

Тукеев У.А., Рахимова Д.Р., Жуманов Ж.М., Сундетова А.М.

МАШИННЫЙ ПЕРЕВОД КАЗАХСКОГО ЯЗЫКА НА АНГЛИЙСКИЙ И
РУССКИЙ ЯЗЫКИ (И ОБРАТНО) НА БАЗЕ ПЛАТФОРМЫ APERTIUM

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	5
1 ОБЗОР ПОДХОДОВ МАШИННОГО ПЕРЕВОДА.....	
1.1 Общая характеристика подходов к решению задачи машинного перевода	
1.2 Подход, основанный на правилах (Rule-based approach).....	
1.3 Статистический подход (Statistical approach).....	
1.4 Подход на основе примеров (Example-based approach).....	
1.5 Подход «память переводов» (Translation memory approach).....	
1.6 Гибридный подход (Hybrid approach).....	
1.7 Обзор программных продуктов компьютерного перевода.....	
2 ОСОБЕННОСТИ ЛЕКСИКИ И СИНТАКСИСА КАЗАХСКОГО ЯЗЫКА.....	42

2.1 Особенности лексики казахского языка.....	
2.2 Особенности синтаксиса казахского языка.....	
3 ПОСТРОЕНИЕ ЛИНГВИСТИЧЕСКИХ ДАННЫХ (СЛОВАРЕЙ, ПРАВИЛ) МАШИННОГО ПЕРЕВОДА НА БАЗЕ ПЛАТФОРМЫ APERTIUM	51
<u>3.1 Описание форматов лингвистических данных платформы Apertium и работы с ними</u>	
<u>3.2 Разработка программного обеспечения автоматизированного пополнения словаря системы машинного перевода</u>	
<u>3.2.1 Методология автоматизированного пополнения словаря системы машинного перевода</u>	
<u>3.2.2 Описание проблемы</u>	
<u>3.2.3 Подход решения проблемы адаптации к казахскому языку</u>	
<u>3.2.4 Применение разработанной методологии автоматизированного пополнения словаря системы к русскому и казахскому языкам</u>	
3.3 Набор регрессионных тестов контроля качества правил для коротких сегментов предложений исходного языка в англо-казахской паре	
<u>Выводы по разделу</u>	
4 РАЗРАБОТКА ОДНОЯЗЫЧНЫХ И ДВУЯЗЫЧНЫХ КОРПУСОВ КАЗАХСКО-АНГЛИЙСКОЙ И КАЗАХСКО-РУССКОЙ ПАР ЯЗЫКОВ	102
<u>4.1 Разработка программного обеспечения поиска и формирования одноязычных лингвистических корпусов</u>	
<u>4.1.1 Модификация открытого программного обеспечения Spiderling поиска и формирования одноязычного корпуса</u>	
<u>4.1.2 Использование одноязычных лингвистических корпусов</u>	
<u>4.2 Разработка программного обеспечения поиска и формирования двуязычных лингвистических корпусов</u>	
<u>4.2.1 Модификация открытого программного обеспечения Vitextor поиска и формирования двуязычного корпуса</u>	
<u>4.2.2 Использование двуязычных лингвистических корпусов</u>	
<u>4.3 Результаты по разработке двуязычных параллельных корпусов англо-казахской пары</u>	
<u>4.4 Результаты по разработке двуязычных параллельных корпусов русско-казахской пары</u>	
<u>Выводы по разделу</u>	
5 РАЗРАБОТКА МОДЕЛИ, АЛГОРИТМА И ПРОГРАММЫ ЛЕКСИЧЕСКОГО ВЫБОРА	123
<u>5.1 Описание проблемы лексического выбора</u>	
<u>5.2 Решение проблемы лексического выбора методом продукционных правил аппарата грамматики ограничений</u>	
<u>5.3 Разработка модели и метода решения проблемы лексического выбора, основанного на статистическом подходе</u>	
<u>5.3.1 Разработка модели решения проблемы лексического выбора</u>	
<u>5.3.2 Разработка метода решения проблемы лексического выбора</u>	
<u>5.4 Результаты решения проблемы лексического выбора</u>	
<u>Выводы по разделу</u>	
6 РАЗРАБОТКА МОДЕЛЕЙ, АЛГОРИТМОВ И ПРОГРАММ СТРУКТУРНЫХ ПРЕОБРАЗОВАНИЙ ПРЕДЛОЖЕНИЙ	150
<u>6.1 Описание проблемы структурных правил преобразования предложений</u>	

6.2 Разработка моделей и методов автоматической генерации структурных правил преобразования предложений для машинного перевода
6.3 Результаты использования технологии автоматической генерации структурных правил преобразования предложений для англо-казахской пары
6.4 Результаты использования технологии автоматической генерации структурных правил преобразования предложений по казахско-русской паре
	Error! Bookmark not defined.
6.5 Исследование проблемы структурных «морфологических чанк» правил преобразования предложений англо-казахской и казахско-русской пар языков	
Выводы по разделу	

7 РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ РЕШЕНИЯ ПРОБЛЕМЫ ОПРЕДЕЛЕНИЯ ИМЕН СОБСТВЕННЫХ ДЛЯ КАЗАХСКО-АНГЛИЙСКОЙ, КАЗАХСКО-РУССКОЙ ПАРЫ ЯЗЫКОВ..... 211

