ИУ7-53Б, 11_KNY, Лысцев Никита

1 Статистический машинный перевод. Методы статистического машинного перевода

В современном мире существует большое количество статей, технической документации, книг, написанных на разных языках. Перевод таких текстов с одного языка на другой человеком вручную требует значительных временных и финансовых затрат, а также наличия соответствующих навыков. Возникает проблема перевода. Решением данной проблемы может стать максимальная автоматизация этого процесса. С этим помогают системы машинного перевода (МП).

1.1 Статистический машинный перевод

Статистический машинный перевод (СМП) – 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].

При статистическом подходе используется такое понятие как «канал помехами»(Noisy-Channel Model)[2]: 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» необходимо 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:

$$argmax_e P(e|f) = argmax_e P(f|e) \cdot P(e) \tag{1.1}$$

где e – предложение перевода, f – предложение оригинала.

Таким образом, в основе перевода, то есть вероятности того, что предложение оригинала f будет переведено конечным предложением e, лежат модель языка ($P(e)$ в формуле 1.6), и модель перевода – ($P(f|e)$ в формуле 1.6).

Модель языка должна присваивать оценку вероятности любому предложению конечного языка, а модель перевода должна присваивать оценку вероятности предложения оригинала при условии определенного предложения на конечном языке [1].

Таким образом, 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.

Формально, алгоритм работы системы статистического машинного перевода можно описать в следующих шагах:

1) Составляются параллельные корпусы [1] для языка источника и языка перевода;

2) Производится выравнивание [1] параллельных корпусов текстов;

3) Согласно выбранной модели перевода ищутся такие значения таблиц переводных соответствий [2], которые максимизируют вероятность части корпуса источника при имеющейся части корпуса языка перевода;

4) На основе корпуса языка перевода составляется модель языка.

5) На основе полученных данных для незнакомого предложения на языке источника ищется предложение на языке перевода, максимизирующее произведение вероятностей, присваеваемых моделью языка и моделью перевода.

Алгоритм перевода можно изобразить следующей схемой.

