РАЗРАБОТКА СВОБОДНОЙ/ОТКРЫТОЙ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА С КАЗАХСКОГО ЯЗЫКА НА АНГЛИЙСКИЙ И РУССКИЙ ЯЗЫКИ (И ОБРАТНО) НА БАЗЕ ПЛАТФОРМЫ «APERTIUM»

А.Т. Колбаева, *магистрант*
А.В. Баширов, *канд. техн. наук, руководитель лаборатории НИИ ЭПИ*
А.С. Цицина, *магистр, старший преподаватель*
Карагандинский университет Казпотребсоюза
(Казахстан, г. Караганда)

DOI:10.24412/2500-1000-2022-2-1-96-99

Аннотация. *Данная статья посвящена описанию процесса разработки свободной/открытой системы машинного перевода с казахского языка на английский и русский языки (и обратно) на базе платформы «Apertium». В процессе разработки были определены конкретные языки программирования и фреймворки, необходимые для разработки веб-приложения и приведены хорошие практики. А также в статье проведен анализ действующих систем машинного перевода.*

Ключевые слова: *язык программирования, фреймворк, веб-приложение, разработка, интеграцию систем, системы машинного перевода.*

Чтобы осуществить выбор наиболее приемлимой платформы для системы автоматического перевода с казахского языка и на другие, целесообразно описать основные особенности и недостатки имеющихся программных средств. С такой целью, при разработке открытой системы машинного перевода, были рассмотрены такие действующие системы, как Apertium, PROMT, SYSTRANet и, конечно же, действующий гигант в этой отрасли Google Переводчик.

Apertium – это открытая платформа машинного перевода. Изначально она была создана для связанных языковых пар, но теперь поддерживает даже такие необычные пары, как английский – каталанский. Но пары нам диктует сама система. Если нет нужной пары для перевода, эта платформа дает возможность самим пользователям добавлять пары для перевода, заполнив трехуровневый словарь. Во время посещения сайта, с целью анализа системы, в списке языков для перевода не было пары английского на русский и наоборот. Хотя эти языки являются одними из самых распространенных языков в мире, что считаю большим минусом системы.

PROMT (PROject об Machine Translation) – это проект машинного перевода, основанный российской компанией, которая занимается разработкой систем машинного перевода. А также компания занимается исследованиями и разработками в области популярного направления искусственного интеллекта. Изначально за основу перевода PROMT была взята технология машинного перевода – Rule-based machine translation, основанная на правилах. Но в ногу со временем, в 2019 году компания представила новую технологию машинного перевода на основе нейронных сетей – PROMT Neural. Алгоритмы данной технологии анализируют текст и решают, какой перевод лучше подходит для перевода предоставленного фрагмента текста. Эта технология дает предоставлять более качественный результат машинного перевода. Одной из главной возможностью PROMT является осуществление перевода документов с сохранением структуры и форматирования загруженного документа и перевод сайтов целиком, с сохранением структуры и гиперссылок страницы.

SYSTRAN – система машинного перевода, основанная в 1968 году Питером Тома. Система является одной из старейших компаний по машинному переводу, кото-

рая была создана для работы по переводу русско-английских текстов для ВВС США во время холодной войны. SYSTRAN проделал большую работу для Министерства обороны США и Европейской комиссии. До 2012 года «Yahoo!» и «Babel Fish» использовали технологию SYSTRAN для перевода. А также до 2007 года был использован языковыми инструментами Google. В 2010 году компания реализовала первую в своем роде на рынке технологию статистического машинного перевода (Statistical Machine Translation, SMT). Система дает возможность переводить конфиденциальные данные, благодаря строгой политике конфиденциальности данных. Не хранят и не используют повторно переводимые данные, которые остаются исключительной собственностью. В SYSTRAN есть возможность получить перевод на 50 популярные языки мира, но казахского языка в списке нет.

Ну и последним объектом анализа действующих систем в области машинного перевода является Google Translate. На сегодня система нейронного машинного перевода Google Translate использует большую сквозную искусственную нейронную сеть, которая пытается выполнять глубокое обучение, в частности, сети с долговременной краткосрочной памятью. По словам исследователей Google, он переводит «целые предложения целиком, а не просто по частям. Он использует этот более широкий контекст, чтобы определить наиболее релевантный перевод, который затем перестраивает и корректирует, чтобы он больше походил на человека, говорящего с правильной грамматикой». В документе Google Neural Machine Translation (GNMT) описывается интересный подход к глубокому обучению в производстве. Документ и архитектура нестандартны, во многих случаях далеко отклоняясь от того, что вы могли бы ожидать от архитектуры, которую вы найдете в академических материалах. Акцент делается на то, чтобы система оставалась практичной, а не на погоне за современным уровнем техники с помощью типичных, но требующих больших вычислительных ресурсов настроек. Во время проводимого анализа действующих систем, у такого гиганта не выявлены недостатки, за исключением не точного перевода слов на казахский и некорретный перевод получаемого переведенного текста.

Таким образом, реализуемая система будет акцентировать свою работу на качество перевода на казахский язык.

