ABSTRACT

OF THE DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD) IN THE SPECIALTY "6D075100 - COMPUTER SCIENCE, COMPUTER ENGINEERING AND MANAGEMENT"

Yerkebulan Gulnur Turataykyzy

«The system of identification of patterns of multilingual texts»

Relevance of the research topic. In the list of languages, English ranks 3rd (379 million people) in terms of the number of native speakers, Russian - 7th (154 million people), and Kazakh - 76th (12.9 million people) [1]. According to the analysis of multilingualism, the number of people who speak three languages fluently is 13% of the world's population, and 42% speak two languages [2]. In the Republic of Kazakhstan, knowledge of three languages is almost a prerequisite for career growth and good pay. The state language of the Republic of Kazakhstan is Kazakh, along with Kazakh, Russian is officially used. They started learning English from the 1st grade in 2014.

Knowledge of several languages opens a window into a large global world with its colossal flow of information and innovations. Multilingualism is the dictate of the time and the state pays great attention to this direction of development.

At the moment, the implementation of the Roadmap for the development of trilingual education for 2015-2020 is underway [3], according to the results of which a schedule for the transition to trilingual education in schools of the Republic of Kazakhstan from 2023 has been prepared [4]. In 2022-2023, it is planned to switch to English-language education in multilingual schools, and in 2023-2024 – the transition of all general secondary educational institutions. The introduction of trilingual education will be implemented by choice on the basis of a collegial conclusion of the pedagogical council of the educational organization and the parents' committee. The transition to teaching in three languages is carried out within the framework of the 79th step of the National Plan "100 steps" [5] and the State Program for the Development of Education and Science of the Republic of Kazakhstan for 2016-2019 [6].

The language industry is a big business. TechNavio analysts in their study "Global Language Services Market 2020-2024" predict a market growth of 9.72 billion US dollars during 2020-2024, with a CAGR of 4% [7].

Today, there are many online translators and browser extensions that help translate an unfamiliar word into a foreign language [8]. Everyone has different translation algorithms, different databases of words and phrases, so the translation results may differ. How do I find the best translation? There is only one browser extension found on the Internet that offers several translation options from different online companies at the same time, its name is MyTranslator. The app offers translations from Google Translate, Microsoft Bing Translator, Yandex Translator and DeepL Translator to see the options you need to switch to the tab "G", "M", "T" and "Y". The extension was launched in January 2020 and has more than 10,000 users [9]. At the same time, to translate the text by sentences, it is necessary to call the extension for each sentence again, there is no way to view all the options from translators without switching tabs, there is no way to somehow fix the selected translation option in order to subsequently correct the translated text yourself (except for the standard method of "Highlight" - "Copy" - "Paste"). The first direction of the dissertation research, "Determining the most reliable translation in working with multilingual texts", does not contain the listed disadvantages. In addition, this direction provides an assessment of each translation option, prompting the user to make the best choice.

Internet resources with automatic text generation [10-12] together with translation sites [13-15] can create many interpretations of the same text. To catch the original, they use the capabilities of neural networks [16-18].

To date, neural networks are widely used in forecasting, classification and management tasks. Their strengths include "solving problems with unknown patterns, resistance to noise in input data, potential ultra-high performance and fault tolerance in the hardware implementation of a neural network" [19]. At the same time, huge computing resources are needed for their training and work [20].

In the second direction of the dissertation research, an alternative method of detecting the translation of a text based on the Renyi entropy is proposed, which does not require significant computer power and a lot of time to search. The Renyi entropy was not chosen by chance as the core of the development. The entropy approach is already being explored and applied when working with texts, but in other contexts. So, in [21] 2019, the results of research on the extraction of concepts for structured text using the entropy weight method are presented, in [22] 2020, the results of research on the entropy relationship between the length of the text and lexical richness are presented, in [23] 2019, the entropy analysis of dubious text sources is investigated on the example of the Voynich manuscript, in [24] 2019, linking entropy estimation with machine translation, methods for solving insufficient translation by two-phase splitting of the process are proposed. More information about these and other works can be found in paragraph 2.3.2 of the dissertation research.

As Internet technologies continue to grow and expand, becoming more accessible and widespread, the value of the site is also growing. A website is a very valuable asset for any company in today's globalized world with its ability to attract new visitors, inform them about products or services, and conduct sales, regardless of the type of business.

