

УДК 519.768

ОБЗОР АНАЛИТИЧЕСКОЙ, СТАТИСТИЧЕСКОЙ И НЕЙРОННОЙ ТЕХНОЛОГИЙ МАШИННОГО ПЕРЕВОДА

Колганов Д.С., Данилов Е.А.

Пензенский государственный технологический университет, Пенза, e-mail: EvgenyDplus@mail.ru

В статье приводится обзор технологий машинного перевода. Рассмотрены области применения машинного перевода. Описана аналитическая технология машинного перевода, вначале рассматривается процесс перевода, затем приводятся достоинства и недостатки данной технологии. Описывается состав системы машинного перевода на основе статистической технологии, отмечаются ее сильные и слабые стороны. Последней рассматривается нейронная технология машинного перевода, которая начала развиваться сравнительно недавно, но уже и применяется в Google и Яндекс. Производится сравнение статистической технологии и нейронная технология машинного перевода, а также указываются достоинства и недостатки нейронной технологии. Приводятся рекомендации по использованию технологий машинного перевода. В завершении говорится о том, что нейронная технология машинного перевода является самой перспективной, но на сегодняшний день не является самой лучшей в первую очередь из-за недостатка обучающих данных в различных предметных областях.

Ключевые слова: аналитический машинный перевод, статистический машинный перевод, нейронный машинный перевод, RBMT, CBMT, NMT

OVERVIEW OF ANALYTICAL, STATISTICAL AND NEURAL MACHINE TRANSLATION TECHNOLOGIES

Kolganov D.S., Danilov E.A.

Penza State Technological University, Penza, e-mail: EvgenyDplus@mail.ru

The article gives an overview of machine translation technologies. Areas of application of machine translation are considered. The rule-based machine translation is described, the translation process is first considered, then the advantages and disadvantages of this technology are given. The composition of the machine translation system is described on the basis of corpus-based machine translation, its strengths and weaknesses are noted. The latter is considered the neural machine translation, which began to develop relatively recently, but already applied in Google and Yandex. A comparison of corpus-based and neural machine translation is performed, as well as the advantages and disadvantages of neural machine translation. Recommendations are given on the use of machine translation technologies. In the end, it is said that the neural machine translation is the most promising, but to date it is not the best, primarily due to the lack of training data in various subject areas.

Keywords: rule-based machine translation, corpus-based machine translation, neural machine translation, RBMT, CBMT, NMT

Об автоматическом переводе текстов впервые заговорили уже с момента появления первых электронных вычислительных машин. Потенциальные сферы применения машинного перевода достаточно многообразны. Например, машинный перевод полезен – на бытовом уровне, упрощая коммуникацию людей различных языковых групп, не обладающих должными языковыми навыками. Машинный перевод актуален и в бизнес-сфере, когда требуется перевод значительных объемов данных.

На данный момент существуют 3 технологии машинного перевода:

- аналитический машинный перевод;
- статистический машинный перевод;
- нейронный машинный перевод.

Аналитический машинный перевод (RBMT, rule-based machine translation, машинный перевод, основанный на правилах) [1] – исторически первая технология машинного перевода. Метод подразумевает использование набора правил перевода из исходного языка в целевой, написанных

лингвистом, и двуязычных словарей – набора лексических элементов обоих языков. Анализ в данном методе зачастую сильно завышен, а процесс перевода проходит следующие этапы:

1. Морфологический анализ, в ходе которого указывается род, число, лицо и другие морфологические характеристики слов. При этом возникает проблема многозначности, когда одно и то же слово может относиться к разным частям речи.

2. Объединение отдельных слов в группы.

3. Синтаксический анализ предложения, то есть определение членов предложения и их места в предложении. Сначала программа ищет сказуемое. Затем перед найденным сказуемым ищется подлежащее. Если его там нет, то алгоритм ищет подлежащее за сказуемым. Если подлежащего нет и там, то считается, что подлежащее отсутствует.

4. Синтез предложений – согласование найденных частей предложений и групп зависимых слов.

К достоинствам аналитического метода можно отнести:

- синтаксическую и морфологическую точность;
- стабильность и предсказуемость результата;
- возможность настройки на предметную область.

Из недостатков можно выделить:

- трудоемкость и длительность разработки: для разработки лингвистических правил, которых в системе аналитического перевода может насчитываться десятки тысяч, необходимо привлечение лингвистов; процесс разработки этих правил может занимать от нескольких месяцев до нескольких лет;

- необходимость поддерживать и актуализировать лингвистические БД;

- «машинный акцент» при переводе – ориентируясь исключительно на правила, такие системы зачастую могут игнорировать контекст, подставляя в целом верные, но не самые уместные варианты перевода отдельных лексем.

Статистический машинный перевод является подвидом метода, основанного на корпусах текста (*corpus-based machine translation, CBMT*) [1]. В основе CBMT состоит работа не с готовыми правилами, то есть не рационалистический (аналитический) подход, а эмпирический подход, то есть формирование правил на основе параллельных двуязычных корпусов текстов.

В основе технологии теорема Байеса: из предложения выделяются отдельные грамматические единицы – слова и фразы, перебираются все варианты перевода для каждого фрагмента и взвешивается вероятность каждого из них.

Система статистического перевода обычно состоит из трёх компонентов:

1. 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.

2. Вероятностная модель языка – это представление знаний системы о языке, на который нужно перевести текст. Она используется для того, чтобы после выбора наиболее вероятных вариантов перевода отдельных слов и фраз исходя из модели перевода, выбрать из этих вариантов наиболее подходящие, исходя из контекста.

