Методы семплирования для генерации естественного текста нейросетевой языковой моделью

Думбай Алексей Дмитриевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dumbay.aleksey.97@gmail.com

Научный руководитель *— Дьяконов Александр Геннадьевич*

Одной из задач обработки естественного языка является генерация текстов, приближенных к человеческим. Для этого используется контекст – начало некоторого текста, для которого необходимо построить продолжение. Сложность данной задачи заключается как в выборе методов генерации новых слов, так и в формулировке критериев для определения нужных текстов. Современные методы опираются на стратегии декодирования некоторой языковой модели, основанной на вероятностях появления слов. Максимизация правдоподобия генерируемого предложения приводит к проблемам повторения и неестественности текста, что приводит к необходимости использования методов семплирования. Основными методами семплирования являются *Top-k* [1] и *Top-p* [2], выбирающие слова из наиболее вероятных.

В данной работе рассматриваются методы семплирования, опирающиеся на фиксированную модель для предсказания вероятности слов и дополнение ее некоторой базовой моделью, а так же предлагаются формальные способы оценки качества сгенерированых текстов.

В общем случае генерация начинается с контекста – некоторого предварительно заданного текста. Модель предлагает вероятности кандидатов, после чего из них выбирается следующее слово. Для качественной генерации необходимо, зачастую, запускать параллельно несколько цепочек слов, выбирая наилучшую по правдоподобию.

В работе рассматривается идея добавления семантической информации в процесс декодирования. Для этого можно использовать более простую модель, которая будет предсказывать не сами слова, а части речи. С помощью полученных таким образом последовательности можно проводить семплирование из языковой модели только из слов, относящихся к данной части речи.

При прямом процессе генерации мы последовательно получаем слова, что не дает нам использовать всю информацию о полученном предложении. Вторая рассмотренная методика выбора опирается на

удаление некоторых слов и попытке заменить их с помощью другой языковой модели. Данный процесс может повторяться некоторое число итераций.

Для верификации результатов в работе рассматриваются как стандартные методы перплексии [3] текста и ассесорской оценки, так и предлагается другой метод, основанный на применении некоторой базовой модели классификации для задачи классификации текста на сгенерированый и естественный. Данная модель должна быть достаточно простой, что бы сохранять некоторые теоретические гарантии. В работе предлагаются методы классификации на основе tf-idf, такие как логистическая регрессия и решающий лес [4]. Важным моментом является то, что контекст для генерации текстов должен быть взят из набора данных, используемого для проверки. В ином же случае качество классификации не может служить разумной метрикой, так как ключевыми факторами для отличия текстов могут стать различные тематики, стиль и др.

Литература

1. Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pp. 889–898, 2018.
2. Holtzman A. et al. The curious case of neural text degeneration //arXiv preprint arXiv:1904.09751. – 2019.
3. Ramaciotti Morales, Pedro; et al. (September 2019). "Role of the Website Structure in the Diversity of Browsing Behaviors". Proceedings of the 30th ACM Conference on Hypertext and Social Media: 133–142. Retrieved 2020-02-13.
4. Bishop C. M. Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.