АННОТАЦИЯ

диссертационной работы Кэрібаевой Айданы Сейілғазықызы на тему: «Разработка и исследование моделей и методов морфологической сегментации текстов казахского языка для нейронного машинного перевода», представленной на соискание степени доктора философии (PhD) по специальности 6D070300 – Информационные системы

Актуальность темы исследования. Множество проблем машинного перевода до конца не изучены и требуют детального рассмотрения в зависимости от специфики языка. Системы машинного перевода не всегда могут решить проблему традиционными методами. В машинном переводе основанный на правилах могут быть не учтены все правила, в статистическом не всегда определяется правильный перевод по контексту. На сегодняшний день использование нейронных сетей стало популярным во многих предметных областях, машинный перевод также не стал исключением.

Существуют разные подходы к решению задач машинного перевода, например, подход, основанный на грамматических правилах языков; метод статистического машинного перевода, основанный на статистическом методе нахождения таблицы вероятностных фраз переводимых языков; нейронный метод машинного перевода, основанный на изучении нейронных сетей целевых языков. Каждый из этих подходов имеет преимущества и недостатки. В последнее время наилучшие результаты в машинном переводе показывает нейронный машинный перевод на основе нейронных сетей. Поскольку проблема машинного перевода еще не решена на достаточно высоком уровне, близком к профессиональному переводу, проблема машинного перевода является весьма актуальной. Следует отметить, что решение проблемы машинного перевода открывает путь к решению других очень важных задач искусственного интеллекта, таких как понимание естественного языка.

В основе нейронного машинного перевода механизм рекуррентных нейронных сетей, построенный на матричных вычислениях, который позволяет строить существенно более сложные вероятностные модели, чем статистические машинные переводчики. Направление нейронного машинного перевода является актуальной темой для обработки естественных языков, так как нейронный машинный перевод превосходит результаты машинного перевода на основе правил и статистического.

В задачах обработки естественного языка имеется ряд актуальных задач. Одной из самых широко распространенных и применяемых задач для совершенствования качества перевода является: сегментация. В большинстве методов предлагается сегментация на основе частоты, не учитывающая морфологические особенности языка. К такому методу относится метод BPE (byte-pair encoding). Сегментация на основе BPE не дают хороших результатов для агглютинативных языков. В работе предлагается метод сегментирования на основе морфологических особенностей казахского языка на основе CSE (complete set of endings)-модели.

Словари играют важную роль в нейронном машинном переводе. Однако большой словарь требует много памяти, что ограничивает удобство использования NMT и может вызвать ошибки памяти. Это ограничение можно снять, разделив каждое слово на морфемы в параллельных корпусах. Во время обучения NMT размер соответствующего словаря NMT быстро увеличивается; поэтому требует большого объема компьютерной памяти. Поэтому в данной диссертации предлагается новый подход к морфологической сегментации тюркских языков, основанный на Complete Set of Endings (CSE), который уменьшает словарный запас исходных корпусов и, следовательно, объем требуемой памяти.

В машинном переводе существует несколько пред обработочных этапов. Сегментация текста является одним из подготовительного этапа для машинного перевода для сокращения объема словаря. Проблема сегментирования исследованы для аналитических языков во многом, в то время для агглютинативных языков, а именно для тюркоязычных количество исследовании малы. Нейронные сети обычно создают большой словарь, для перевода большинство слов на целевом языке. В нейронном машинном переводе задача сегментации появляется при обучении нейронных сетей для уменьшения объёма словаря, когда объем словаря требуют большего объема памяти, а также для решения задач неизвестных и редких слов. Разбиение на сегменты для данного текста является одним из такого решения. Поэтому **актуальность** задачи сегментирования текста в нейронном машинном перевода возрастает.

Цель диссертационной работы. Разработка моделей, алгоритмов и программных средств для совершенствование нейронного машинного перевода казахского языка на основе лингвистических особенностей.

Задачи исследования. Для достижения целей исследования решаются следующие вопросы:

1. Совершенствование (расширение списка возможных окончаний казахского языка) языковой модели морфологии казахского языка на основе полной системы окончаний (CSE – complete set of endings);
2. Разработка модели и алгоритма морфологической сегментации на основе CSE- модели морфологии казахского языка;
3. Разработка и проведение экспериментов по указанным задачам на платформе нейронного машинного перевода.

Объект исследования. казахский язык.

Предмет исследования. Нейронный машинный перевод казахского языка

Методы исследования. В качестве методов исследования использовались численные методы комбинаторного анализа, методы машинного обучения, глубокого обучения и нейронных сетей.

Научная новизна полученных результатов:

1) Разработана усовершенствованная модель морфологии казахского языка на основе рассмотрения возможных окончаний, отличающихся созданием полного набора языковых окончаний.
2) Разработан метод и алгоритм морфологической сегментации на основе усовершенствованной вычислительной модели морфологии казахского языка, который отличается от популярных методов и алгоритмов созданием полного набора связей в виде таблицы решений и позволяет уменьшить размер словаря нейрона машинного перевода.

