Новосибирский государственный университет

Обработка естественного смешанного языка

Выполнил: студент гр. 19123
Александр Бочкарев

Новосибирск 2022

Оглавление

Введение

Обработка естественного языка (*Natural Language Processing, NLP*) – область исследования искусственного интеллекта, основные задачи в которой имеют очень естественную постановку: распознавание смыслов речи и текста на естественном языке, а также генерация грамотного текста, передающего нужные смыслы. Сейчас для исследований в этой области применяются нейронные сети со сложнейшими архитектурами и огромным числом параметров.

В научной среде на данный момент сформирован консенсус относительно того, что без качественного скачка в технологии нейронных сетей не стоит ожидать столь стремительного прогресса в решении задач обработки языка, каким он был в последние 20 лет: количественный прирост обучающих данных и настраиваемых параметров не приведет к принципиально новым результатам. Так, GPT-3 (алгоритм третьего поколения обработки естественного языка от OpenAI) превосходящий своего предшественника, GPT-2 более чем в 100 раз по количеству параметров, и более чем в 10 по объему датасета, не дал качественно новых результатов [1].

Вот что думает французский ученый, один из основателей глубокого машинного обучения (*Deep Learning, DL*) Ян Лекун относительно завышенных ожиданий от GPT-3 и вообще масштабирования нейросетей:

> *"...trying to build intelligent machines by scaling up language models is like building a high-altitude airplanes to go to the moon. You might beat altitude records, but going to the moon will require a completely different approach."*

Можно сказать, что решение задачи с такой широкой постановкой уже столкнулось с так называемой «проблемой хрупкости искусственного интеллекта» [2], и дальнейший прирост эффективности алгоритмов требует фундаментальных исследований в этой области. В то время, как более узконаправленные задачи, рассматривающие частные случаи естественного языка, еще не настолько подробно изучены в актуальных исследованиях.

Мотивация

В мире сотни миллионов [3] людей испольют смешанные языком в том или ином виде. На данный момент не существует всеобъемлющей классификации таких языков, но к ним, например, относятся креольские языки, возникшие в период колониализма в условиях двуязычной среды. Креолы являются родными для носителей.

Другой пример смешанного языка – пиджин (контактивный язык). Часто это третий язык, который используется для вынужденного общения между двумя (и более) этническими группами, говорящими на взаимно непонятных языках. Пиджин представляет собой упрощённое средство коммуникации и, в отличие от креольского языка, не является родным для носителей. Понятно, что многие случаи использования смешанного языка находятся где-то между этими категориями, или вообще вне существующей классификации.

Так, например, в Казахстане государственным языком признан казахский, однако частота его использования в общении сопоставима с русским языком [4]. Таким образом, в условиях широко распространенного двуязычия в Казахстане существует некоторый промежуточный диалект. И если существенная часть населения общается, используя слова и конструкции сразу из обоих языков (причем их соотношение может существенно варьироваться), ясно, что для исследователя задача интерпретации такого диалекта представляет особый интерес.

Кроме того, что существование хороших предсказательных моделей для смешанных языков ценно само по себе, исследования в этой области представляют интерес для разработчиков мультиязычных нейросетей. Качественная модель для смешанных языков очевидным образом может быть полезна при создании алгоритмов, претендующих на итерпретацию текстов без привязки к конкретному языку.

Специфика задачи

Один из важнейших этапов создания языковой модели – сбор и предобработка данных. Спецификой исследований смешанных языков, очевидно, является то, что не всегда получится собрать достаточно большой по объему корпус текста, потому что часто на смешанных языках не пишутся статьи и книги, не ведется документооборот (особенно в случае с пиджином, который часто используется почти только при устном общении).

