ИУ7-53Б, 11_KNY, Лысцев Никита

Статистический машинный перевод (СМТ) – разновидность МП, которая основана на поиске наиболее вероятного перевода предложения с использованием данных, полученных из двуязычной совокупности текстов (параллельного корпуса) в результате обучения (по языковым парам)[1].

Языковые пары можно определить как тексты на языке оригинала и соответствующие им переводные тексты. Чем больше языковых пар накоплено в базе данных, тем лучше получается текст на выходе. Анализируя языковые пары, система определяет наиболее вероятный вариант перевода.

При статистическом подходе используется такое понятие как «канал с помехами»[1]: рассматривая перевод с английского на русский, предполагается, что английское предложение на самом деле – русское предложение, но искаженное неким шумом. Для корректного перевода или же «расшифровки шума» нам нужно знать, что в рассматриваемом случае говорят русскоязычные и как рассматриваемый текст искажается, что превращается в текст на английском языке. Перевод осуществляется после поиска такого русского предложения, которое максимизирует вероятность «встречи» английского предложения и соответствующего ему русского предложения. В основе такого поиска лежит теорема Байеса:

$$P(e|f) = P(f|e) \times P(e) \tag{1}$$

где e – русское предложение, f – английское предложение.

Таким образом, в основе перевода, то есть вероятности того, что английское предложение e будет переведено русским предложением f, лежат модель языка, то есть P(e), и модель перевода – P(f|e). Задача, которая ставится перед системой статистического машинного перевода – не перевод, как таковой, а декодирование. То есть, в данном случае, перевод – не что иное, как расшифровка исходного текста.

Алгоритм перевода можно изобразить следующей схемой.

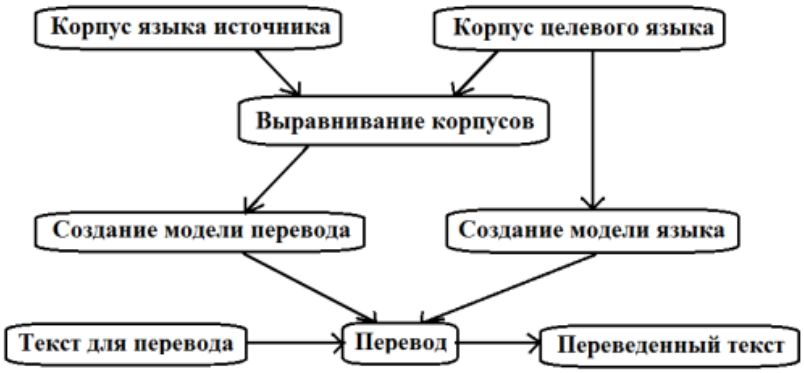

Рисунок 1 – Алгоритм статистического перевода [2]

Итак, весь алгоритм основывается на выравнивании, создании моделей и самом переводе. Рассмотрим данные составляющие более подробно

Выравнивание – процесс, в результате которого на выходе мы получаем проработанный текст, в котором за каждым предложением на языке оригинала идет соответствующее предложение на языке перевода. Для выравнивания текстов система использует инструментальное средство автоматического выравнивания параллельных текстов (alignment tool). С помощью таких инструментов система может проанализировать и привести в соответствие тексты оригинала и перевода [2].

Подготовленные таким образом тексты используются для обучения модели перевода.

Существуют 5 основных моделей статистического перевода: IBM 1, IBM 2, IBM 3, IBM 4, IBM 5 [1]. В первой модели для перевода предложения с одного языка на другой достаточно перевести все слова, а за их расстановку в предложении отвечает модель языка. Такая модель оперирует таблицей вероятностей попарно переводных соответствий двух языков. Вторая модель призвана сохранить порядок слов при переводе. Помимо таблицы вероятностей попарно переводных соответствий, используется таблица вероятностей обратных смещений. Третья модель устраняла недостаток второй модели, связанный с тем, что одному слову оригинала не могло соответствовать сразу несколько слов перевода. В моделях 4 и 5 появляются включения грамматики. Результатом формирования модели перевода является таблица вероятностей слов и фраз.

Модель языка оценивает фразы целевого языка и дает им соответствующую вероятность. В качестве модели языка в статистическом переводе используются n-граммные модели (в частности, триграммные), определяющие выбор слова на основе предшествующих слов.

На этапе перевода предложения в составе текста делятся на фразы. Для каждой фразы производится поиск наиболее вероятного перевода в таблице фраз, который максимизирует произведение вероятностей текста оригинала и текста перевода согласно теореме Байеса.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Статистическая система маашинного перевода [Электронный ресурс]. — — Режим доступа: `https://elib.bsu.by/bitstream/123456789/92831/1/58.pdf` (дата обращения: 07.10.2023).
2. Современные системы машинного перевода. Статический машинный перевод [Электронный ресурс]. — — Режим доступа: `https://elibrary.ru/download/elibrary_32330475_28416879.pdf` (дата обращения: 08.10.2023).