Internet and full-text databases, on the one hand, and the availability and simplification of tools for their processing, on the other, substantially complicate a qualitative solution of the task of automatic recognition of adopted text fragments, due to the need of identifying both text fragments of documents, represented in the language of an input document, and in other languages from the analyzed language set, translated to the language of an input document beforehand, in the analyzing document. All this is referred to as cross-language functionality. The quality of the translated equivalent is the most important thing, and the approaches of recognition of adopted text fragments used further will greatly depend on it.

The usage of machine translation systems with the objective to translate one of the couple of analyzing documents beforehand to the language of the second one seems to be the most natural. It's needed to take into consideration the fact that these systems need to be multilingual, and therefore, the concept of building such a complex machine translation system should take into account many characteristics of the problem, including the state of the utilized natural languages distribution in relation to the text search database. Thus, the existed applied solutions (for example, Promt [10], SysTran [11], Retrans [12], Belazar [13], SMP B/R [3], Yandex.Translate [14], Google.Translate [15] etc.) can be utilized for the machine translation of text documents. They implement, as a rule, the following fundamental approaches: statistical approaches [16, 17], including example-based methods [18, 19], and linguistic approaches – rule-based methods [20, 21].

The first group of methods implies the usage of parallel text corpora as the basis for equivalence calculations of lexical items in different languages and their statistical characteristics, as well as certain translation model construction. The second implies a translation model construction according to the set of linguistic rules, which specify the necessary depth of text analysis, as well as the admissible transformation of the grammatical structure of the input text into the equivalent structures of the output one.

The performed analysis, as well as the experience of application of machine translation tools, which are available in open access in the Internet, has revealed that until recently their qualitative characteristics for arbitrary texts were relatively low. The certain machine translation systems for "too close" natural languages (for example, Russian and Belarusian) are the only exception. Therefore, recently the researchers have made some efforts to solve the designated problem. In the context of overall task, the alternative solutions, related to methods, which take into consideration syntactical characteristics of natural languages [22] and utilize various dictionaries and thesauri [23, 24], as well as comparable and parallel corpora [25, 26], are proposed.

In this context the choice of an acceptable task solution provides the implementation of cross-language functionality in the system of automatic recognition of adopted text fragments and depends on the cardinality of natural languages set, the volume of analyzing documents, the available computational and informational resources, as well as the temporal limitations. In case of insufficient quality, preference may be given to the information retrieval approach and the machine translation of a document search profile (DSP), rather than text documents.

The analysis of well-known contemporary Internet search services (for example, Google [27], Yandex [28], Bing [29], Baidu [30], Yahoo [31], Mail.ru [32], AOL [33] etc.) has revealed, that it is sensible to focus on Google search engine, in spite of its peculiarities and imposed restrictions that are the most significant in terms of the problem to be solved. For example, according to [34, 35] (as of April 2017) the global marketing share percentage, in terms of the use of search engines heavily favours Google, with over 77%. It's interesting to note that Google's large market share is still on the increase. Last year the market share for Google was 67%, so Google has taken another 10% of the market from its rivals in just the past 12 months ("Fig. 1").

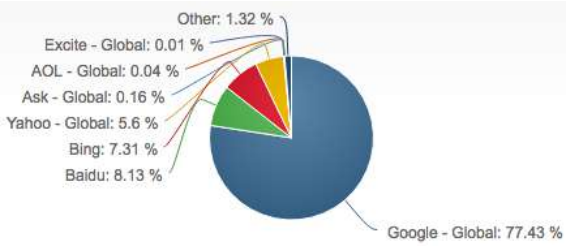

Figure 1. The global marketing share of search engines.

The graph below highlights the usage of search engines in Belarus in past 12 months ("Fig. 2") [36]. During the survey period, it was found that Google accounted for 69 %, Yandex.ru – 26,77 %, Mail.ru – 3,28 %, Bing – 0,35 %, Yahoo! – 0,31 % and Other – 0,29 % of search queries in the country.