Если в качестве промежуточного этапа решения задачи происходит векторизация языка, то количество слов может существенно влиять на качество предсказания этой модели. Популярная в машинном обучении идея векторизации набора слов основывается на том, что в результате обучения некоторой модели на достаточно большом объеме данных можно построить такое отображение из множества слов обучающей выборки в многомерное векторное пространство, что схожие по семантике слова будут расположены близко к друг другу по заданной метрике в этом пространстве. Модель, делающая предсказания на основании расстояний между словами, вероятно, часто будет «находить» сразу несколько слов очень близких по значению, потому что для обозначения одного и того же понятия существуют слова в разных языках, образующих смешанный.

Другая проблема заключается в том, что если объектом исследования являются смешанный язык по типу того, что используется в Казахстане, его вариативность от региона к региону может сильно меняться. Но, в отличие от диалектов более стабильных языков (типа английского, испанского, русского), будет меняться еще и доля использования слов из разных языков, что делает набор и без того непросто агрегируемых данных еще менее однородным.

Подходы к решению

Обученные мультиязычные модели

Сейчас широко изучаются мультиязычные нейронные сети, которые извлекают смыслы из текста независимо от того, на каком языке он написан. Существует также множество исследований на тему перевода с одного языка на другой, такой «модуль перевода» как раз может быть частью мультиязычной нейросети. Но это все отличается от задачи интерпретации текстов на смешанных языках.

Если мы хотим извлекать смыслы из текстов на, например, креольских языках, где в простейшем случае будут встречаться слова из разных языков уже в пределах одного предложения, то уже обученная мультиязычная нейронная сеть (пусть она обучалась на каждом языке, образующем креольский, в отдельности) по понятным причинам будет выдавать качество ниже того, котогое она показывается на задачах, под которые была обучена. Это связано, по крайней мере, с тем, что в обучающей выборке такой нейросети не было текстов, включающих достаточно предложений, внутри которых содержалось бы несколько языков.

Даже если в обученных мультиязычных нейросетях уже заложена информация о семантической близости отдельных слов из разных языков, это еще не обеспечивает их применимости к смешанным языкам. В процессе образования креолов могут создаваться новые синтаксические конструкции [5], совершенно нетипичные для языков, на которые заточена мультиязычная модель.

Увеличение объема данных за счет образующих языков

Естественным предположением кажется, что если данных на смешанных языках недостаточно, то их можно дополнить текстами на чистых языках, влияющих на смешанный. Такой метод использовали исследователи в недавней статье от сентября 2021 года [6]. Они в своем первом подходе использовали тексты из СМС-сообщений, журналов религиозных общин, расшифровки публичных выступлений и некоторые открытые датасеты. Обучение моделей велось в отдельности для каждого из трех языков: нигерийский пиджин-английский, сингапурский разговорный английский («синглиш») и гаитянский креольский. В каждой группе было около нескольких десятков тысяч текстов на смешанных языках. Они были разбиты в соотношении 95 к 5 (тренировочная и валидационная часть), также каждая тренировочная группа была дополнена текстами на чистых языках, образующих смешанные, в количестве примерно в 4.5 раза большем, чем было текстов в каждой группе изнчально. Причем соотношение таких языков соответствовало частоте встречаемости слов из них в текстах на смешанных.

Понятно, что при таком распредении данных, тексты, на которых будет осуществляться валидация, будут в меньшей степени представлены в обучающей части. Для решение этой проблемы уже существуют методы, однако даже при использовании DRO (*Distributionally Robust Optimization*), корректирующего обучение с учетом недопредаствленности важного класса данных, модель не показала удовлетворительных результатов.

Стандартный подход

В том же исследовании приводятся результаты обучения языковых моделей на тех же данных, но без добавления текстов на чистых образующих языках. Вопреки ожиданиям исследователей и несмотря на небольшой объем данных, по всем трем смешанным языкам результаты оказались лучше, чем в предыдущем эксперименте. Причем по всем метрикам:

1. Предсказние k слов, случайно убранных из предложения на смешанном языке - **P@k**
2. Предсказние k слов, которые принадлежат заранее подготовленному набору слов, характерных для конкретного смешанного языка, случайно убранных из предложения - $\mathbf{P}_D@K$
3. Взвешенная сумма логарифмов условных вероятностей обнаружения правильного слова на месте пропуска в предложении при известной оставшейся его части - **PLL**

В таблице ниже приведены оценки качества моделей тремя различными метриками в отдельности для каждого языка. Тут №1 - первый эксперимент (с добавлением в датасет текстов на чистых языках), №2 - второй эксперимент (использовались только смешанные языки).