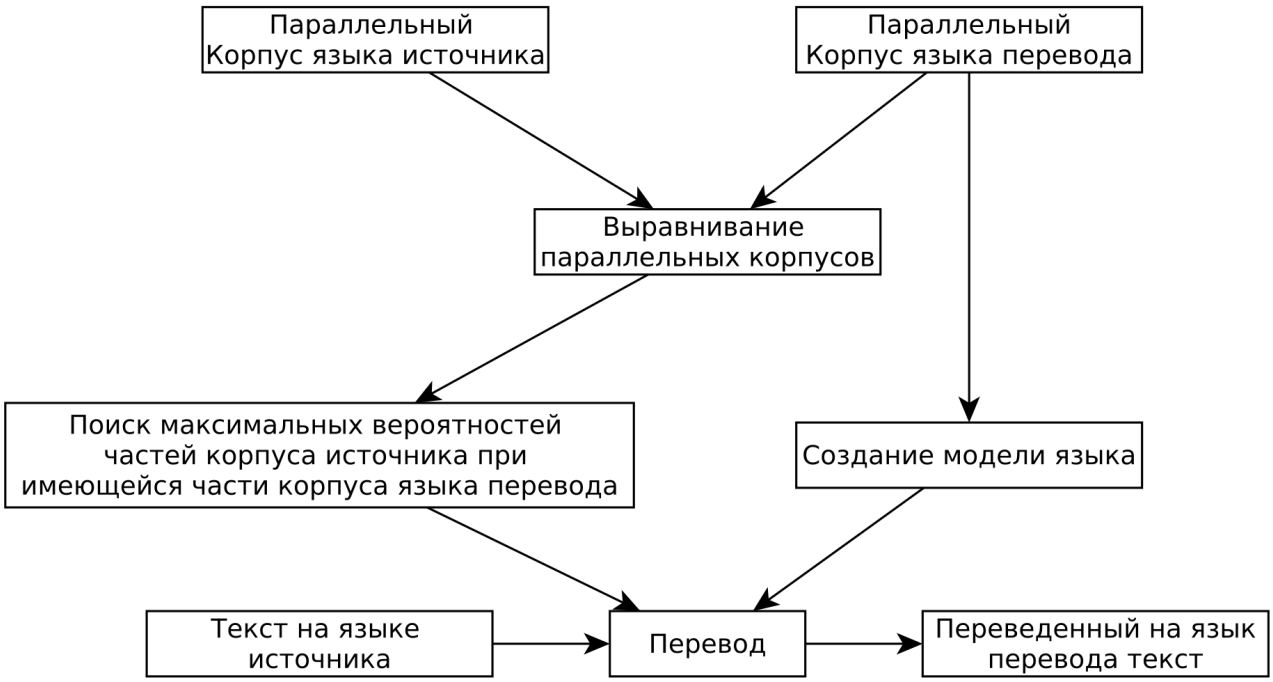

Рисунок 1.1 – Алгоритм работы системы статистического машинного перевода

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

1.1.1 Выравнивание параллельных корпусов

Параллельный текст – текст на одном языке вместе с его переводом на другой язык. Параллельный корпус – большие собрания параллельных текстов [1].

Выравнивание параллельных корпусов – процесс, в котором для каждого параллельного текста каждому его предложению на языке источника ставится в соответствие 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 [3].

Подготовленные таким образом тексты используются для обучения модели

перевода.

1.1.2 Модель перевода

Модель перевода, или таблиц перевода – это таблица-словарь, в которой для всех известных системе слов и фраз на одном языке перечислены все возможные их переводы на другой язык и указана вероятность этих переводов [3].

1.2 Модель языка

Модель языка оценивает фразы целевого языка и дает им соответствующую вероятность. В качестве модели языка в системах статистического машинного перевода используются преимущественно используются различные модификации n - граммной модели, утверждающей, что грамматичность выбора очередного слова определяется лишь тем, какие $(n-1)$ слов идут перед ним. Вероятность каждого n - грамма определяется по его встречаемости в параллельном корпусе[1].

Рассмотрим математику n - граммных моделей с $n = 3$:

$$\begin{gathered} P(e) = P(e_1, e_2, \ldots, e_n) = \\ P(e_1) \cdot P(e_2|e_1) \cdots P(e_n|e_1, e_2, \ldots, e_{n-1}) \simeq \\ P(e_1) \cdot P(e_2|e_1) \cdots P(e_n|e_{n-2}, e_{n-1}) \end{gathered} \tag{1.2}$$

где e – предложение на языке перевода, $P(e)$ – вероятность перевода всего предложения e, e_i – слово в предложении e, $i = \overline{1,n}$, $P(e_i)$ – вероятность перевода слова e_i, $i = \overline{1,n}$.

1.2.1 Декодер

Декодер – составляющая переводчика, которая непосредственно переводом. Для каждого предложения исходного текста он подбирает все варианты перевода, сочетая между собой фразы из модели перевода, и сортирует их по убыванию вероятности. Затем все получившиеся варианты декодер оценивает с помощью модели языка.

1.3 Методы статистического машинного перевода

В статистическом машинном переводе существует два основных метода перевода:

– метод перевода, основанный на словах;

– метод перевода, основанный на фразах.

Каждому из вышеперечисленных методов соответствуют свои модели перевода. Рассмотрим каждый из методов более подробно.

1.3.1 Метод перевода, основанный на словах

Введем некоторые обозначения:

– f – слово на языке источника, с которого нужно переводить;

– e – слово на языке приемника, на который нужно переводить.

Понятие статистического машинного перевода подразумевает использование статистики. На основе анализа параллельных корпусов для каждого слова f в предложении выполняется подсчет количества переводов для каждого возможного варианта перевода e на язык приемника. Далее, на основе этих подсчетов, производится оценка распределения вероятности лексического перевода.

Формально, находится функция [2]

$$p_f : e \to p_f(e) \tag{1.3}$$

что для слова f возвращает вероятность для каждого выбора перевода e, которая указывает, насколько вероятен этот перевод.

Функция 1.3 должна обладать следующим свойством:

$$\sum_e p_f(e) = 1 \tag{1.4}$$

После получения вероятностного распределения для каждого слова f в предложении выбирается наиболее вероятный перевод e. Сопоставление слова f

и его перевода e в предложении на языках источника и приемника определяются функцией выравнивания:

$$a : j \to i \tag{1.5}$$

которая отображает каждое выходное слово в позиции i к входному слову в позиции j.

IBM Model 1

Генеративный метод моделирования – метод разбиение процесса генерации данных на более мелкие шаги, моделирование более мелких шагов с помощью вероятностных распределений и объединение шагов в последовательную историю [2].