Based on the analysis of the answers to the question "Would you make a purchase on a website that has content in your native language if the quality of this content was low?" of a sample group (3,000 people, 50% men / 50% women in 8 countries, on about 3 continents, aged 25 to 65 years and with different levels of English proficiency), it was found that 45.3% would not agree to make a purchase on a website with poor translation of content in their native language [25]. This means that the issue of site localization is becoming very relevant, so new disciplines are emerging in universities dedicated to this topic, with the study of linguistic and marketing features, as well as various technical aspects [26, 27].

Every educational institution is interested in attracting foreign students and investors, so the university's website should be an appropriate representation for international cooperation, as well as the websites of public and private organizations.

In the system of identification of patterns of multilingual texts, another direction is considered, which determines the missing translations of a multilingual site in all its language versions with the possibility of generating machine translation of the missing information.

News aggregators are also multilingual sites, but in the dissertation research they are considered not in the third, but in the fourth direction. This is due to the fact that the algorithm applies to the generation of information during post-parsing and pre-publication, and not to published materials.

The latter direction, dedicated to the creation of tests and training materials for several languages, considers the importance of authenticity of information in different languages. Authenticity of information is a property that guarantees that the translation and the original are identical. There are scientific experiments, the results of which have shown that materials in the target language are assimilated better if the material is authentic to the original [28,29]. It is difficult to find a program that allows you to check the correctness of the translation in automatic mode. Until now, online translators contain errors or lack of translation options for some words and expressions in the Kazakh language, therefore, double-checking by an expert group is sometimes used to verify translations.

In the methodological recommendations [30, 31] on the preparation of test tasks, paragraph 6.5 of the sixth chapter states: "Translation of test tasks – when developing, updating test tasks in the state, Russian and other languages, questions and answers must be authentic and adhere to terminological dictionaries approved by the State Terminological Commission under the Government of the Republic of Kazakhstan." In the dissertation work, it is proposed to consider the possibility of automatic verification and hints in the formation of translations of tests and educational materials based on a series of synonyms and terminological dictionaries, as well as a general analysis of information in general, in order to identify missing or erroneously translated parts of the original text.

As a result, in the dissertation "The system of identification of patterns of multilingual texts" (hereinafter – SIPPT), five main areas of research were identified:

1) determining the most reliable translation in working with multilingual texts among online translators;

2) determining the difference between a real and fake translation of the source text based on the entropy approach;

3) identification of missing translations of a multilingual site in all its language versions;

4) identification of missing translations in the work of multilingual news aggregators;

5) creation of test and training materials for several languages.

All these facts indicate that the identification of the translation and the original text in working with multilingual texts is relevant.

The purpose of the study: to develop a management model and algorithms for identifying patterns of multilingual texts in areas related to multilingual texts using the "sentence" and "paragraph" patterns, as part of the implementation of the 79th step of the National Plan "100 steps".

To achieve the goal, the following research objectives were set:

1. Analysis of plagiarism detection systems and online translation services for use in SIPPT, systematization of data to identify the main directions of SIPPT.

2. Performing calculations to select the most accurate fuzzy string comparison algorithm for comparing text and its translation (from online translators) based on software implementation and expert assessments. Creation of a corpus of parallel texts in Russian, Kazakh, and English for calculations.

3. Development of an algorithm for determining the most reliable translation in working with multilingual texts among online translators and an algorithm for creating tests and training materials for several languages to prevent translation errors using the selected fuzzy string comparison algorithm.

4. Performing calculations to test the efficiency of the entropy approach (formation of key series of high-frequency words of texts, calculation of entropy coordinates for the "sentence" and "paragraph" patterns, calculation of distances between sets of entropy texts in accordance with the Minkowski metric) to determine the proximity of texts in different languages with a software implementation of coordinate calculation.

5. Development of an algorithm for determining the difference between a real and fake translation of the source text, an algorithm for determining missing translations of a multilingual site in all its language versions and an algorithm for determining missing translations in the work of multilingual news aggregators using an entropy approach.

6. Development of a management model of a system for identifying patterns of multilingual texts based on the results of the above studies.

The object of the study is an information field containing essential information blocks of text in various languages.

The subject of the study is a control model and algorithms for identifying patterns of multilingual texts.

Research methods: Oliver's fuzzy string comparison algorithm, FuzzyWuzzy fuzzy string comparison algorithm, Porter stemmers, normalization, Shannon entropy, Renyi entropy, Minkowski metric, Hamming distance, Cartesian distance, distance between centers of mass, distance between geometric centers, distance between centers of parametric averages.

Scientific novelty: the scientific novelty lies in the substantiation of the SIPPT proposed by the author as a result of the synthesis of methods for identifying patterns of text materials, taking into account the peculiarities of multilingualism, parametrizable entropy and well-known php solutions.

