ANNOTATION

to the thesis research to seek a PhD (PhD)
majoring in «6D070400 – Computer engineering and software»

Mukhsina Kuralay Zhenisbekovna
«Development of multi-language text information analysis system based on the computerized learning»

Relevance of work. The problem of languages contingence, searching for effective and sustainable programs in the field of languages on society consolidation is actual more than ever in the modern multilingual and multicultural world. Trinity of language project initiated by the Head of the state and executed within the Republic of Kazakhstan is a base for data exchange both domestically and internationally. This project promotes the brisk growth of information community within the country and integration of Kazakhstan into the world global information community by considering Kazakh language as a state language and Russian language as a dominant spoken language and English language as a language of successful integration into the global economics. It shall be deemed to develop as special information applications on processing Kazakh language so the possibilities of its inclusion into the applications of multi-language processing to execute international engagement, product promotion, knowledge and communication of the Republic of Kazakhstan in the global information space.

At present, there is sufficient quantity of applications processing Kazakh language. At the same time, a fair amount of problems concerning the need in improving the quality of automatic processing of Kazakh language is kept.

By now, the following leading scientists made considerable contribution in solving problems of automatic processing of Kazakh language as: Kaldybai Bektaev, Sharipbaev Altynbek Amirovich, Amirgaliev Edilkhan Nesipkhanovich, Tukeev Ualsher Anuarbekovich, Khusein Atakan Varol, Rakhimova Diana Ramazanovna, Musabaev Rustam Rafikovich, Mansurova Madina Esimkhanovna, Musiralieva Shynar Zhenisbekovna and many others.

However, the main part of current investigations is focused on the analysis of morphology automation and syntax when the problem of its semantic analysis remains as not solved. Scientific research analysis in the field of computer linguistics, data mining and artificial intelligence associated with development of knowledge and including Kazakh language into the multi-language projects demonstrates that solutions existing in this prospect is not sufficient to meet the actual requirements on developing the system of automatic processing Kazakh, Russian and English languages in a good manner.

Predominantly, the current multi-language NLP applications use only grammatical phase of language processing, while the semantic text analysis and analysis of natural language content remain as one of the key problems as artificial intelligence theory so computer linguistics.

As far as methods based on the rules require much intelligent expenditures and computer processor units are more powerful, machine learning methods become more common. Though, the previously made grammatically and semantically tagged corpus of natural language are required to use machine learning methods while semantic and grammatical processing of multi-language information.

All the above mentioned stipulates the relevance of thesis work concerned with comprehensive research and solving problems of multi-language text information analysis based on the machine learning.

Objective of thesis work is improving the quality of automatic processing the text information of Kazakh, Russian and English languages using the models of intelligent analysis and machine learning methods. The scientific task of modeling the processes of intelligent processing multi-language texts, development of models, methods, algorithms, performing the analysis of multi-language text information in order to determine the general properties of texts to construct machine learning models is solved within the stated objective. Obtained algorithms shall reach their practical implementation in form of the research systems of text processing with possibility of assessing their work on the basis of the made tagged corpus.

The **tasks** given below are solved to obtain the set objectives of the thesis work.

- Development of fact extracting model from semi structured and unstructured text bases and its adaptation for Kazakh, Russian and English languages;
- Modification of the method of random POS-tagging using the hidden Markov's model;
- Development of method for determining the semantic distance of multi-language text documents based using VSM;
- Formation of method for appraisal of quality of the text semantic distance analysis work;
- Building software application implementing the developed models, methods and algorithms.

Scientific novelty of thesis work.

- The hybrid method of automatic morphologic and semantic tagging of text corpus of Kazakh language was modified, the distinctive feature of which is simultaneous use of HMM and rules represented by the regular expressions; that allowed to remove the part of morphologic polysemy and to increase the entity and accuracy of tagging;
- The logical-linguistic model of semantic analysis was developed identifying the facts in multilingual texts, that allowed to extract knowledge from the texts of the Kazakh, Russian and English languages, explicitly represented as the form of RDF triplets and to form semantically marked learning corpus.
- Method of determining the semantic distance of multilingual text documents based on VSM was improved, which is distinguished by using the impulse response PPMI to determine whether a text belongs to a highly specialized subject area;
- Information technology was made for determining the semantic distance of texts to a given highly specialized topic, based on the proposed method of

calculating the average value of the cosine similarity of vectors of learning corpus documents.

Methods of research are based on the complex use of intelligence theory, general systems theory, systematic analysis, finite predicate algebra and machine learning methods. The finite predicate algebra is used to formalize the semantic information transferred by natural sentences, with the subsequent formation of a learnt corpus. Machine learning methods are used to build models for determining the belonging of texts of a narrow subject area and an algorithm for semantic tagging the multi-language corpus.

Object of the research. Systems of automatic processing of text information in Kazakh, Russian and English languages.

Subject of the research is the models and algorithms for intelligent semantic analysis of multi-language text information.

