АННОТАЦИЯ
диссертационной работы Шормаковой Асем Ноябревны
на тему «Разработка и исследования моделей, методов и программных средств пост-редактирования машинного перевода английского языка на казахский язык» представленной
на соискание степени доктора философии (PhD)
по специальности «6D070300 – Информационные системы»

Актуальность темы исследования связана с современным развитием машинного перевода и постредактирования в сфере информационных систем. Информационные системы используются практически во всех сферах современного общества. Кроме того, информационные технологии повышают эффективность и производительность в каждой области и имеют множество преимуществ, поэтому машинный перевод актуален для роста образования и других областей. Сегодня качество машинного перевода играет важную роль для пользователей, особенно в сфере интерактивных информационных систем.

Машинный перевод — одно из ведущих направлений искусственного интеллекта в сфере информационных систем. Машинный перевод играет важную роль в решении глобальной проблемы улучшения коммуникации между народами и странами по всему миру. Качество машинного перевода растет из года в год, но до качества профессионального перевода еще не дошло.

Одним из наиболее важных и практичных способов повышения качества машинного перевода является процесс постредактирования, то есть исправления машинного перевода с целью повышения качества машинного перевода. Постредактирование машинного перевода может производиться как вручную, так и в автоматизированных версиях. Ручное постредактирование машинного перевода – трудоемкий процесс. Автоматизированный постредактирование машинного перевода является одним из актуальных направлений машинного перевода естественных языков.

В последние годы число пользователей машинного перевода стремительно растет, особенно часто им пользуются образовательные учреждения, частные предприятия, центры переводов. Кроме того, подавляющее большинство зарубежных компаний используют машинный перевод. Кроме того, многие пользователи используют машинный перевод в повседневной жизни.

Машинный перевод казахского языка еще не достиг уровня профессиональных переводчиков, поэтому повышение качества машинного перевода казахского языка с использованием направления постредактирования в настоящее время является очень актуальным вопросом.

Научным вкладом данной работы заключается в разработке технологии автоматического постредактирования для казахского языка, основанной на поиске неверно переведенного слова, формирование списка слов с близким значением (каталога) и выборе из них наиболее вероятнее правильного слова с использованием технологии лексического выбора.

Цель диссертационный работы. Основной целью данной работы является повышение качества машинного перевода с английского на казахский язык за счет автоматического частичного пост-редактирования машинного перевода на основе лексического отбора.

Задачи исследования. Для достижения поставленной цели рассматривались 4 задачи:

1 – определить неправильно переведенное слово в казахском переведенном предложении;

2 – автоматическое формирование каталога синонимов неправильно переведенных слов;

3 – неправильно переведенное слово заменив синонимом близким по смыслу вывести машинный перевод исправленного предложения.

4 – объединить три вышеуказанные задачи для создания технологии постредактирования.

Методы исследования: модели и методы обработки естественных языков.

Объект исследования: тексты машинного перевода с английского на казахский язык.

Предмет исследования: автоматическое пост-редактирование машинного перевода с английского на казахский язык.

Научная новизна исследования:

1) Впервые разработана технология автоматического постредактирования Post Edit - Lexical Choice (PE-LC) машинного перевода английского текста на казахский язык.

2) Усовершенствован метод поиска неправильно переведенных слов с английского на казахский путем обратного перевода.

3) Впервые разработан метод автоматического формирования каталога синонимов неправильно переведенных казахских слов.

4) Адаптирована модель и алгоритм метода семантического куба выбора синонима с высокой вероятностью для неправильно переведенного слова.

Теоретическая значимость исследования. Теоретическая значимость исследования заключается в развитии и объединении известных методов обработки текстов в постредактирования машинного перевода с английского на казахский язык.

Практическая значимость исследования. Практическая значимость исследования заключается в создании технологии постредактирования текста переведенного с английского на казахский язык и в разработке программных средств.

Основные положения, выносимые на защиту.

1. Новая технология автоматического постредактирования англо-казахского машинного перевода.

2. Усовершенствованный метод выявления слов неправильно переведенных с английского на казахский язык.