Innovations of the results obtained:

- a corpus of parallel texts in Russian, Kazakh, and English has been developed;
- an algorithm has been developed to determine the most reliable translation in working with multilingual texts among online translators and an algorithm for creating tests and training materials for several translation languages using a fuzzy string comparison algorithm;

- an entropic approach to detecting the proximity of multilingual texts has been developed;

- an algorithm has been developed to determine the difference between a real and fake translation of the source text, an algorithm for determining missing translations of a multilingual site in all its language versions and an algorithm for determining missing translations in the work of multilingual news aggregators using an entropy approach;

- The SIPPT management model has been developed as part of the implementation of the 79th step of the National Plan "100 steps".

Theoretical significance: The main theoretical discoveries were the discovery of the ability of the author's entropy approach to determine the proximity of texts in different languages, as well as the experimental identification of a more accurate algorithm for fuzzy string comparison when working with multilingual texts. Recommendations on building a corpus of texts will be useful to researchers in the field of multilingual texts, using online translators. The developed management model and algorithms of the System for identifying patterns of multilingual texts are a description of original solutions to existing problems related to the multilingualism of the population.

Practical significance: The practical significance of the work lies in the applicability of the developed system both in organizations interacting with the information field, taking into account multilingualism, and any interested user, since the software elements of this system are publicly available on the Internet. The developed algorithms with the possibility of using the specified patterns can be applied in terms of identifying the most reliable translation in working with multilingual texts in the interests of translators, analysts and other interested users by placing the necessary scripts in open access on the Internet (with the generation of translations based on translation options of several online translators, usually requiring payment for software use).

In terms of determining the detection of a translation adequate to the source text, the developed algorithms with parametrizable entropy can be proposed:

- analytical companies (search for source article/news, breaking articles/news on part of the borrowing and the author's work, etc.);

- organizations from the field of information security (search for duplicates, search of the primary sources of materials in other languages, which represent a threat to national security, etc.);

- in higher education institutions (search translated plagiarism in student papers), etc.

In the areas of determining the missing translations in the work of multilingual news aggregators and determining the missing translations of a multilingual site in all its language versions, the developed algorithms can be applied primarily for owners of news agencies, as well as for public and private multilingual Internet resources.

In terms of creating tests and training materials in different languages, the proposed algorithms can be applied to a wide range of interested persons dealing with the processing of voluminous text materials.

The provisions of the dissertation submitted for defense (scientific results):

- recommendations on the construction of a corpus of texts as a result of the study of ways to search for patterns of multilingual texts using online translation services;
- an entropic approach to detecting the proximity of multilingual texts;
- control model and algorithms of the Multilingual text pattern Identification System;
- software implementation of calculations for comparing patterns of text and its translation using fuzzy string comparison, as well as software implementation of calculation of entropy coordinates.

Personal contribution of the author it consists in conducting research that substantiates the main provisions put forward for protection, as well as a significant role in the generalization and analysis of the results obtained.

The structure and scope of the dissertation. The dissertation has a classical structure: an introductory part, the main part (three chapters), a conclusion, a list of sources used and appendices. The work includes 65 figures, 14 tables and 113 names of the sources used.

In the introduction the choice of the research topic is justified, the relevance of the five directions of SIPPT is revealed, the purpose of the study is formulated, its tasks are determined, the object and subject of the study are presented, the scientific novelty and practical significance of the work are revealed.

In the first chapter a comparison of plagiarism detection systems in Runet is carried out. A detailed analysis of the work of the module of search for transferable borrowings of the system "Antiplagiat. University". Disclosed are methods for finding patterns of multilingual texts with Google and Yandex. The tasks of the dissertation research were formulated.

In the second chapter models and methods of SIPPT are proposed. The direction of determining the most reliable translation using php code is investigated. The classification of texts with five stages is described in detail. The application of the entropy approach as an assessment of the degree of coherence of texts is considered. The efficiency of the entropy approach for the tasks of dissertation research and the choice of an algorithm for fuzzy string comparison for SIPPT directions are experimentally proved. As a result of the experiments, scripts were created for which two certificates were obtained on entering information into the state register of rights to objects protected by copyright (Appendices D, E of the Dissertation).

The best fuzzy string comparison algorithm identified in the experiments was applied in the first and fifth directions of the SIPPT, and the developed entropy approach was applied in the second, third and fourth directions of the SIPPT.