По сравнению с традиционными алгоритмами статистического и основанного на правилах перевода, это программа позволит получить более точный и естественно звучащий текст на казахский язык. В разработке использовалось статистический перевод (Statistical Machine Translation, SMT), который был доминирующей в машинном переводе на протяжении долгих десятков лет. Модель статистического машинного перевода, осуществляет поиск наиболее вероятного перевода предложения с использованием данных, полученных из параллельных корпусов системы. А также разрабатываемая систем будет иметь следующие достоинства:

- недопустимость грамматических ошибок, логических нарушений и синтаксиса;
- отсутствие противоречий в терминологии (единая терминология согласно словарю);
- восприятие переведенного текста, как родного (например, перевод теста с казахского на русский требует изменения порядка слов и грамматических конструкций).

Особенностью разрабатываемого приложения - предоставить пользователям новый удобный и функционально достаточный web-интерфейс для работы с открытой системы машинного перевода с казахского языка на английский и русский языки (и обратно). Отличительной функциональностью системы от конкуретов будет возможность получения перевода текста с казахского языка на латиницу.

После проведенного анализа и определения алгоритма работы, переходим к следующему этапу в разработке. Первым нужно определить вид продукта, структуру базы данных и архитектуру взаимодействия частей системы.

Для создания доступности приложения всем желающим и набора большой популярности, было решено, создать веб-приложение. Разработка Web-приложения уже говорит о том, что архитектура системы будет создана на основе клиент-серверной архитектуры. Архитектура системы клиент-сервер формулирует принципы виртуального общения между локальными компьютерами, а все правила и принципы взаимодействия находятся внутри протокола.

Для ускорения работы приложения, бизнес-логика приложения будет разделено на frontend и backend часть. Backend часть и есть серверная часть, где будет проводиться основная бизнес-логика, расчеты и алгоритмы. Через Frontend часть будет осуществлена интерфейсная часть, где пользователь будет отправлять необходимые запросы на серверную часть. Сервер в свою очередь будет генерировать и отправляет HTML-код в зависимости от запроса пользователя.

Реализация серверной части приложения были использованы такие инструменты, как Spring Framework, Hibernate.

Контейнером для внедрения зависимостей, с несколькими удобными слоями выступил Spring Framework. Spring – самая популярная среда разработки приложений для корпоративного Java. Spring Framework – это платформа Java с открытым исходным кодом. Spring Framework можно использовать при разработке любого приложения Java, но существуют расширения для создания веб-приложений поверх платформы Java EE. Платформа Spring нацелена на упрощение разработки J2EE и продвигает передовые методы программирования, используя модель программирования на основе POJO.

Связка ООП и реляционной базы данных машинного переводчика сопровождена Hibernate Framework. Hibernate не только заботится об отображении классов Java в таблицы базы данных (и типов данных Java в типы данных SQL), но также предоставляет средства запроса и извлечения данных. В этом учебном пособии вы узнаете, как использовать Hibernate для разработки веб-приложений на основе баз данных простыми и легкими шагами.

Клиентская часть приложения реализованы с помощью Angular, PHP, JavaScript, HTML и CSS. JavaScript, HTML и CSS не требуют представления, так как они являются основными строительными блоками веб-страниц.

Подводя итоги, можно сделать вывод, что выбор платформы для разработки, должен соответствовать определенным критериям, а сам продукт иметь свою специфичность и аудиторию. Выбор метода статистического перевода связан со спецификой функционирования казахского языка, который невозможен в действующей платформе «APERTIUM». Выбор разработки веб-приложения связан с желанием охвата большой аудитории и быстрого роста количества пользователей.

Библиографический список

1. Никольский А.П. JavaScript на примерах. Практика, практика и только практика. – Санкт-Петербург: Наука и Техника, 2018. – 272 с.

2. Кириченко А.В., Дубовик Е.В. Динамические сайты на HTML, CSS, Javascript и Bootstrap. Практика, практика и только практика. – 2-е изд. – Санкт-Петербург: Наука и Техника, 2018. – 272 с.

3. Сидорова Е.А., Загорулько М.Ю. Программный инструментарий разработки лингвистических ресурсов // Труды III Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» OSTIS2013. – Минск: БГУИР, 2013. – С. 159-164.

4. Яворский В.В., Баширов А.В., Емелина Н.К., Рахимбекова А.Е., Чванова А.О., Байдикова Н.В. Развитие смешанной формы обучения в процессе совершенствования информационно-коммуникационного обеспечения ВУЗа // Международный журнал экспериментального образования. – 2017. – №7. – С. 60-64.

DEVELOPMENT OF A FREE/OPEN MACHINE TRANSLATION SYSTEM FROM KAZAKH INTO ENGLISH AND RUSSIAN (AND VICE VERSA) ON THE BASIS OF THE «APERTIUM» PLATFORM

A.T. Kolbayeva, *Graduate Student*
A.V. Bashirov, *Candidate of Technical Sciences, Head of Laboratory, Research Institute of EPI*
A.S. Tsitsina, *Senior Lecturer, Master*
Karaganda University of Kazpotrebsoyuz
(Kazakhstan, Karaganda)

Abstract. *This article is devoted to the description of the process of developing a free/an open machine translation system from Kazakh to English and Russian (and vice versa) based on the «Apertium» platform. During the development process, specific programming languages and frameworks necessary for the development of a web application were identified and good practices were given. The article also analyzes machine translation systems.*

Keywords: *programming language, framework, web application, development, system integration, machine translation systems.*