Денисова Дарья Сергеевна
Denisova Daria Sergeevna
Магистрант
Master's Degree student
Балтийский федеральный университет имени Иммануила Канта
Immanuel Kant Baltic Federal University

МЕТОДЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА В СТАТИСТИЧЕСКОМ МАШИННОМ ПЕРЕВОДЕ
METHODS OF NATURAL LANGUAGE PROCESSING IN STATISTICAL MACHINE TRANSLATION

Аннотация на русском языке: Статья посвящена статистическогому машинному переводу, а именно вопросу качества выходного текста и возможным способам решения данной проблемы. Целью исследования является изучение технологий лингвистического анализа при переводе статистическим методом. В данной статье проанализированы основные принципы работы таких технологий как регулярные выражения, вычисление минимального расстояния при редактировании, н-граммы и анализ эмоциональной окраски текста. С помощью технологий лингвистического анализа может осуществляться качественный пред-переводческий анализ, а также анализ качества перевода для статистического машинного перевода, что, несомненно, лежит в основе качественного перевода.

The summary in English: The article is devoted to statistical machine translation, namely the questionable quality of the target text and possible solutions to this issue. The aim of the research is to study the technologies of linguistic analysis implemented in statistical machine translation. The basic principles of such technologies- as regular expressions, minimum edit distance, N-grams and sentiment analysis of the text have been analyzed in the article. With the help of technologies of linguistic analysis pre-translational analysis of high quality can be performed, as well as analysis of the quality of translation for statistical machine translation, which is undoubtedly the basis of a quality translation.

Ключевые слова: статистический машинный перевод, обработка естественного языка, регулярные выражения, вычисление минимального расстояния при редактировании, анализ эмоциональной окраски текста.

Key words: statistical machine translation, natural language processing, regular expressions, minimum edit distance, sentiment analysis.

Современные технологии могут в значительной степени облегчить труд человека. Также и труд переводчика. Порой переводчику нужно выполнять рутинные монотонные операции, а также предоставить качественный перевод в сжатые сроки. Но в век научно-технического прогресса эти операции можно поручить автоматизированному переводчику. Проблема в том, что перевод, полученный благодаря таким средствам, не всегда отвечает требованиям качества. Именно поэтому на сегодняшний день данная тема является актуальной.

Целью нашего исследования является изучение технологий лингвистического анализа при переводе статистическим методом. Основными методами исследования являются метод лингвистического описания, метод вычисления и сопоставительный метод. Теоретической базой являются работы Ю.Н. Марчука по компьютерной лингвистике и теории перевода, а также исследования профессоров Стэндфордского университета Д. Джаравски и К. Маннинга по обработке естественного языка. В основе исследования лежит анализ существующих технологий по обработке естественного языка и решение таких задач как исправление системой ошибок в оригинальном тексте, то есть пред–переводческий анализ, и анализ тональности текста. Изучение данных алгоритмов является обязательным условием дальнейшего исследования и улучшения качества перевода.

Но задумывались ли мы, каким образом система может решать определенные задачи? Обладает ли программа компьютера интеллектом или же, попросту, мышлением? Если система может решить определенные задачи, то целесообразно предположить, что система может мыслить, то есть обладает искусственным интеллектом.

Прежде чем обучить систему, необходимо знать, какого рода проблему вы собираетесь решить, а также суметь ее описать, пользуясь специальным языком, понятным этой программе, преобразовывая естественный язык в искусственный [1, с. 89].

Такая технология лингвистического анализа как регулярные выражения (Regular expressions) используется для поиска и редактирования слов и выражений текстовыми редакторами, что, несомненно, может являться частью оценки качества перевода. Регулярные выражения – это язык поиска определенной информации путем использования специальных символов и подстрок [4]. Программа, работающая с регулярными

выражениями, называется «Regexpal» или «Regex Tester». Допустим, во всем тексте нам нужно найти артикль «the». Если в верхнем окне мы зададим следующее условие поиска: [Tt]he, где квадратные скобки означают, что система будет искать не только «the» с прописной, но и с заглавной буквой, то помимо «the» как и с прописной, так и с заглавной буквы, программа нам выделит также сочетание этих букв внутри слова: «they», «other», «their», «them». Причина кроется в том, что мы не указали в правиле поиска, что нам нужно отдельное слово, а не сочетание этих букв. Используя символ ^, который означает отрицание, перепишем условие поиска: [^A-Za-z][Tt]he[^A-Za-z]. В результате, система выделит нам «the» как артикль. А чтобы выделить «the» в начале предложения, потребуется ввести такую комбинацию символов: The[^A-Za-z].