Huge volumes of search space and severe time limitations imposed on the response duration of the system of automatic recognition of adopted text fragments require preliminarily the quick and efficient minimization

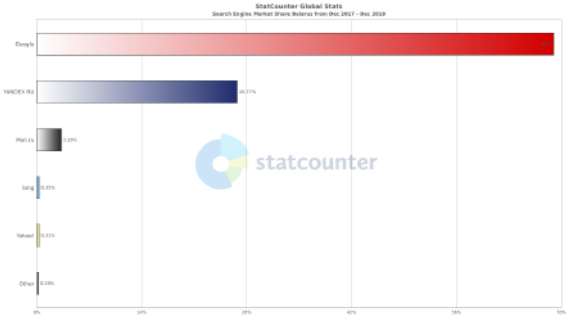

Figure 2. Search engine market share of Belarus in the period of December 2017 – December 2018.

of the search space and then start the recognition of the adopted text fragments. It's necessary to take into consideration that we are talking about multilingual information environment and cross-language functionality, in other words the search space must contain all relevant documents from the search space, regardless of language representation. And this, in turn, requires the use of machine translation functionality either in relation to text documents, or to their search profiles.

Thus, in [37] the method, based on bilingual dictionaries of concepts, actions, attributes that provides an efficient solution for the task of automatic recognition of adopted fragments even at a semantic level, it is proposed to use. The following steps are performed ("Fig. 3"):

- the key words are marked out from the input document and the query search profile is built (QSP);
- the QSP is translated into the remaining (n-1) languages via a multilingual lexical database;
- the search for relevant documents;
- the formation of the minimized search space for the purpose of fast and effective selection of documents belonging to it;
- the search for the adopted fragments in the text documents from the minimized search space.

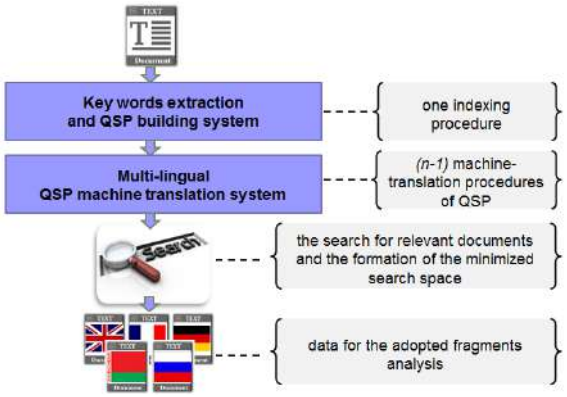

Figure 3. The flowchart of the minimization procedure of the search space.

A well-known TF*IDF method [38] was used to automatically create a query search profile (the input information for the search engine). This method forms the set of key words marked out from the input document automatically, or, in other words, it forms a search profile of a proper text document. The analysis of search queries has revealed a new downward trend of the number of keywords usage. One of the examples is the average number of typed search terms during online search in the United States as of January 2016. During that month, 28.24 percent of all U.S. online search queries contained two keywords ("Fig. 4") [39]. This trend is obviously a consequence of the optimization of the used indexing mechanisms for both documents and user search queries.

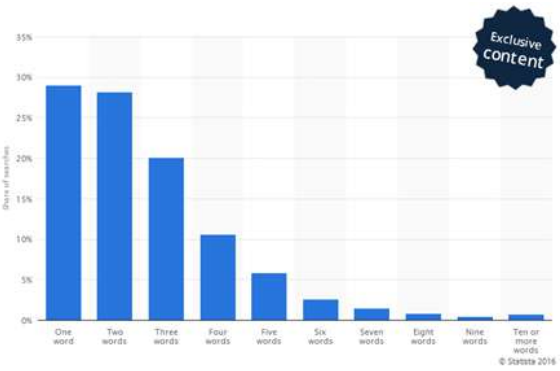

Figure 4. Average number of search terms for online search queries in the United States as of January 2016.

So, the procedure of searching for the relevant documents is a mechanism of the search space minimizing. The list of keywords, achieved by TF*IDF method, can be improved via the usage of synonymy relations, which are defined with the help of checking of lists of synonyms in the MModWN database [40], as well as by the corrections of the weight coefficients, which take in consideration lexical units membership of the most informative lexico-grammatical, syntactical and semantic classes ("Fig. 5").