Practical significance of work consists in the development based on the rules presented for defense of a software application, which allows to execute automatic semantic tagging of multi-language corpus of texts of the Kazakh, Russian and English languages, and an application allowing to determine the possible criminal component of the analyzed text. And also consists in the development of semantically tagged corpus of criminalized texts of the Kazakh, Russian and English languages.

The applied value of the work results consists in the possibility of identifying crime-colored texts in computer networks by any interested state bodies.

The main findings to be defended. The following below tasks were solved based on the research results:

- The model of extracting facts from semi-structured and unstructured text bases which was adapted for the Kazakh, Russian and English languages was developed. The selection of the mathematical tool of the algebra of finite predicates for modeling the semantics of natural language sentences was proved.
- Method of random POS-tagging using the hidden Markov's model was modified.
- Method for determining the semantic distance of multi-language text documents based using VSM was developed;
- Method for appraisal of quality of the text semantic distance analysis work was formed;
- Software complex allowing to determine the presence of a criminal meaning in incoming texts and performs semantic tagging was built.

Connection between the topic and the plans of scientific and research programs

The thesis work was carried out according to the schedule plan of scientific research grant works: "Methods and models of search and analysis of criminally significant information in unstructured and semi-structured text bases" of the Institute of Information and Computational Technologies of the Science Committee and the Ministry of Education and Science of the Republic of Kazakhstan.

Publications. The main results of research on the thesis topic are presented in 17 publications, 4 of them in scientific publications recommended by the KN MoES of the Republic of Kazakhstan, 6 – in the international scientific publications being the part of data base of Scopus and Web of Science, 7 – in the materials of international scientific and practical conferences.

Thesis structure consists of introduction, 4 sections, conclusion, bibliography and five annexes. The total volume of the thesis is 121 pages, 26 figures, 6 applications. The list of references consists of 145 sources.

Introduction contains justification of the relevance of the chosen topic of the thesis work. The objective, object, subject and tasks of research work are formulated. The results of the conducted were described, their scientific novelty and practical significance were shown. The approbation data on the main results of the thesis work is given.

The first section of thesis work contains a review of modern linguistic resources and systems ща automatic processing of Kazakh texts, an analysis of the existing problems of formalization and algorithmization of automatic processing of Kazakh texts is made. Analysis of modern machine learning methods used in processing text information is done in this section and approaches of the Open IE allowing to obtain information from multi-language unstructured texts were highlighted. The research task was set based on the carried out analysis.

The second section of thesis work contains the justification of selecting the mathematical apparatus of the algebra of finite predicates for modeling the processes of intelligent processing multi-language text information. The fundamentals of the algebra toolkit of the predicates and predicate operations are considered concerning its use to formalize the constructions of natural languages that identify relations between participants of the action in a sentence. The developed logical-linguistic model of Open IE, which describes the semantic functions of the sentence participants through the relations of the grammatical and semantic characteristics of the words of sentences of a given language is given in this section. The adaptation of this model for the automatic generation of structured machine-readable information from the texts of the Kazakh, Russian and English languages is demonstrated. The section describes an algorithm for paraphrasing the motivation fact to action in English sentences English obtained on the basis of the developed mathematical model for extracting facts from multi-language text information.

The third section of thesis work is devoted to the development of methods and algorithms for morphological and semantic analysis of multi-language texts, based on machine learning models. The section describes a probabilistic morphological and semantic tagging method using the hidden Markov's model (HMM). The function of evaluating the probability of a tag chain used in the method depends on two probabilities: the conditional probability of the tag sequence and the conditional probability of the token designation by this tag. The algorithm for semantic marking of Kazakh texts is described. The primary tagging of Kazakh language corpus, on the basis of which the training is carried out, is based on the use of a list of suffixes and linguistic rules. The third section also presents a method for determining the semantic proximity of multi-language text documents, based on calculating the cosine similarity between two vectors of VSM documents, which uses the PPMI measure representing the weight function as the coordinates of the vectors.

The fourth section of thesis work contains practical realization of obtained results. The section proves the use of the metric of numerical estimates, using as objectively measured indicators of the effectiveness of the developed models a tuple including the rates of completeness, accuracy and the Van Riesbergen measure. Also, the practical results of the implementing the developed Open IE model on three corpuses of Russian, Kazakh and English texts are also shown. The total volume of built corpuses are 6,000 texts consisting of 700,000 words. The accuracy of extracting a fact triplet for English language is 87.2%, for Russian language is 82.4% and for Kazakh language is71.0%. In addition, the section describes the developed information technology for determining the semantic proximity of texts to a given highly specialized topic, based on the machine learning methods proposed in the thesis work, and provides a model for assessing the quality of this technology.

Conclusion contains the main results and conclusions of this thesis work.

Validity of scientific provisions, conclusions and recommendations submitted to defense is approved by the proper use of the mathematical apparatus; proper design of experiments; qualitative and quantitative correspondence of the results of theoretical studies and experimental data; practical application of research results.