7.1 Классификация имен собственных.....
7.2 Паттерны распознавания для различных классов имен собственных
7.3 Модель, метод и программа максимальной энтропии определения имен собственных на основе параллельных корпусов
7.4 Решение проблемы распознавания имен собственных для казахско-русской языковой пары на основе правил.
Выводы по разделу

8 РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ ПРЕОБРАЗОВАНИЯ ТЕМПОРАЛЬНЫХ (ВРЕМЕННЫХ) ВЫРАЖЕНИЙ КАЗАХСКОГО ЯЗЫКА НА АНГЛИЙСКИЙ ЯЗЫК (И НАОБОРОТ), КАЗАХСКОГО ЯЗЫКА НА РУССКИЙ ЯЗЫК (И НАОБОРОТ). 227

8.1 Описание темпоральных выражений казахского и английского языков с использованием нотации Рейхенбаха
8.2 Модель, метод и программа для определения темпоральных выражений
8.3 Разработка моделей и методов преобразования темпоральных (временных) выражений казахского языка на русский язык (и наоборот)
Выводы по разделу

9 РАЗВЕРТЫВАНИЕ И ОЦЕНКА СИСТЕМЫ МАШИННОГО ПЕРЕВОДА КАЗАХСКО-АНГЛИЙСКОЙ, КАЗАХСКО-РУССКОЙ ПАРЫ ЯЗЫКОВ В СЦЕНАРИЯХ АССИМИЛЯЦИИ (УСВОЕНИЕ), ИНТЕРАКТИВНОГО ПРОГНОЗНОГО ПЕРЕВОДА (INTERACTIVE TRANSLATION PREDICTION), КОРРЕКЦИИ НЕЧЕТКОГО СООТВЕТСТВИЯ (FUZZY-MATCH REPAIR)..... 265

9.1 Развертывание и оценка системы машинного перевода казахско-английской и казахско-русской пар языков в сценарии ассимиляции
9.2 Развертывание и оценка системы машинного перевода казахско-английской и казахско-русской пар языков в сценарии интерактивного прогнозного перевода
9.3 Развертывание и оценка системы машинного перевода казахско-английской и казахско-русской пар языков в сценарии коррекции нечеткого соответствия
Выводы по разделу.....
ЗАКЛЮЧЕНИЕ.....

ВВЕДЕНИЕ

Активная интеграция Казахстана в мировое сообщество и увеличивающимся объемом информационных потоков между нашей страной и ее зарубежными партнерами, реальная потребность для различных слоев населения в оперативном компьютерном переводе при работе в Интернете определяют актуальность вопросов машинного (компьютерного) перевода казахского языка на различные ведущие мировые языки, такие как английский, русский, французский, немецкий, и последнее время, китайский языки, а также и обратного машинного перевода. Первоочередными задачами для информационного взаимодействия населения Казахстана с зарубежными партнерами и внутри страны определены взаимодействия по трем языкам: казахскому, английскому и русскому. В связи с этим является весьма актуальным является высокоэффективная инструментальная поддержка машинным переводом такого трехязычного языкового взаимодействия. В связи с чем актуальным являются исследования и разработка систем машинного перевода промышленного качества с казахского языка на английский и русский языки и обратно.

Анализ состояния исследований в области машинного перевода с казахского языка на английский, русский языки и обратно показывает, что исследования в данной области практически отсутствуют, несмотря на наличие двух-трех коммерческих программных продуктов, качество перевода которых является не достаточно высоким.

Область машинного перевода в мировой науке является достаточно зрелой, разработаны многие формальные модели, алгоритмы, существуют достаточно высокого уровня системы машинного перевода. Однако, большинство результатов машинного перевода получены и применены к ведущим мировым языкам, таким как английский, французский, немецкий, русский, китайский. В приложении к казахскому языку разработанных моделей и алгоритмов, или применений уже разработанных моделей и алгоритмов весьма мало.

Анализ реализации машинного перевода естественных языков показывает всю сложность проблемы машинного перевода естественных языков, как одну из главных проблем создания систем искусственного интеллекта. Выявилось, что основной проблемой машинного перевода естественных языков является недостаточное качество перевода. На это влияют множество вопросов таких как неоднозначность слов в естественных языках, различие грамматических структурах предложений в различных языках, вопросы определения имен собственных, вопросы различного представления темпоральных выражений в различных языках.

Кроме этих языковых проблем, при реализации систем машинного перевода возникают и проблемы реализации программного обеспечения машинного перевода промышленного качества, что требует разработки или выбора соответствующей платформы, использующей уже принятые стандарты хранения и обмена данными в области машинного перевода.

Для решения данной проблемы авторы остановили свой выбор на свободной/открытой платформе машинного перевода Apertium. Выбор свободной(бесплатной) открытой платформы машинного перевода был обусловлен тем, использование коммерческих систем требуют значительных первоначальных вложений. И кроме этого, оказалось, что среди свободных открытых платформ машинного перевода

система Apertium является наиболее известной платформой, основанной на правилах, и которая активно расширяется, исследуя и внедряя статистические методы машинного перевода. Разработка собственной платформы машинного перевода является, с нашей точки зрения, нецелесообразным по соображениям существенных дополнительных значительных финансовых и временных затрат.

В данной работе авторы описывают опыт реализации системы машинного перевода казахского языка на английский и русский языки (и обратно), включая модели, алгоритмы и программное обеспечение.

Благодарности. Авторы выражают благодарность МОН РК за финансовую поддержку данного исследования (проект 0749/ГФ4), а также всех участников проекта за их вклад в разработку проекта.