3. 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.

Помимо этого, может также иметься таблица изменения порядка, указывающая, как можно менять порядок слов при переносе на целевой язык. Иногда также необходимо включать дополнительную лингвистическую информацию для языков с богатым словоизменением, например, русский язык.

К достоинствам статистических систем можно отнести:

- быстрая настройка: поскольку система обучается сама, лингвисты необходимы только для помощи в написании алгоритма анализа корпусов текстов; для дальнейшего обучения системы используются тексты, которые можно найти в свободном доступе;

- такие системы хорошо справляются с переводом сложных и редких слов, терминов и устойчивых выражений;

- легко добавлять новые направления перевода: если язык начнёт меняться, система заметит это как только к ней попадут соответствующие тексты;

- отсутствие глубокого анализа текста экономит вычислительные ресурсы.

Из недостатков можно отметить:

- статистические системы гораздо хуже работают для сильно непохожих друг на друга языков без использования сложных моделей типа *tree-to-tree/tree-to-string* (например, при переводе с английского на японский);

- «дефицит» параллельных корпусов: качество перевода сильно зависит от количества параллельных корпусов, для корректных переводов статистической системе необходимо как минимум 500 тысяч, в идеале от нескольких миллионов параллельных текстов;

- нестабильность перевода: несмотря на способность переводить устойчивые выражения, эти самые выражения и термины могут переводиться по-разному исходя из контекста;

- зачастую результат перевода похож на «собранный пазл»: хотя общий смысл предложения понятен, но части предложений существуют отдельно друг от друга.

Несмотря на сохранение актуальности аналитической и статистической технологий (в том числе и гибридной), последнее время активно развиваются методы перевода с использованием искусственных ней-

ронных сетей (neural machine translation, NMT) [1, 2].

На первый взгляд нейронный перевод очень похож на статистический, поскольку также использует анализ параллельных данных и формирует на основе этого анализа определённые зависимости и закономерности. Однако в основе данного метода лежат совершенно иные принципы.

В основе нейронного переводчика лежит механизм двунаправленных рекуррентных нейронных сетей, построенный на матричных вычислениях, который позволяет строить существенно более сложные вероятностные модели, чем статистические машинные переводчики.

Хотя нейронный перевод также используется для обучения параллельные корпуса, в процессе обучения он оперирует не отдельными фразами, но целыми предложениями. Одна из главных проблем состоит в том, что нейронной сети требуется гораздо больше корпусов для обучения, чем статистической системе: по меньшей мере порядка 100 миллионов токенов для адекватной работы, для переводов же надлежащего качества – не менее порядка 500 миллионов. Также для обучения подобной системы требуется куда больше вычислительных мощностей.

Однако главной причиной того, что нейронный перевод начал проявляться сравнительно недавно, является не столько аппаратные ограничения, поскольку тренировать подобные системы можно было и раньше, пускай ценой больших временных затрат, а скорее пришедшая мода на нейронные сети.

Ведущие разработчики систем нейронного перевода вели свои исследования уже достаточно давно, однако высокие ожидания общественности от возможностей нейронных сетей и опасения, что недостаточно совершенные системы нейронного перевода не оправдают эти ожидания, заставляли разработчиков не оглашать результаты своих разработок раньше времени. Однако же в районе 2015–2016 годов все разработчики подобных систем начали представлять свои варианты нейронных переводчиков один за другим. Так, Яндекс объявил о постепенном переходе на нейронный перевод с сентября 2017 года, а Google сделал это ещё в марте.

Что касается достоинств и недостатков нейронных сетей, они во многом схожи

с таковыми у статистических систем. С одной стороны, нейронная сеть, при должном количестве входных данных способна выдать практически идеальный или близкий к таковому перевод, поскольку она не просто обучается, она «понимает» принципы, по которым строится перевод. С другой стороны, во-первых, она требует куда больше вычислительных мощностей для своего обучения, и, во-вторых, в условиях недостатка параллельных данных, коих требует на порядок больше, чем статистическая сеть, нейронная сеть выдаст перевод крайне низкого качества, то есть нейронные переводчики ещё более требовательны к объёму входных данных, чем статистические системы.

В завершение можно выделить ряд тезисов:

- в общем случае нейронный автоматический перевод дает результат более высокого качества, чем «чисто» статистический подход;
- автоматический перевод через нейронную сеть лучше подходит для решения задачи «универсального перевода»;
- ни один из подходов к машинному переводу сам по себе не является идеальным универсальным инструментом для решения любой задачи перевода;
- для решения задач по переводу в бизнесе только специализированные решения могут гарантировать соответствие всем требованиям.

Несмотря на очевидные перспективы нейронных сетей в качестве основного инструмента машинного перевода в будущем, на текущий момент недостаток в первую очередь должного объёма данных в различных предметных областях не позволяет говорить о безоговорочном преимуществе нейронных сетей во всех сферах применения машинного перевода. Поэтому на текущий момент по-прежнему сохраняют актуальности как статистические системы, требующие меньший объём данных для обучения, так и аналитические системы, дающие стабильный и точный результат в узких областях.

Список литературы

1. How does Neural Machine Translation work? // SYSTRAN Blog: [Электронный ресурс], 2016. – URL: <http://blog.systransoft.com/how-does-neural-machine-translation-work>.
2. eMPTY Pages: A Deep Dive into SYSTRAN's Neural Machine Translation (NMT) Technology: [Электронный ресурс], 2016. – URL: <https://kv-emptypages.blogspot.ru/2016/09/a-deep-dive-into-systrans-neural.html>.