3.Технология автоматического формирования каталога синонимов казахских слов неправильно переведенных с английского языка.

4. Адаптированный метод заключающийся в подборе высоковероятностного синонима на основе семантического куба.

Уровень достоверности и результаты апробации. Достоверность полученных результатов показывается результатами экспериментов разработанной технологии пост-редактирования и апробациями результатов публикациями в журналах и трудах международных конференций.

Научные результаты работы были представлены и обсуждены на следующих международных научных конференциях и научных семинарах:

- 9-я Азиатская конференция по интеллектуальной информации и системам баз данных ACIIDS 2017;

- 11-я Международная конференция по вычислительному коллективному разуму ICCCI 2019;
- Международная научная конференция студентов и молодых ученых «Мир Фараби», Алматы, 2014, 2015, 2017, 2018.

Также эта тема обсуждалась на кафедре информационных систем Казахского национального университета имени аль-Фараби и на научных семинарах факультета информационных технологий.

Связь темы диссертации с планами научно-исследовательских работ. Диссертационная работа выполнена в соответствии с планом докторской диссертации и планом НИР проекта грантового финансирования «Разработка информационно-аналитической поисковой системы для казахского языка ». (2018-2020, государственный регистрационный номер: №AP05132950). Результаты проведенных некоторых исследований по данной диссертационной работе включены в отчеты данного проекта за 2018-2020 годы.

Вклад докторанта в подготовке каждой публикации. В опубликованных статьях и научных трудах описаны результаты исследования по теме диссертации. За время научной работы было написано 14 научных работ, в том числе: 1 научная статья в журнале, индексируемом 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 статьи в журналах, рекомендованных Комитетом по Контролю в Сфере Образования и Науки Министерства образования и науки Республики Казахстан:

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.

В сборниках международных научно-практических конференций, индексируемых на базе Scopus, опубликовано 2 научных статьи:

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%)

В сборниках международных научных конференций опубликовано 6 научных статей и 1 статья в научно-техническом журнале:

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

Структура и объем исследовательской работы. Диссертационная работа состоит из введения, 6 глав, заключения, списка использованной литературы и 2 приложений. Диссертация составляет письменный текст объемом 77 страниц, в том числе 12 таблиц, 7 рисунков.

Во введении обосновывается актуальность диссертации. Сформулированы цель работы, объект и предмет исследования. Выявлены научная новизна и практическая значимость. Описаны результаты исследования. Приводится информация об апробации результатов исследования и публикации.

В первом разделе представлен обзор машинного перевода и автоматического постредактирования. Приведены термины и понятия, используемые применительно к диссертации. Описаны новые научные работы по постредактированию. Имеется обзор научных работ по данной теме.

Во втором разделе описывается структура и алгоритм технологии постредактирования PE-LC. Дана краткая информация о трех задачах, поставленных в диссертации. Общий алгоритм предлагаемой технологии PE-LC описан в исследовательской работе.

В третьем разделе всесторонне рассматривалось решение первой задачи: выявление неправильно переведенного слова в переведенном предложении. Описан усовершенствованный метод выявления слов, неправильно переведенных с английского на казахский язык.

В четвертом разделе описывается вторая задача, заключающаяся в создании автоматического каталога (списка) синонимов, которые создаются из неправильно переведенных слов. Представлены инструменты и ссылки для создания автоматизированного каталога. Приведены примеры синонимов к неправильно переведенным словам в задаче на автоматическое создание каталога.

В пятом разделе описывается третья задача, проблема лексического выбора неправильно переведенного слова. Представлена усовершенствованные модель и алгоритм метода семантического куба, выбора наиболее подходящего слова для данного неправильно переведенного слова. Таблицы и примеры используются для создания семантического куба для найденных неправильно переведенных слов.

В шестом разделе представлены результаты, полученные после экспериментов с предложенной технологией PE-LC и сравнения их с Google Translate. Статистическая значимость экспериментальных данных была рассчитана для определения улучшений в предлагаемой работе. Для сравнения результатов исследования было использовано несколько инструментов и метриков.

В заключении сформулированы основные полученные результаты в диссертации.