РАЗРАБОТКА И ИССЛЕДОВАНИЕ СИСТЕМ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА С РУССКОГО НА КАЗАХСКИЙ ЯЗЫК

Бохаева А.С.

Бохаева Асем Сактапбергеновна – магистрант,
кафедра информационных систем, факультет информационных технологий,
Казахский национальный университет им. аль-Фараби,
г. Алматы, Республика Казахстан

Аннотация: *в статье рассматриваются различные методы и исследования в области машинного перевода, приводятся определения нейронного и статистического машинного перевода, рассматриваются проблемы перевода для русско-казахской пары языков, анализируются методы улучшения качества перевода для русско-казахской пары языков, рассматривается применение в повседневной жизни обученных моделей нейронного машинного перевода, приводится пример обучения без сложных математических аспектов нейронного машинного перевода и примерный алгоритм системы.*

Ключевые слова: *нейронная система, машинный перевод, BLEU метрика.*

УДК 004.852

В последнее время нейронные системы используются во многих областях, которые могут обрабатывать данные быстрее и проще, чем ранее использовавшиеся системы статистического анализа. Примером широкого использования этих нейронных сетей является индустрия машинного перевода. За последние пять лет в этой области широко используется ряд готовых систем. Нейронный машинный перевод - это новое поколение машинного перевода, основанного на корпусе (также называемого управляемым данными, реже машинным переводом, основанным на корпусе). Он обучается на огромных корпусах пар сегментов исходного языка (обычно предложений) и их переводах, то есть, в основном, из огромной памяти переводов, содержащей сотни тысяч или даже миллионы единиц перевода. В этом смысле она похожа на технологию статистического машинного перевода, которая была самой современной до недавнего времени, но использует совершенно другой вычислительный подход: нейронные сети. Нейронный машинный перевод, несмотря на свое название, имеет только очень расплывчатую связь с нейронами или с тем, как работает мозг человека (или мозг переводчика). Название происходит от того факта, что нейронные сети (которые должны называться искусственными нейронными сетями), на которых основан NMT, состоят

из тысяч искусственных единиц, которые напоминают нейроны в том, что их выход или активация (то есть степень, в которой они возбуждены или подавлены) зависит от стимулов, которые они получают от других нейронов, и от силы связей, по которым эти стимулы передаются [1].

В NMT слова или единицы подслов, такие как символы или последовательности коротких символов, обрабатываются параллельным, распределенным способом: фактические состояния активации каждого нейрона в больших наборах нейронов обучаются строить распределенные представления слов и их контекстов, оба в контексте обрабатываемого исходного предложения и в контексте создаваемого целевого предложения. Представление - это снимок состояния активации каждого нейрона в определенной группе из них, обычно называемой слоем: список фиксированного размера (вектор) таких величин, как (+0.2, 0,1, -0.13, +0.01) Фактический вывод перевода производится из этих представлений. Чтобы получить представление о том, как векторы могут использоваться для представления знаний, представьте прямоугольную комнату, идеально выровненную с точками компаса. Любая точка внутри комнаты может быть расположена в юго-западном углу комнаты («источник»), используя три числа: сколько сантиметров далеко на север, сколько сантиметров на восток и сколько сантиметров в высоту над полом. Например, положение часов на тумбочке может быть представлено трехмерным вектором, например «(40, 120, 87)». 3. Теперь представьте, что, подобно лампочке, понятия (слова, предложения) могут быть помещены в пространство внутри этой комнаты: два одинаковых понятия в идеале должны быть близко друг к другу и, следовательно, иметь схожие координаты; очень разные понятия будут далеко друг от друга и, следовательно, имеют разные координаты. Трех измерений недостаточно для богатства, наблюдаемого в языке: кодировки слов и репрезентации предложений нуждаются во многих других измерениях, чтобы приспособить их и их взаимоотношения, обычно сотни. Большинству из нас трудно представить пространства с более чем тремя измерениями, но геометрия и математика приятно выходят за пределы трех измерений, и поэтому вычисление и хранение этих представлений - это только вопрос вычислительной мощности и памяти. Анализируя многие из современных систем машинного перевода основанных на нейросетях (например, NEMATUS, KERAS, Tensorflow, Theano), мы обнаружили, что использование системы Tensorflow для пары казахско-русского языка более приемлемо, чтобы изучить их преимущества и недостатки. TensorFlow - это программная библиотека с открытым исходным кодом для высокопроизводительных численных расчетов. Его гибкая архитектура позволяет легко развертывать