Такие технологии, как вычисление минимального расстояния при редактировании (Minimum Edit Distance) и н-граммы (N-grams) представляют собой наибольший интерес. Именно на этих технологиях основывается решение такой задачи, как исправление ошибок. N-граммы представляют собой лингвистические модели, целью которых является вычисление вероятности предложения или последовательности слов. Минимальное расстояние при редактировании – наименьшее количество действий для преобразования одного слова в другое.

Система подразделяет ошибки на два вида: несловесные (слова с ошибкой не существует или, по крайней мере, его нет в словаре) и словесные (слово с ошибкой – реальное слово [5]. При исправлении несловесной ошибки система ищет всех возможных «кандидатов», используя минимальное расстояние при редактировании по Домеро-Ливенштейну, а затем вычисляет тип ошибки (удаление (deletion), вставка (insertion), замена (substitution), транспозиция (transposition)). В основе вычисления правильного слова лежит формула Байеса: P (w/x) = P (W) × P (x/w), где P (w/x) – вероятность, что неправильная буква x в слове будет исправлена на

правильную букву w; P (W) – вероятность самого правильного слова; P (x/w) – вероятность замены w на x, то есть вероятность допущения ошибки, а также формулы для каждого типа ошибки, как показано ниже .

$$P(x/w) = \begin{cases} \dfrac{del\ [w_{i-1}, w_i]}{[w_{i-1}\ w_i]} \\ \dfrac{ins\ [w_{i-1}, x_i]}{[w_{i-1}]} \\ \dfrac{sub\ [x_i, w_i]}{[w_i]} \\ \dfrac{trans\ [w_i, w_{i+1}]}{[w_i\ w_{i+1}]} \end{cases}$$

Рисунок 1. Формула Байеса

Источник [5]

Для вычисления P (x/w) система прибегает к помощи матрицы ошибок (confusion matrix) и корпуса сервиса GitHub. Для вычисления P (W) система смотрит вероятность слова в сервисе Google N-grams. После вычисления P (w/x), система оценивает каждого кандидата исходя из контекста, также используя сервис Google N-grams.

Возьмем предложение с ошибкой: «He did it with his bary hands». Возможные кандидаты и типы ошибок представлены ниже.

Таблица 1. Типы ошибок

Ошибка	Кандидаты	Правильная буква (w)	Неправильная буква (x)	Тип ошибки
bary	bar	-	y	вставка
bary	bury	u	a	замена
bary	bare	e	y	замена
bary	bark	k	y	замена
bary	bray	ra	ar	перестановка
bary	barky	k	-	удаление

Источник: анализ автор

Интересно отметить, что на начальном этапе анализа наибольшую вероятность имело слово «bar», на втором месте – «bare». Это можно объяснить тем, что в данном примере такой тип ошибки как вставка «y»

является более типичной, нежели замена «e» на «y». Но после анализа слов в контексте с помощью сервиса Google N-grams система все же исправит «bary» на «bare», а не на «bar», и, соответственно, поступит правильно. Данный пример является демонстрацией того, что контекст при исправлении ошибок имеет огромную роль и очень важно обучить систему оценивать контекст.

В основе исправления словесной ошибки также лежит контекст. В отличие от исправления несловесных ошибок, в данном случае анализируются не только кандидаты с минимальным расстоянием, но и само слово, где допущена предполагаемая ошибка. При исправлении данного типа ошибки также используется сервис Google N-grams.

Возьмем, например, словосочетание «three was an» из предложения «Three was an accident on the street». Поиск кандидатов можно наглядно представить в виде следующей схемы.

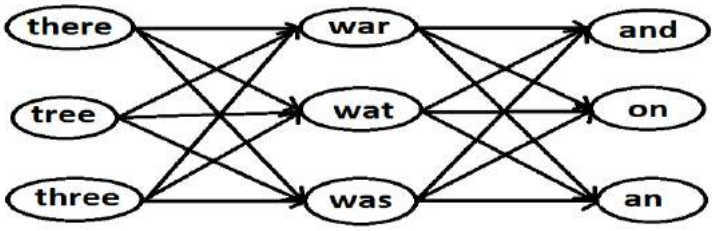

Рисунок 2. Поиск кандидатов

Источник: анализ автора

Итак, оценив все возможные варианты, система исправит первое слово на «there» и наше предложение примет вид «There was an accident on the street».

Анализ эмоциональной окраски текста является, пожалуй, одним из самых ярких направлений обработки текста. Его можно определить как автоматическое извлечение особенностей текста, то есть эмоционально окрашенную лексику и отношение авторов по отношению к тому, что

сказано в тексте. Анализ полярности и субъективности являются основными задачами анализа эмоциональной окраски текста. Первый предполагает наличие словаря «хороших» и «плохих» слов. В анализируемом тексте каждому слову присваивается оценка: обычно +1 в случае позитивной тональности, и -1 – в случае негативной. Данный подход имеет ограничения, такие как, пренебрежение контекстом и близлежащими словами. Второй анализ определяет, субъективен текст или нет. Стоит отметить, что под субъективным текстом мы подразумеваем тот текст, в котором содержится личное мнение автора, а объективный текст содержит факты. Именно тот текст, в котором выражается мнение автора и подлежит анализу.