Synset number	Definition	Synonymic sets		
		English	French	German
109459609	any mechanical force that tends to retard or oppose motion	• RESISTANCE	• RESISTANCE	• WIDERSTAND
109421558	the resistance encountered when one body is moved in contact with another	• FRICTION • RUBBING	• FRICTION • FROTTEMENT	• REIBUNG • FRIKTION
109421888	the process of wearing down or rubbing away by means of friction	• ABRASION • ATTRITION • GRINDING	• ABRASION • BROYAGE • MEULAGE	• ABRIEB • ABNUTZUNG • VERSCHLEI • SCHLEIFEN • SCHLIFF

Figure 5. A fragment of multilingual lexical database MModWN.

For example, semantic super-subordinate relations between the concepts with numbers 109459609 →

109421558 and 109421558 → 109421888 define the relationships between proper synonymic sets in different languages.

Thereby, the machine translation of the query search profile is performed as follows:

- the search for the key words in the multilingual lexical database;
- the selection of the equivalent synonyms, including the synonyms in different languages.

CONCLUSION

An involvement of the well-developed linguistic text analysis that is based on the knowledge of natural language, together with the existed effective information retrieval and machine translation tools provide a facility of qualitative solving the task of automatic recognition of adopted fragments in multilingual information environment with the cross-language functionality and focus on the detection of both explicit and implicit adoptions of text.

REFERENCES

[1] Mamchich, A.A. Modeli i algoritmy informatsionnogo poiska v mnogoyazychnoi srede na osnove tematicheskikh i dinamicheskikh korpusov tekstov, *dis. kand. tekh. nauk*, Minsk, 2011. – 122 p.

[2] Poisk znanii v Internet. Available at: http://poiskbook.kiev.ua/pbs.html. (accessed 2018, Oct).

[3] Voronkov N.V. Metody, algoritmy i modeli sistem avtomaticheskogo referirovanija tekstovyh dokumentov, *dis. kand. teh. nauk*, Minsk, 2007. 165 p.

[4] AntiPlagiat. Available at: http://www.antiplagiat.ru/. (accessed 2017, Dec).

[5] Ephorus. Available at: http://www.ephorus.com/. (accessed 2017, Dec).

[6] WCopyfind. Available at: http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/. (accessed 2018, Nov).

[7] JPlag. Available at: https://jplag.ipd.kit.edu/. (accessed 2019, Jan).

[8] Copyscape Plagiarism Checker - Duplicate Content Detection Software. Available at: http://copyscape.com /. (accessed 2019, Jan).

[9] Krapivin, Yu.B. K zadache avtomaticheskogo raspoznavaniya vosproizvedennykh fragmentov tekstovykh dokumentov [To a task of automatic recognition of reproduced fragments of the textual documents], *Vestnik BrGTU*, 2009, vol. 59, no. 5, pp. 109–112.

[10] PROMT Translator. Available at: http://www.promt.com. (accessed 2018, Oct).

[11] SYSTRAN. Available at : http://www.systransoft.com. (accessed 2018, Oct).

[12] RETRANS. Available at: http://www.retrans.ru/. (accessed 2018, Apr).

[13] Belazar. Available at: http://belazar.info. (accessed 2018, Oct).

[14] Yandex.Translate. Available at: https://translate.yandex.by/help. (accessed 2018, Dec).

[15] Google.Translate. Available at: https://translate.google.com/intl/en/about/. (accessed 2018, Dec).

[16] Brown, P.F. [et al.] A Statistical Approach to Machine Translation, *Computational Linguistics*, 1990, vol. 16, no. 2, pp. 79–85.

[17] Vogel, S. [et al.] Statistical Methods for Machine Translation,*Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag: Berlin, 2000, pp. 377–393.

[18] Beyond Translation Memories. Available at: http://www.eamt.org/events/summitVIII/papers/schaeler.pdf. (accessed 2018, Feb).

[19] Turcato, D., Popowich, F. What is Example-Based Machine Translation?,*Recent Advances of EBMT* / M. Carl, A. Way; ed.: M. Carl [et al.], Kluwer Academic Publishers, 2003, pp. 59–82.