вычисления на различных платформах, от настольных компьютеров до кластеров серверов, мобильных и периферийных устройств [2].

NMT TensorFlow использует архитектуру кодер-декодер. Исходный текст дается на вход кодера, так как машина не распознает слова, здесь используется word embeddings. Это технология преобразования слов в вектор по близости значения.

Во-первых, давайте рассмотрим кодировку исходного предложения: представьте, что мы хотим перевести предложение «Я люблю осень.» на казахский. Представление для предложения рекурсивно формируется из векторных вложений отдельных словследующим образом *e*('Я'), *e*('люблю'), *e*('осень'), *e*('</s>') и *e*('.') (обратите внимание, что мы используем «e (...)» в качестве сокращенной записи для вектора кодирования, который может иметь сотни компонентов):

1. Сеть кодировщиков объединяет ранее существовавшее (предварительно изученное) кодирование для пустого предложения E (‘’) с встраиванием первого слова e («Я») для создания кодировки E («Я»).

2. Затем сеть кодировщика объединяет представление E («Я») и вложение *e*(«люблю») для создания кодировки E («Я люблю»).

3. В последовательных шагах E («Я люблю») и *e*(«осень») приводят к E («Я люблю осень»).

На нейронном языке такая сеть называется рекуррентной нейронной сетью с дискретным временем: она применяется неоднократно, и часть выходных данных, вычисленных за один шаг, возвращается на следующий шаг. Кодеры размещают свои слои в определенных структурах стробирования, которые наделены определенной способностью учиться забывать прошлые входные данные, которые не имеют отношения к определенной точке, или запоминать прошлые входные данные. Наиболее часто используемые конфигурации стробирования - это долговременные воспоминания (LSTM: Hochreiter and Schmidhuber 1997) и рекуррентные стробированные блоки (GRU: Cho et al. 2014).

Теперь рассмотрим работу декодера.

1. Начиная с кодирования всего предложения E («Я люблю осень.»), Декодер создает два вектора: один - это начальное состояние декодера D («Я люблю осень», «»), где «» представляет пустая последовательность целевых слов и вектор вероятностей для всех возможных слов x в первой позиции целевого предложения.

2. Декодер читает D («Я люблю осень», «») и слово «Мен» и выдает два вектора: следующее состояние декодера D («Я люблю осень», «Мен») и вектор вероятности всех возможных выходных слов x во второй позиции предложения, p (x | «Я люблю осень», «Мен»). и т.д. Пока не достигнет конца предложения.

Список литературы

1. *Mikel L.F*. Making sense of neural machine translation // Translation Spaces 6:2 (2017) 291–309, DOI 10.1075/ts.6.2.06for.
2. An open source machine learning framework for everyone. [Электронный ресурс]. Режим доступа: https://www.tensorflow.org/(дата обращения: 07.02.2019).
3. *He W., Wu Hua., Wang H.* Improved neural machine translation with SMT features // Thirtieth AAAI conference on artificial intelligence, 2016.
4. *Luong M., Manning C.* Achieving open vocabulary neural machine translation with hybrid word-character models, 2016. arXiv preprint.
5. *Sennrich R., Haddow R.* Improving neural machine translation models with monolingual data // 54th annual meeting of the association for computational linguistics. Berlin, 2016. Pp. 86-96.