Evaluation of work. The results of thesis work were reported at international scientific conferences, annual scientific conferences of the Institute of Computational and Information Technologies, scientific conferences of young scientists and specialists of the Kazakh National University, as well as scientific seminars of the Department of Informatics of the Kazakh National University named after Al-Farabi.

Certificates of state registration of copyright object rights were received.

Scientific publications:

1. Мамырбаев О.Ж., Мухсина К.Ж. Мәтін үндесітілігін анықтауға арналған қолданыстағы жүйелерді талдау//«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы», 2017. - №5 (315). – Б.149-155.

2. Мамырбаев О.Ж., Мухсина К.Ж. Анализ текстовых сообщений с применением векторной формы//"Международная научно-практическая конференции «Математический методы и информационные технологии макроэкономического анализа и экономической политики», посвященной празднования 80-летнего юбилея академика НАН РК Абдыкаппара Ашимовича Ашимова", Алматы, 11.04.2017-12.04.2017.-С.136-144.

3.Хайрова Н.Ф., Избасаров Е.Ж., Мамырбаев О.Ж., Мухсина К.Ж. Формальная модель оценивания качества экстракции и идентификации знаний из слабоструктурированной текстовой информации// Материалы научной конференции ИИВТ МОН РК «Современные проблемы информатики и Вычислительных технологий». – 2018. - С.306 – 310.

4.Мамырбаев О.Ж., Хайрова Н. Ф., Мухсина К.Ж. Моделирование грамматических способов выражения семантики факта в английском предложении // III Международной научной конференции «Информатика и прикладная математика», посвященная 80-летнему юбилею профессора Бияшева Р.Г.и 70-летию профессора Айдарханова М.Б. 26-29 сентября 2018 года, Алматы. -С.136-144.

5.Petrasova S., Khairova, N., Lewoniewski W., Mamyrbaev O., Mukhsina K. Similar text fragments extraction for identifying common wikipedia communities// MDPI № 66 от 10.12.2018 https://doi.org/10.3390/data3040066.

6.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Automatic extraction of synonymous collocation pairs from a text corpus // Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, 2018, DOI: 10.15439/2018F186 Номер статьи 8511195, -P. 485-488.

7.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Comparative analysis of the informativeness and encyclopedic style of the popular web information sources// Lecture Notes in Business Information Processing 320, 2018, DOI: 10.1007/978-3-319-93931-5_24 -P. 333-344.

8.Mamyrbaev O., Turdalyuly M., Mekebayev N., Mukhsina K., Keylan A., Bagher B., Nabieva G., Duisenbayeva A., Akhmetov B. Continuous Speech Recognition of Kazakh Language // AMCSE 2018 - International Conference on Applied Mathematics, Computational Science and Systems Engineering. Vol. 24 – 2019.

9.Мамырбаев О.Ж., Мухсина К.Ж., Хайрова Н. Ф., Колесник А.С. Лингвистические инструменты выявления криминально окрашенной текстовой информации веб-контента // Қазақстан-Британ техникалық университетінің Хабаршысы – 2018. - №3(46). – Б. 112-117.

10.Хайрова Н. Ф., Мамырбаев О.Ж., Мухсина К.Ж., Колесник А.С. Автоматическая генерация структурированной машинно-читаемой информации из мультиязычных текстов // Информатика и прикладная математика: Матер. IV междунар. науч. конф. – Алматы, 2019. – Ч.2. - С. 509 – 519.

11.Мамырбаев О.Ж., Хайрова Н.Ф., Мухсина К.Ж. Қазақ тіліндегі мәтіндердегі қылмыстық мәнді коллакцияларды анықтау // Вестник КазАТК. – 2019. – № 3 (110). – С. 170-175.

12.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme // 3rd International Conference on Computational Linguistics and Intelligent Systems, 2019, Volume 2362.

13.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Detecting Collocations Similarity via Logical-Linguistic Model // In Proceedings of the Workshop on meaning relations between phrases and sentences - May 23, 2019, Gothenburg, Sweden, pp. 15-22.

14. Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Influence of Various Text Characteristics on the Readability and Content Informativeness // In Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-372-8, DOI: 10.5220/0007755004620469 - pp. 462-469.

15.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Open Information Extraction as Additional Source for Kazakh Ontology Generation // ACIIDS 2020, LNAI 12033, 2020. https://doi.org/10.1007/978-3-030-41964-6_8 - P. 86–96,

16.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. Logical-linguistic model for multilingual Open Information Extraction // Cogent Engineering (2020), https://doi.org/10.1080/23311916.2020.1714829 00: 1714829.

17. Хайрова Н. Ф., Колесник А.С., Мамырбаев О.Ж., Мухсина К.Ж. Выровненный казахско-русский параллельный корпус, ориентированный на криминальную тематику// Вестник Алматинского университета энергетики и связи № 1 (48) 2020- С. 84-92.

Certificates of state registration of rights to a copyright object:

Certificate No. 9180 dated April 8, 2020 on entering information into the State Register of Rights to Objects Protected by Copyright, authors: Mamyrbaev O. Zh., Zhumazhanov B. Zh., Mukhsina K. Zh.