№	нигерийский пиджин			сингапурский английский			гаитянский креольский		
	P@1	$\mathbf{P}_D@1$	**PLL**	**P@1**	$\mathbf{P}_D@1$	**PLL**	**P@1**	$\mathbf{P}_D@1$	**PLL**
1	63.83	59.97	42.41	46.77	42.89	41.06	68.09	43.35	55.04
2	73.72	71.38	28.14	53.80	52.26	34.22	73.15	55.50	55.51

Исследователи далее подробно показывают, что такие результаты были получены не из-за отстуствия регуляризации модели или других технических ошибок. Они заключают, что самый обыкновенный подход хорошо показал себя в данном исследовании, потому что выбранные ими языки (или по крайней мере отобранные ими данные) оказались досточно стабильными, не имеющие существенных «недостатков», присущих некоторым смешанным языкам, которые обсуждались в пункте о специфике задачи.

Заключение

Если грубо интерпретировать процесс обучения нейросети как выявление некоторых закономерностей в наборе данных, для предсказания на их основе некоторого целевого признака, то можно сказать, что в случае со смешанными языками выявление этих самых закономерностей осложняется несколькими факторами:

1. Данных просто мало, так как каждый смешанный язык в отдельности используют не более нескольких десятков миллионов людей
2. Сложность выявление закономерностей в данных из-за высокой вариативности диалектов
3. Отстутствие стандартов письменности, влекущее ухудшение однородности данных
4. Из-за взаимозаменяемости слов из языков, образующих смешанный, при фиксированном количестве смыслов возрастает количество слов в языковой модели

Несмотря на все названные сложности, по итогам некоторых исследований [6,7] выяснилось, что как минимум некоторые смешанные языки не так уж сильно отличаются от гораздо более широко используемых в мире языков, по крайней мере в плане построения для них языковой модели. Возможно, те языки, на которых стандартный подход показал бы себя хуже, недостаточно широко и долго используются, чтобы для проведения подобного исследования можно было бы собрать необходимое количество текстов на них.

В подтверждение того, что по крайней мере к наиболее стабильным смешанным языкам, возможно, не стоит подходить каким-то особым образом, можно привести распространенное в научной среде мнение российского лингвиста Олега Мудрака:

> *"Смешения языков никогда не бывает, нельзя говорить, что один язык появился из-за смешения двух других языков, такого в природе не отмечено <...> каждый язык — это отдельная система, которая сама по себе существует и развивается по своим собственным законам."*

Список литературы

1. Chuan Li. OpenAI's GPT-3 Language Model: A Technical Overview - lambdalabs.com/blog/demystifying-gpt-3

2. Yejin Choi. Trying to give AI some common sense - www.axios.com/2018/03/15/the-quest-to-give-ai-some-common-sense-1521085175 (*and others her interview about that*)

3. Matras, Yaron and Peter Bakker, eds. (2003) The Mixed Language Debate: Theoretical and Empirical Advances, Berlin: Walter de Gruyter. ISBN 3-11-017776-5

4. Национальный состав, владение языками, гражданство, вероисповедание // Аналитический отчет. «Итоги Национальной переписи населения Республики Казахстан 2009 года» : [рус.] / Под ред. А. А. Смаилова. — Астана : Агентство Республики Казахстан по статистике, 2011. — С. 20—24. — 65 с.

5. Bickerton, Derek (1983), «Creole Languages», Scientific American 249 (8): 116—122, doi:10.1038/scientificamerican 0783-116

6. Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, Anders Søgaard, On Language Models for Creoles, 2021

7. Mohammad Mahdi Jaghoori, Tom Chothia. Timed Automata Semantics for Analyzing Creol, DOI 10.4204/EPTCS.30.8