Chapter Three it is devoted to the control model and algorithms of the five directions of SIPPT using methods and techniques from the second chapter. The third chapter describes in detail the software implementation of calculations for comparing patterns of text and its translation using fuzzy string comparison, as well as the software implementation of calculating entropy coordinates (scripts for which copyrights have been obtained).

In conclusion the results of the research, including the main conclusions based on the results of the dissertation research, are presented.

Approbation of the work. The results of the dissertation research were reported and discussed at scientific conferences:

- The VIIIth International Scientific and Practical Conference "GLOBAL SCIENCE AND INNOVATIONS 2019: CENTRAL ASIA" in Nur-Sultan;

- International conference in Warsaw within the framework of the publication MODERN SCIENTIFIC CHALLENGES AND TRENDS (2019);

- International scientific and Methodological conference "Modern University as a space of digital thinking" in Novosibirsk (2020).

A scientific internship was conducted. Received 2 certificates of entering information into the state register of rights to objects protected by copyright.

Publications and author's certificates. The main results of the research have been published in 8 scientific papers, including 2 articles published in international peer-reviewed scientific journals (Scopus), 3 articles in scientific publications included in the List of scientific publications recommended for publication of the main results of scientific activity, approved by the authorized body, 3 papers - in the proceedings of international scientific conferences. Received 2 copyright certificates.

List of scientific publications:

1. Yerkebulan G.T., Kulikova V.P. Features of the implementation of the search for transferable borrowings in the "Anti-Plagiarism" system: strengths and weaknesses // Mater. International Scientific Conference "Modern scientific challenges and trends" Polish Science Journal. – Warsaw, 2019. – № 9(20). – P. 42-46.

2. Erkebulan G. T., Kulikov V. P. Search engines as detecting borrowings in multilingual texts // Global Science and Innovations 2019: Central Asia. – Nur-Sultan, 2019. – № 2(3). – P. 171-173.

3. Yerkebulan G.T., Kulikova V.P. Comparative analysis of systems for detecting cross-language (translated) plagiarism // Bulletin of KazNITU named after K. Satpayev. - Almaty, 2019. – № 6(136). – Pp. 178-183.

4. Yerkebulan G.T., Kulikova V.P., Kulikov V.P. On the use of Google Custom Search and Google Translate API in detecting cross-language plagiarism // Bulletin of the AUES. The series "Information technologies". - Almaty, 2019. – № 4(47). – Pp. 109-116.

5. Yerkebulan G.T., Kulikova V.P., Kulikov V.P. About the application Яндекс.XML and the Yandex.Translator API in the system of identification of patterns of multilingual texts // Bulletin of the AUES. The series "Information technologies". - Almaty, 2020. – № 1(48).– Pp. 110-117.

6. Yerkebulan G.T., Kulikova V.P., Kulikov V.P., Krylova E.M. Models and methods of classification of text queries in the system of identification of patterns of multilingual texts // Mater. international scientific conf. "Modern University as a space of digital thinking". - Novosibirsk: SGUGiT, 2020. - pp. 130-134.

7. G. Yerkebulan, V. Kulikova, V. Kulikov. Google/Yandex Translation Detection in the Patterns Identifying System of Multilingual Texts // Research Institute for Intelligent Computer Systems, West Ukrainian National University, журнал «International Journal of Computing», ISSN 1727-6209 (print). – 2021. – Vol. 20, Issue 1. – P. 72-77. *// https://doi.org/10.47839/ijc.20.1.2094*

8. G. Yerkebulan, V. Kulikova, V. Kulikov, Z. Kulsharipova. Devising an entropy based approach for identifying patterns in multilingual texts // PC TECHNOLOGY CENTER, Kharkiv, Ukraine, журнал «Eastern-European Journal of Enterprise Technologies», ISSN 1729-3774 (print). – 2021. – №2/2 (110). – P. 16-22. // *https://doi.org/10.15587/1729-4061.2021.228695*

Certificates of entry of information into the State register of rights to objects protected by copyright:

1. Certificate of entry of information into the state register of rights to objects protected by copyright No.19775 dated August 17, 2021 "Program for determining the most reliable translation in working with multilingual texts" (computer program), authors: Yerkebulan G.T., Kulikova V.P., Kulikov V.P.

2. Certificate of entry of information into the state register of rights to copyrighted objects No.19562 dated July 30, 2021 "Program for calculating entropy coordinates (using the "paragraph" and "sentence" patterns) to determine the proximity of multilingual texts" (computer program), authors: Yerkebulan G.T., Kulikova V.P., Kulikov V.P.