ANNOTATION
dissertation work Shormakova Assem Noyabrevna
on the topic "Development and research of models, methods and tools of post-editing machine translation from English into Kazakh language"presented
for the degree of Doctor of Philosophy (PhD)
in the specialty "6D070300 - Information systems"

The relevance of the research topic is related to the modern development of machine translation and post-editing in the field of information systems. Information systems are used in almost all spheres of modern society. In addition, information technology increases efficiency and productivity in every field and has many benefits, so machine translation is relevant for the growth of education and other areas. Today, the quality of machine translation plays an important role for users, especially in the field of interactive information systems.

Machine translation is one of the leading areas of artificial intelligence in the field of information systems. Machine translation plays an important role in solving the global problem of improving communication between peoples and countries around the world. The quality of machine translation is growing from year to year, but the quality of professional translation has not yet reached.

One of the most important and practical ways to improve the quality of machine translation is the process of post-editing, i.e. correcting machine translation in order to improve the quality of machine translation. Post-editing of machine translation can be done both manually and in automated versions. Manual post-editing of machine translation – laborious process. Automated post-editing of machine translation is one of the current trends in machine translation of natural languages.

In recent years, the number of users of machine translation has been growing rapidly, especially among educational institutions, private enterprises, and translation centers. In addition, the vast majority of foreign companies use machine translation. In addition, many users use machine translation in their daily lives.

Machine translation of the Kazakh language has not yet reached the level of professional translators, so improving the quality of machine translation of the Kazakh language using the post-editing direction is currently a very topical issue.

The scientific contribution of this work is the development of an automatic post-editing technology for the Kazakh language, based on the search for an incorrectly translated word, the formation of a list of words with a similar meaning (catalog) and the selection of the most likely correct word from them using lexical selection technology.

The purpose of the dissertation work. The main goal of this work is to improve the quality of machine translation from English into Kazakh through automatic partial post-editing of machine translation based on lexical selection.

Research objectives. To achieve this goal, 4 tasks were considered:

1 – determine the incorrectly translated word in the Kazakh translated sentence;

2 – automatic formation of a catalog of synonyms for incorrectly translated words;

3 – replacing the incorrectly translated word with a synonym close in meaning, output a machine translation of the corrected sentence.

4 – combine the above three tasks to create a post-editing technology.

Research methods: models and methods of natural language processing.

Object of research: texts of machine translation from English into Kazakh.

Subject of research: automatic post-editing of machine translation from English into Kazakh.

Scientific novelty of the work:

1) The technology of automatic post-editing Post Edit - Lexical Choice (PE-LC) for machine translation of English text into Kazakh was developed for the first time.

2) The method of searching for incorrectly translated words from English into Kazakh has been improved by reverse translation.

3) For the first time, a method has been developed for automatically generating a catalog of synonyms for incorrectly translated Kazakh words.

4) The model and algorithm of the semantic cube method for choosing a synonym with a high probability of a mistranslated word have been adapted.

Theoretical significance of the study. The theoretical significance of the study lies in the development and integration of known text processing methods in post-editing machine translation from English into Kazakh.

The practical significance of the study. The practical significance of the study lies in the creation of a technology for post-editing the text translated from English into Kazakh and in the development of software.

Basic provisions for defense.

1. New technology for automatic post-editing of English-Kazakh machine translation.
2. An improved method for detecting words incorrectly translated from English into Kazakh.
3. Technology of automatic formation of a catalog of synonyms of Kazakh words incorrectly translated from English.
4. The adapted method based on the semantic cube of a selection a high-probability synonym for incorrectly translated words.

The degree of reliability and results of approbation. The reliability of the obtained results is shown by the results of experiments of the developed post-editing technology and approbation of the results by publications in journals and proceedings of international conferences.

The scientific results of the work were presented and discussed at the following international scientific conferences and scientific seminars:

- 9th Asian Conference on Intelligent Information and Database Systems ACIIDS 2017;
- 11th International Conference on Computational Collective Intelligence ICCCI 2019;
- International scientific conference of students and young scientists "World of Farabi", Almaty, 2014, 2015, 2017, 2018.

Also, this topic was discussed at the Department of Information Systems of the Kazakh National University named after al-Farabi and at scientific seminars of the Faculty of Information Technologies.

The connection of the dissertation topic with the plans of scientific research. The dissertation work was carried out in accordance with the plan of the doctoral dissertation and the research plan of the grant funding project "Development of an information and analytical search system for the Kazakh language". (2018-2020, state registration number: No. AP05132950). The results of some studies carried out on this dissertation work are included in the reports of this project for 2018-2020.

The contribution of the doctoral student in the preparation of each publication. Published articles and scientific papers describe the results of research on the topic of the dissertation. During the scientific work, 14 scientific papers were written, including: 1 scientific article in a journal indexed by Scopus:

1. Shormakova A., Zhumanov Z.H., Rakhimova D. "Post-editing of words in Kazakh sentences for information retrieval". *Journal of Theoretical and Applied Information Technology,* 2019, 97(6), p. 1896–1908. (Scopus 2021: Q4, CiteScore-1.3; Percentile- 30%)

4 articles in journals recommended by the Committee for Control in the Sphere of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan:

1.Абеустанова (Шормакова) А.Н. "Машиналық аударманың нарықтағы және Қазақстандағы күйі". ҚазҰТУ хабаршысы № 6(106), 2014. –150-152 б.