Теоретическое и практическое значение работы. Теоретическая значимость данной работы заключается в создании универсального нового метода морфологической сегментации с учетом лингвистических особенностей казахского языка. Метод морфологической сегментации на основе созданной CSE-модели может быть применен и к другим тюркским языкам.

Практическая значимость работы заключается в том, что обучение нейронному машинному переводу на основе сегментированного текста снижает объем памяти и позволяет избежать ошибок с памятью.

Основное положение, выносимое на защиту. Новая модель и алгоритм морфологической сегментации слов казахского языка. результаты экспериментов нейронного машинного перевода казахского языка, подтверждающих эффективность предложенной модели и алгоритма сегментации слов казахского языка.

Степень доверия и результаты апробации. Достоверность и обоснованность результатов исследования подкрепляются обоснованной ответственностью постановки задач, экспертизой критериев и состояния исследований в данной области, а также улучшением результатов в нейронный машинном переводе казахского языка. Результаты диссертации были опубликованы в следующих изданиях.

Журнальная статья в базе данных Scopus:

1) Tukeyev U., Karibayeva A., Zhumanov Zh. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 2020, 1 том, номер №1. (Scopus:Q2; CiteScore-2.5; Percentile- 73%))

2) Turgangayeva A., Rakhimova D., Karyukin V., Karibayeva A, Turarbek A. Semantic Connections in the Complex Sentences for Post-Editing Machine Translation in the Kazakh Language. *Information* 2022, *13*(9), 411; https://doi.org/10.3390/info13090411(Scopus: Q2; CiteScore 4.2; Percentile-64%)

3) Rakhimova D., Karibayeva A. Aligning and extending technologies of parallel corpora for the Kazakh languge. *Eastern-European Journal of Enterprise Technologies*, 2022, 4(2-118), стр. 32–39 (Scopus: Q3; Citescore: 2.0; Percentile: 37%)

В журналах, рекомендованных Комитетом по Контролю в Сфере Образования и Науки Министерства образования и науки:

1) Karibayeva A., Rakhimova D., Abduali B., Amirova. Анализ машинного перевода казахского языка. Вестник КазНИТУ №3 (127), КазНИТУ имени К. И. Сатпаева, 2018, 90 - 96 б. (CCES)

2) Рахимова Д., Тұрарбек А., Карюкин В., Карибаева А., Тұрғанбаева Ә. Қазақ тіліне арнлаған заманауи машиналық аударма технологияларына шолу. Вестник КазНИТУ, №5 (141) 2020. -стр. 103-110.

3) Абдуали Б.А., Әмірова Д.Т., Рахимова Д.Р., Кәрібаева А.С. Қазақ тіліндегі мәтінді ресурстар мен құжаттарды аналитикалық өңдеу. Вестник КазНИТУ, №2(132), 2019, стр. 356-362. (CCES)

4) Karibayeva A., Karyukin V., A. Turgynbayeva, A. Turarbek. The translation quality problems of machine translation systems for the Kazakh language. Journal of Mathematics, Mechanics and Computer Science, [S.l.], v. 111, n. 3, p. 132-140, oct. 2021. ISSN 2617-4871. (CCES)

В конференциях базы Web Science и Scopus:

1) Tukeyev U., Amirova D., Karibayeva A., Sundetova A., Abduali B. Combined technology of lexical selection in rule-based machine translation. Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017, Proceedings, Part II (Lecture Notes in Computer Science) 1st ed. 2017 Edition, Springer, p. 491-500 (**Q3, SJR=0.25, CS=1.8, Percentile-50%).**

2) Tukeyev U., Karibayeva A., Abduali B. Neural machine translation system based on synthetic corpora. CMES-2018, Poland, Kazimeirz Dolny, 2018, MATEC Web of Conferences. 252. 03006. 10.1051/matecconf/201925203006 (**Web of Science**).

3) Tukeyev U., Turganbayeva A., Abduali B., Rakhimova D., Amirova D., Karibayeva A. Lexicon-free stemming for Kazakh language information retrieval. DOI:10.1109/ICAICT.2018.8747021.AICT-2018, Kazakhstan, Almaty (**Scopus)**.

4) Tukeyev U., Karibayeva A. Inferring the Complete Set of Kazakh Endings as a Language Resource. Proceedings of International Conference on Computational Collective Intelligence, 2020, p. 741-751 **(Q4, SJR=0.209, CS=0.9, Percentile – 16%)**

5) Tukeyev U., Karibayeva A., Turganbayeva A., Amirova D. Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words. Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, vol 12876. Springer,

Cham. https://doi.org/10.1007/978-3-030-88081-1_48 -p. 643–654 (**Q3, SJR=0.25, CS=1.8, Percentile-50%**).

