Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.021

Зазноба
Евгений Васильевич

Статистический метод машинного перевода

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1–40 80 03 «Вычислительные машины и системы»

_____

Научный руководитель
Одинец Дмитрий Николаевич
доцент каф.ЭВМ, к.т.н., доцент

_____

Минск 2015

ВВЕДЕНИЕ

В наше время информационные технологии прочно вошли в нашу жизнь. Компьютер теперь выполняет роль не только рабочего инструмента, он занимает важное место и в повседневной жизни. Быстрое развитие новых информационных технологий свидетельствует о всевозрастающей роли компьютерной техники в мировом информационном пространстве.

С каждым днем увеличивается число пользователей Интернета, сетевые технологии оказывают влияние на развитие самой науки и техники. В последнее время образование стало возможно получить дистанционно. Однако, развитие науки замедляется из-за существования языкового барьера. Эта проблема пока не нашла своего кардинального решения.

Последние годы объем предназначенной для перевода информации увеличился. Создание универсальных языков типа Эсперанто не привело к решению проблемы. Поскольку в наши дни данные появляются и изменяются круглосуточно, широко применяются электронные средства связи. В такой ситуации классический подход к осуществлению перевода часто не оправдывает себя, поскольку он требует значительных финансовых и временных затрат. В некоторых случаях более целесообразным представляется использование машинного или автоматического перевода.

Целью данной работы является создание модифицированного метода статистического машинного перевода.

Объектом исследования являются системы машинного перевода. Рассматриваются их особенности, разновидности, достоинства и недостатки.

Предметом исследований выступает модифицированный статистический метод машинного перевода. Также предметом исследований является анализ эффективности полученного метода с целью выявления его эффекта.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В исследовании детально изучается процесс осуществления машинного перевода. Изучаются существующие методы машинного перевода, их преимущества и недостатки. Также рассматриваются уже существующие наиболее популярные системы машинного перевода.

Больше всего внимания уделяется статистическому методу машинного перевода. Этот метод является наиболее эффективным, что подтверждается результатами работы систем, построенных на его основе. Был рассмотрен математический аппарат, лежащий в основе данного метода, подробно разобраны этапы обучения, вычисления модели перевода и декодирования результатов системы. Разобраны наиболее популярные методы поиска подходящего перевода в базе. Также были рассмотрены этапы EM-алгоритма, применяющегося для определения параметров модели и выравнивания.

Исследование затронуло тему модификации метода статистического перевода. Был продемонстрирован модифицированный алгоритм статистического машинного перевода, разобраны его отличия от оригинального метода. В исследовании были описан принцип работы синтаксического анализатора, лежащего в основе модифицированного метода. Также была рассмотрена грамматика связей, с использованием которой из декодированных (то есть переведенных) слов в итоге составляется предложение.

После изучения теоретических основ грамматики связей был спроектирован программный модуль, позволяющий визуализировать анализ текста с помощью грамматики связей. Модуль имеет потенциал, что в будущем позволить собрать связный текст на основе полученного от декодера набора переведенных слов.

СОДЕРЖАНИЕ РАБОТЫ

Исследование состоит из 4 глав. Уклон исследования направлен на область машинного перевода.

В первой главе приведено исследование методов машинного перевода, существующих в настоящее время. Изучаются все доступные методы, рассматриваются их достоинства и недостатки. Также приводится обзор наиболее популярных систем машинного перевода, реализующих эти методы. Помимо этого был проведен обзор существующих систем машинного перевода.

Вторая глава содержит детальное исследование математического аппарата статистического метода машинного перевода. Поочередно рассматриваются этапы работы алгоритма, такие как обучение системы, вычисления модели перевода и декодирование полученных результатов. Приводятся формулы, по которым вычисляется модель перевода, выполняется поиск наилучшего результата перевода, а также определение параметров модели и выравнивания.

В третьей главе детально описывается модифицированный метод статистического машинного перевода. Показываются отличия модифицированного метода от базового. Рассматриваются теоретические основы синтаксического анализа, который применяется для разбора заданного предложения и разбиения его на N-граммы на его основе. Также разбирается конкретный пример работы анализатора. Помимо этого, подробно и на примере рассматривается применение грамматики связей для составления предложения из полученных от декодера слов.

Четвертая глава описывает разработку модуля, работающего с грамматикой связей. Рассматриваются особенности реализации этого модуля. Рассматривается создание специализированного словаря, содержащего коннекторы. Были приведены фрагменты исходного кода, демонстрирующие разбор предложения, выделение из него дизъюнктов и грамматических связей.

ЗАКЛЮЧЕНИЕ

В работе была решена задача разработки модифицированного метода статистического перевода текста. Проанализированы основные подходы к машинному переводу, проведен их сравнительный анализ. Был рассмотрен математический аппарат базового метода статистического перевода, а именно особенности обучения, вычисления модели перевода и декодирование полученных результатов. Приведен модифицированный метод статистического перевода, рассмотрены его основные механизмы.

В рамках работы были разработаны модифицированный метод статистического машинного перевода, а также модуль работы с грамматикой связей. Модифицированный метод позволяет получать результаты более высокой точности по сравнению с базовым.

Область знаний, подвергнутая исследованию, на сегодняшний день является очень актуальной благодаря непрерывному росту объема информации на всех языках мира и необходимости ее моментальной обработки.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1-А.] Зазноба, Е.В. Статистический метод машинного перевода / Е.В. Зазноба, Д.Н. Одинец // II Международная научно-практическая конференция «Образование и наука в современных условиях»: Тезисы – Чебоксары, 2015.