[20] Kaji, H. An Efficient Method for Rule-Based Machine Translation, *Computational Linguistics: proceedings of the 12th Conference, Budapest, 22-27 August 1988*, ACL, 1988, pp. 824–829.

[21] Tamas, G., Gabor, H., Balazs, K. MetaMorpho TM: A Rule-Based Translation Corpus, *Language Resources and Evalutation: proceedings of the 4th International Conference, Lisbon, 2004*, pp. 339–342.

[22] Potthast, M. [et al.] Cross-language plagiarism detection, *Language Resources and Evaluation*, 2011. vol. 45, no. 1, pp. 45–62.

[23] Steinberger, R., Pouliquen. B., Hagman, J. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc, *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Computational Linguistics and Intelligent Text Processing*: Springer Berlin Heidelberg, Berlin-Heidelberg, 2002, pp. 415–424.

[24] Pataki, M. A new approach for searching for translated plagiarism, *Proceedings of the 5th International. Plagiarism Conference, Newcastle-upon-Tyne*, UK, 2012, pp. 49–64.

[25] Potthast, M. Cross-language plagiarism detection, *Language Resources and Evaluation*, 2011, vol. 45, no. 1, pp. 45–62.

[26] Muhr. M. External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010, *Notebook Papers of CLEF 2010 LABs and Workshops*, 2010, vol. 1176, pp. 1–10.

[27] Google. Available at: https://www.google.com/. (accessed 2018, May).

[28] Yandeks. Available at: https://www.yandex.ru/. (accessed 2018, May).

[29] Bing. Available at: https://www.bing.com/. (accessed 2018, May).

[30] Baidu. Available at: https://www.baidu.com/. (accessed 2018, Dec).

[31] Yahoo. Available at: https://www.yahoo.com/. (accessed 2018, Dec).

[32] Mail.ru. Available at: https://www.mail.ru/. (accessed 2018, May).

[33] AOL. Available at: https://www.aol.com/. (accessed 2018, May).

[34] Search Engine Statistics 2017. Available at: https://www.airsassociation.org/airs-articles/item/19297-search-engine-statistics-2017. (accessed 2018, Dec).

[35] NetMarketShare. Available at: https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0. (accessed 2018, Dec).

[36] StatCounter Global Stats. Available at: http://gs.statcounter.com/search-engine-market-share/all/belarus#monthly-201712-201812. (accessed 2019, Jan).

[37] Krapivin, Yu.B. Funktsional'nost' cross-language v zadache avtomaticheskogo raspoznavaniya semanticheski ekvivalentnykh fragmentov tekstovykh dokumentov [Cross-language Functionality in the Problem of the Automatic Identification of the Semantically Equivalent Fragments of the Text Documents], *Iskusstvennyi intellect [Artificial Intelligence]*, 2013, vol. 62, no. 4, pp. 187–194.

[38] Robertson, S. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF, *Journal of Documentation*, 2004, vol. 5, no 60, pp. 503-520.

[39] U.S. online search query size 2017. Available at: http://www.statista.com. (accessed 2017, Dec).

[40] Krapivin, Yu.B. Lingvisticheskii analiz teksta v zadache avtomaticheskogo raspoznavaniya zaimstvovannykh fragmentov tekstovykh dokumentov [The linguistic analysis of text in a problem of automatic recognition of the borrowed fragments of text documents], *Vestnik BrGTU*, 2017, vol. 107, no 5, pp. 54–58.

ИНФОРМАЦИОННЫЙ ПОИСК И МАШИННЫЙ ПЕРЕВОД В РЕШЕНИИ ЗАДАЧИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ЗАИМСТВОВАННЫХ ФРАГМЕНТОВ ТЕКСТОВЫХ ДОКУМЕНТОВ

Крапивин Ю. Б.

В статье предложено решение задачи автоматического распознавания заимствованных фрагментов в многоязычной информационной среде с функциональностью cross-language и ориентацией на обнаружение не только явных, но и неявных заимствований фрагментов текста, на основании применения средств развитого лингвистического анализа текста, опирающихся на знания о естественном языке в сочетании с существующими эффективными инструментами информационного поиска и машинного перевода.

Received 10.01.19