6) Rakhimova D., Karyukin V., Karibayeva A., Turarbek A., Turganbayeva A. The Development of the Light Post-editing Module for English-Kazakh Translation. DOI: https://doi.org/10.1145/3492547.3492651. ICEMIS'21: The 7th International Conference on Engineering & MIS 2021, Almaty, Kazakhstan, October 2021 (**Scopus**)

В международных конференциях:

1. Tukeyev U., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules generation, based on the complete set of Kazakh endings // Информатика және қолданбалы информатика: Халықаралық ғылыми конференция материалдары (27-30 қыркүйек 2017 ж). 2-бөлім. - Алматы, 2017, - 38 б.
2. Tukeyev U., Zhumanov Zh., Karibayeva A., Amirova D., Sundetova A., Abduali B. Формирование двуязычного словаря многозначных слов для машинного перевода казахского языка" The Vth International Conference on Computer Processing of Turkic Languages "TurkLang 2017", 18-21 October, Kazan, Tatarstan.
3. Tukeyev U., Zhumanov Zh., Sundetova A., Abduali B., Karibayeva A., Amirova D. Technology of the structural machine translation rules genearation, based on the complete set of Kazakh endings. The II International Conference "Computer Science and Applied Mathematics", 2017, Part II, Almaty, Kazakhstan.
4. Tukeyev U., Zhumanov Zh., Rakhimova D., Karibayvea A, Amirova D. Complex technology of machine translation resources extension for the Kazakh language. Varia Informatica 2017 №1, ISBN 978-83-936692-3-3, Lublin, 14 стр
5. Karibayeva A., Abduali B., Amirova D. Formation of the synthetic corpora for Kazakh on the base of endings complete system. Turklang-2018, Uzbekistan, Tashkent, pp. 153 – 161.
6. Кәрібаева А.С, Абдуали Б.А, Тукеев У. А. Разработка программы морфологической сегментации текста казахского языка на основе полной системы окончаний. «Фараби әдемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Казахстан, Алматы, 2020, - 53 стр.
7. Әмірова Д.Т., Кәрібаева А.С. Исследование технологии машинного перевода казахско-английской пары языков и обратно на основе трансферной модели нейронной сети. «Фараби әлемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференция", Қазақстан, Алматы, 2020, -45 стр.
8. Рахимова Д.Р., Турарбек А., Карибаева А., Карюкин В. Технологий машинного перевода и постредактирования казахского языка. Глава в коллективной монографии «Современные методы и подходы обработки казахского языка» КГТУ, Бишкек 2021.

Личный вклад соискателя. Соискатель решила задачи диссертационной работы. Разработана модель и метод морфологической сегментации текста в нейромашинном переводе казахского языка. Для обучения и тестирования в системе нейронного машинного перевода был составлен корпус параллельных текстов на казахском языке. Были проведены эксперименты по определению эффективности разработанной модели и метода. Для казахского языка был создан полный список окончаний на основе модели CSE (полный набор окончаний).

Связь темы диссертации с планами научно-исследовательской работы. Исследовательские работы по диссертации были выполнены в рамках проекта грантового финансирования «Разработка и исследование системы нейронного машинного перевода казахского языка» (2017-2020, №AP05131415) в Научно-исследовательском Институте Математики и механики.

Объем и структура работы. Диссертация состоит из введения, 4 глав и заключения. Общий объем диссертации составляет 172 страниц, 7 рисунков, 54 таблиц. Список литературы состоит из 79 наименований.

Во введении была определена актуальность работы и показаны проблемы, связанные с темой. Показаны идея работы, цель и задачи исследования, научная новизна и практическая ценность исследования, методы исследования.

Первая глава описывает исследования и анализ существующих технологии для совершенствования нейронного машинного перевода. Приводится аналитический обзор по исследованиям в области сегментации и языковой модели морфологии в машинном переводе. Определяются основные преимущества и недостатки методов.

Во второй главе анализируются модели, используемые при описании языковой морфологии. С учетом морфологических моделей описана работа по созданию языковой модели морфологии казахского языка на основе целостной системы союзов.

В третьей главе проведена работа по созданию модели и алгоритма морфологической сегментации с использованием модели полной системы оконаний (CSE). Разработан пошаговый алгоритм морфологической сегментации. Был создан словарь исключений для устранения ошибочного сегментирования основы слова, который решает многозначность основы слова.

В четвертой главе выбрано программное обеспечение для обучения системы нейронного машинного перевода казахского языка, описана программа сегментации текстов на казахском языке. Описаны основные нейросетевые модели, используемые при обучении нейронных систем машинного перевода. Процесс обучения был реализован с помощью модели seq2seq на основе рекуррентных нейронных сетей в библиотеке Tensorflow. Описаны эксперименты с построенным методом (CSE) и другим методом (BPE). Анализируются результаты эксперимента. Для определения качества нейронного машинного перевода были проведены экспериментальные работы с сравнением метода сегментирования, а именно с BPE, были получены результаты качества в метрике BLEU.

В заключение приводятся основные результаты и выводы диссертации.

Полученные научные результаты подтверждены экспериментами с разными конфигурациями обучения. Обоснованность и достоверность исследования соответствуют полученными результатами разработанного метода и имеющихся в машинном переводе.