Поскольку прямое моделирование распределения вероятностей перевода для полных предложений сложно оценить (большинство предложений встречается только один раз, даже в больших текстовых коллекциях), применяется генеративный метод моделирования: процесс разбивается на более мелкие этапы, в нашем случае на перевод отдельных слов, а их расстановку в правильном порядке обеспечит модель языка.

IBM Model 1 – генеративная модель перевода предложений, основанная исключительно на распределении вероятностей лексического перевода [2]. Данная модель генерирует несколько различных переводов предложения, каждый из которых имеет разную вероятность. Единственным массивом данных, которым оперирует IBM Model 1 является таблица вероятностей попарно переводимых соответствий слов двух языков [1].

Обучение IBM Model 1 производится на параллельных корпусах, выровненных н уровне предложений. IBM Model 1 допускает ситуацию, в которой наиболее употребительным переводом нескольких смысловых слов может быть признано одно высокочастотное – например, служебное – слово конечного языка, и данная модель не всегда правильно учитывает порядок слов в предложении [1].

IBM Model 2

Чтобы сохранить при переводе информацию, заключенную в порядке слов, была предложена IBM Model 2.

В IBM Model 2 помимо таблица вероятностей попарно переводимых соответствий слов двух языков вводится таблица распределения вероятностей выравнивания, то есть вероятностей, что при определенной длине предложения на языке перевода l_e и длине предложения на языке источника l_f отображает каждое выходное слово в позиции i к входному слову в позиции j.

Таким образом, перевод с помощью IBM Model 2 можно описать в двух этапах:

1) Выполнение лексического перевода, как в IBM Model 1;

2) Выполнение этапа выравнивания.

1.3.2 Метод перевода, основанный на фразах

Наиболее эффективные в настоящее время системы статистического машинного перевода основаны на моделях, основанных на фразах: моделях, которые переводят небольшие последовательности слов за раз [2].

Под фразой понимается любая многословная единица предложения [2]. В методе перевода, основанном на фразах, наименьшей единицей перевода является не слово, как в первом методе, а фраза.

Алгоритм перевода следующий:

1) Входное предложение на языке источника разбивается на фразы;

2) Каждая фраза переводится на фразу в языке перевода;

3) Фразы на языке перевода переупорядочиваются согласно с моделью языка.

В основе перевода на основе фраз лежит правило Байеса [2]:

$$e_{best} = argmax_e P(e|f) = argmax_e P(f|e) \cdot P_{LM}(e) \tag{1.6}$$

где e_{best} – наиболее вероятный перевод предложения f на языке источника, $P_{LM}(e)$ – модель языка, $P(f|e)$ – модель перевода.

Данный метод имеет несколько преимуществ. Во-первых, слова, возможно, не являются лучшим атомарным средством единиц для перевода из-за частых сопоставлений один ко многим (и наоборот).Во-вторых, перевод групп слов вместо отдельных слов помогает устранить неоднозначность перевода. Есть и третье преимущество: если у нас есть большие учебные корпуса, мы можем выучить все более и более длинные полезные фразы, иногда даже запоминать перевод целых предложений [2].

IBM Model 3

IBM Model 2 не допускает возможности, что одному слову из предложения на языке источника соответствует соответствует несколько слов в предложении на языке перевода. Этот недостаток устраняется в IBM Model 3, где вводится понятие коэффициента деления слова оригинала, и, соответственно, таблица вероятностей каждого значения коэффициента деления для каждого слова [1].

IBM Model 4 и IBM Model 5

В IBM Model 4 и близкой к ней IBM Model 5 делается следующий шаг к включению понятия грамматики в систему статистического машинного перевода. В IBM Model 4 появляется понятие класса слов, определяемое автоматически для всех слов языка оригинала и языка перевода.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Статистическая система машинного перевода [Электронный ресурс]. — — Режим доступа: `https://elib.bsu.by/bitstream/123456789/92831/1/58.pdf` (дата обращения: 08.11.2023).
2. *Koehn P.* Statistical Machine Translation. — The Edinburgh Building, Cambridge CB2 8RU, UK : CAMBRIDGE UNIVERSITY PRESS, 2010.
3. Обзор аналитической, статистической и нейронной технологий машинного перевода [Электронный ресурс]. — — Режим доступа: `https://elibrary.ru/download/elibrary_32782386_78190438.pdf` (дата обращения: 08.11.2023).