Также к анализу тональности относится определение степени полярности, то есть насколько текст позитивен и наоборот. Например, слову «болеть» может быть присвоена такая оценка как -2, что будет означать среднюю степень негативности, а слову «мучать» – -5, что означает сильно отрицательное слово.

Альтернативным методом определения полярности является определение наивероятнейшей общей полярности слов в тексте. В рамках данного метода оценивается то, как часто эти слова встречаются с множеством слов с заранее известной и недвусмысленной полярностью, например, к таким слова относятся – «хороший», «отвратительный» и т.д. Смысл заключается в том, что слова, имеющие позитивную окраску, чаще всего появляются с другими словами с такой же окраской и наоборот. Это значит также, что если нам уже известна эмоциональная окраска слова «великолепный» и это слово соединяется со следующим прилагательным союзами «и» и «или», то система автоматически отнесет второе слово в базу данных позитивно окрашенных слов. А если после такого слова следует союз «но», система отнесет второе слово, находящееся после союза, в другую базу данных, нежели первое слово. [**Ошибка! Источник ссылки не найден.**]

Популярной онлайн-программой по оценке тональности является Sentiment Analysis with Python NLTK Text Classification. Данная программа дает оценку по трем составляющим: положительная, отрицательная и нейтральная. К тому же поддерживает тексты в 50 000 символов. Программа работает с тремя языками: английский, голландский и французский. Система дает оценку по шкале от 0 до 1. Субъективность (subjectivity) означает, насколько оценка эмоционально окрашена, полярность (polarity), в свою очередь, – насколько позитивна или негативна оцениваемая информация.

Текст для анализа возьмем с сайта BBC News: в данном тексте речь идет о землетрясении, которое повлекло за собой много смертей. С точки зрения логики, текст несет негативную оценку. Давайте проверим, как система проанализировала данный текст. Программа отнесла текст к нейтральному. Возникает естественный вопрос: почему? Возможно, потому что в тексте много фактов, которые автоматически присваивают тексту «статус» нейтрального. Также влияние может оказать тот факт, что программа основана на юниграммах, то есть анализирует отдельно каждое слово. О последнем также говорит следующий пример, где «not bad» оценивается как отрицательное высказывание, хотя это не так.

Также немалый интерес представляет программа SentiStrength. Данный сервис имеет три варианта анализа тональности: быстрый тест, тест, ориентированный на определенные слова и тест, ориентированный на определенную тематику. Рассмотрим первый вариант. В данном случае, в отличие от других, оценка может производиться по трем составляющим: положительно, отрицательно окрашенные тексты, а также нейтральные. Анализ осуществляется по шкале от 1 до 5. К сожалению, работая с данной программой онлайн, можно анализировать только небольшие тесты, буквально размером с одно среднее предложение.

Интересной особенностью данного сервиса является возможность учитывать количество букв. В первую очередь это относится к текстам личной переписки, где вместо «happy» мы можем написать «haaapppy». Второй вариант, как нам известно, более эмоционально окрашен и система это распознает. Также стоит отметить слова, тональность которых может быть неоднозначной. Например, слово «miss», которое может означать «скучать» и «потерять». И как показывает практика, при анализе такого рода слов позитивная и негативная оценки кардинально не отличаются.

Итак, нами были рассмотрены и охарактеризованы одни из основополагающих алгоритмов лингвистического анализа. Одно из достоинств статистического машинного перевода – в том, что развивается вместе с языком [2, с. 163]. Основываясь на технологиях лингвистического анализа, вышеперечисленные алгоритмы могут осуществить качественный пред-переводческий анализ, а также анализ качества выходного текста для статистического машинного перевода, что, несомненно, лежит в основе качественного перевода.

Литература:

1. Марчук Ю.Н. Модели перевода: учеб. пособие для студ. Учреждений высш. проф. Образования. – М.: Издательский центр «Академия», 2010. – 176 с.

2. Мифтахова Р.Г. Влияние информационных технологий на развитие лингвистических норм // Вестник БашГУ, 2012, № 1. – С. 162-164.

3. URL: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf (Дата обращения: 8.04.2016).

4. URL: http://www.youtube.com/watch?v=hwDhO1GLb_4 (Дата обращения: 10.03.2016)

5. URL: http://www.youtube.com/watch?v=Z1m7McLIP9c (Дата обращения: 10.03.2016)