**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «АСТРАХАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. В.Н.ТАТИЩЕВА»**

Факультет цифровых технологий и кибербезопасности

Кафедра цифровых технологий

**ЛАБОРАТОРНАЯ РАБОТА №1**

**Разработка архитектуры системы**

**«Машинный перевод с английского языка на русский язык»**

выполнена в рамках изучения дисциплины

«Архитектура систем ИИ»

Направление подготовки: 09.04.02 Информационные системы и технологии Направленность (профиль): «Проектирование и разработка

систем искусственного интеллекта»

Исполнитель:

студент группы ИИ-15

\_\_\_\_\_\_\_\_\_ Исназаров Р.К.

Научный руководитель:

к.т.н., доцент кафедры ПМИ,

\_\_\_\_\_\_\_\_\_\_\_\_ Гордеев И.И.

Астрахань – 2023

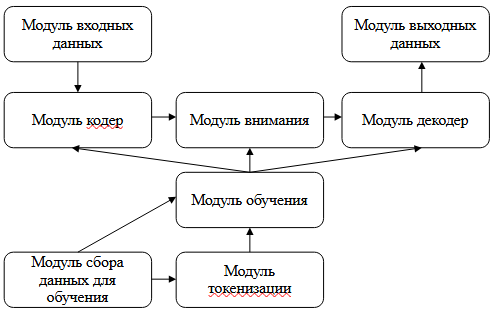


Рис. 1 Схема архитектуры проектируемой системы

Система будет состоять из следующих компонентов: модуль сбора данных для обучения, модуль токенизации, модуль обучения, модуль входных данных, модуль кодер, модуль внимания, модуль декодер, модуль выходных данных.

Модуль сбора данных для обучения будет принимать файл, состоящий из переводческих пар слов, фраз или предложений на обоих языках. Данный модуль будет проводить предобработку данных:

* удаление лишних символов, таких как знаки препинания, перенос строк и т.д.;
* перевод всех символов в нижний регистр;
* удаление строк, в которых нет переводческой пары (пустые строки или строки, в которых слово или предложение на одном языке).

После предобработки мы получим тренировочный корпус данных, который также будет представлять из себя словарь слов и предложений.

Модуль токенизации предназначен для создания словаря токенов для обоих языков. Токенизация — это разделение большого количества слов в тексте на еще более маленькие фрагменты, которые называют токенами. Токены позволяют лучше находить закономерности между словами. В нашем словаре будут пары токенов из обоих языков, которые наиболее близки по смыслу друг другу.

Модуль обучения предназначен для тренировки кодера, декодера и модуля внимания с помощью словаря токенов переводить слова или предложения на исходном языке в наиболее точный вариант на переводческом языке.

Модуль входных данных предназначен для приема и предобработки слов или предложений, которые мы хотим перевести на другой язык. При предобработке все слова переводятся в нижний регистр и удаляются знаки препинания. Далее предобработанные данные передаются кодеру.

Модуль кодер предназначен для разбиения предобработанных данных на токены, которые есть в словаре полученные после работы модуля токенизации. Кодер, в зависимости от степени своей обученности, будет стремиться к наиболее подходящему варианту разбиения. В кодере и декодере используются рекуррентные сети (Recurrent neural network, RNN) с долгой краткосрочной памятью (Long Short-Term Memory, LSTM). Слои LSTM состоят из цепочки ячеек. Каждая ячейка обрабатывает входные данные на основе результатов предыдущей цепочки. В нашем случае каждое слово будет обработано в зависимости от результатов обработки предыдущего слова. Это позволит лучше распознать зависимости между словами и контекст предложения. Кодер, как и декодер, состоят из нескольких слоев LSTM. Каждый слой LSTM результат своей обработки передает следующему слою. Чем больше слоев, тем более точнее перевод у нас получиться, но также тем меньше у нас будет скорость обучения. Наиболее оптимальным количеством слоев считается 8.

Модуль внимания принимает данные от кодера, для вычисления для каждого фрагмента его смысловой нагрузки, в зависимости от степени своей обученности.

Модуль декодер принимает закодированные данные с информацией о смысловой нагрузке для каждого фрагмента. На основе этих данных декодер, в зависимости от степени своей обученности, переводит его в наиболее вероятный вариант слова или предложения на переводческом языке, опираясь на словарь токенов. Как и в кодере, в декодере используется рекуррентные сети с 8 слоями LSTM.

Модуль выходных данных возвращает знаки препинания и верхний регистр в словах, где его перевели в нижний, и выводит нам итоговый вариант.