2.Абеустанова (Шормакова) А.Н. "Қазақ тіліндегі көпмағыналы сөздердің бірін анықтаудың бір болжамы". ҚазҰТУ хабаршысы №4(110) 2015. –625-628 б.

3.Абеустанова (Шормакова) А.Н. "Ағылшын тілінен қазақ тіліне аударылған қазақша қате сөздерді анықтау және баламалар каталогын құру". ҚазҰТУ хабаршысы №6 2017. –313-317 б.

4. Шормакова А.Н. "Екі табиғи тілдегі аударылған мәтінді туралау". ҚазҰТУ хабаршысы, №4(128), 2018. –344-349 б.

In the collections of international scientific and practical conferences indexed on the basis of Scopus, 2 scientific articles have been published:

1. Abeustanova (Shormakova) A., Tukeyev U. "Automatic Post-editing of Kazakh Sentences Machine Translated from English". *Studies in Computational Intelligence/Advanced Topics in Intelligent Information and Database Systems*, vol. 710 – Springer International Publishing, 2017. – p. 283-295. (Scopus 2021: Q4, CiteScore-1.8; Percentile- 27%).

2. Rakhimova D., Assem S. "Problems of Semantics of Words of the Kazakh Language in the Information Retrieval". *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11684 LNAI, p. 70–81.(Scopus 2021: Q2, SJR=0.25, CS=2.1, Percentile-50%)

6 scientific articles and 1 article in a scientific and technical journal were published in the collections of international scientific conferences:

1.Shormakova A. "Machine translation and post-editing". Материалы международной конференции студентов и молодых ученых «Мир науки», 17-19 апреля 2013г. – Алматы: Қазақ университеті, 2013. – с. 222

2.Шормакова А.Н. "Информатика терминдерінің мемлекеттік тілге аудару ерекшеліктері". Материалы III международного конгресса студентов и молодых ученых «Мир науки», 23-28 апреля, 2009г.-Алматы: Қазақ университеті,- с. 249.

3.Шормакова А.Н., Тукеев У.А. "Технология машинного перевода с обучением английского языка на казахский язык". Материалы международной конференции студентов и молодых ученых «Мир науки», 23-26 апреля 2012г. – Алматы: Қазақ университеті, – с. 154.

4. Sundetova A., Forcada M.L., Shormakova A., Aitkulova A. "Structural transfer rules for English-Kazakh machine translation in the free/open-source platform Apertium", in Proceedings of the I International Conference on Computer Processing of Turkic Languages (TurkLang-2013) (Astana, 3-4 oct. 2013) , p. 322-331.

5. Шормакова А.Н., Айткулова А. "Добавление новой англо-казахской языковой пары в платформу машинного перевода Апертиум". 51-я Международная научная студенческая конференция «Студент и научно-технический прогресс» , Новосибирск, 12-18 апреля 2013, Секция "Информационные технологии".- с. 241.

6.Тукеев У.А., Абеустанова (Шормакова) А.Н., Сундетова А. "Ағылшын-қазақ тілдік жұбы үшін Apertium платформасындағы сөйлемді синтаксистік құрылымдық түрлендіру ережелері және мәселелері".IV международная научно-практическая конференция: (секция «Искусственный интеллект»). Қоғамды ақпараттандыру IV Халықаралық ғылыми-практикалық конференция еңбектері, Астана 2014 ,127-129 б.

7.Шормакова А.Н. Қазақ тіліндегі автоматтандырылған синонимдер тізімін құру [Мәтін] / А.Н. Шормакова, У.А. Тукеев // Механика және технологиялар / Ғылыми журнал. – 2022. – №3(77). – Б.44-49. https://doi.org/ 10.55956/AEQO3045

Structure and scope of research work. The dissertation work consists of an introduction, 6 chapters, a conclusion, a list of references and 2 appendices. The dissertation is a written text of 77 pages, including 12 tables, 7 figures.

The first section provides an overview of machine translation and automatic post-editing. The terms and concepts used in relation to the dissertation are given. New scientific works on post-editing are described. There is a review of scientific works on this topic.

The second section describes the structure and algorithm of the PE-LC post-editing technology. Brief information about the three tasks set in the dissertation is given. The general algorithm of the proposed PE-LC technology is described in the research paper.

In the third section the solution of the first task was comprehensively considered: identifying an incorrectly translated word in a translated sentence. An improved method for detecting words incorrectly translated from English into Kazakh is described.

The fourth section describes the second task, which is to create an automatic catalog (list) of synonyms that are created from incorrectly translated words. Tools and links for creating an automated catalog are presented. Examples of synonyms for incorrectly translated words in the task of automatically creating a catalog are given.

The fifth section describes the third task, the problem of the lexical choice of an incorrectly translated word. An improved model and algorithm of the semantic cube method, the choice of the most appropriate word for a given incorrectly translated word, is presented. Tables and examples are used to create a semantic cube for mistranslated words found.

The sixth section presents the results obtained after experimenting with the proposed PE-LC technology and comparing them with Google Translate. The statistical significance of the experimental data was calculated to determine the improvements in the proposed work. Several tools and metrics were used to compare study results.

In conclusion the main results obtained in the dissertation are formulated.