Название статьи: Vocabulary Learning via Optimal Transport for Neural Machine Translation
Авторы статьи: Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, Lei Li
Автор исследования: Шапкин Антон

Первая версия данной работы была опубликована в сентябре 2020 года. Авторы планировали выступить на конференции ICLR 2021, однако заявка была отозвана. В начале августа 2021 года исследователи успешно выступили на онлайн конференции, организованной ACL (Association for Computational Linguistics) и AFNLP (Asian Federation of Natural Language Processing), где их статья была отмечена как лучшая. Ответственными за продвижение статьи (corresponding authors) были Lei Li и Hao Zhou.

Над статьей работали Jingjing Xu, Zaixiang Zheng, Lei Li, Hao Zhou и Chun Gan, работающие в ByteDance AI Lab (ByteDance - интернет-компания, владеющая рядом популярных сервисов, включая TikTok). Кроме того, Zaixiang Zheng аспирант университета Нанджунга, а Lei Li - преподаватель Калифорнийского университет в Санта-Барбаре. Основываясь на тексте статьи и истории публикаций авторов, я предполагаю, что идея данной статьи выглядит как случайная находка, нежели чем прямое улучшение их предыдущей работы. Каждый из авторов имеет большое количество публикаций связанных с NLP и не только, однако про построение словаря у каждого из них лишь эта работа.

Данное предположение подтверждают и статьи, которые оказывают наибольшее влияние на исследуемую работу. Одной из таких публикаций является работа [1], вдохновившись которой авторы изобрели метрику MUV (Marginal Utility of Vocabularization). Как и предельная полезность, мера MUV оценивает общую полезность (энтропию) при потреблении дополнительных единиц блага (новых токенов в словаре). Статьи про эффективное решение транспортной задачи [2] и обобщение метода Синхорна [3] сильно влияют на работу, ведь именно благодаря им возможно решение исходной задачи дискретной оптимизации за приемлемое время. Работы, описывающие разные способы построения словаря также важны, однако они не играют ключевую роль, ведь предложенный авторами алгоритм не зависит от способа генерации токенов.

В силу того, что исследуемая статья была опубликована недавно, данная публикация цитируется лишь в 2 работах. Первая работа [4] изучает чувствительность моделей к пертурбациям в порядке слов и символов. В статье предлагается 2 меры, которые оценивают локальные и глобальные перестановки. Авторы обнаружили, что для ряда моделей локальные перестановки играют большую роль, чем глобальные. Отметив, что текущие исследования в области построения словаря пытаются найти альтернативы и улучшения BPE, они предложили свой подход: использовать словари меньшего размера с большой детализацией без потери качества, ведь

локальные структуры могут помочь улучшить показатели метрик в некоторых задачах. Авторы статьи [5] отмечают, что на данный момент токенизация является отдельным шагом, а современные методы подходят не для каждого языка. В публикации предлагается нейронная сеть кодировщик CANINE, которая работает с последовательностями символов без предварительной токенизации и построения словаря. Работа [6] от сотрудников Bytedance AI Lab также ссылается на исследуемую статью, так как они используют VOLT в предобработке данных. В публикации исследователи предлагают систему для параллельного перевода текстов. В статье [7], которую Google Scholar помечает как похожую, авторы оценивают количество словесных токенов в разных языках, необходимых для того, чтобы энтропия сошлась к стабильному значению.

Оценив, что происходит вокруг в области, можно сделать вывод, что авторы проделали большую работу. Они оценили VOLT на разных данных в задаче машинного перевода и показали преимущество предложенного метода, однако хотелось бы увидеть исследования и в других задачах NLP. Как отмечалось в работе [5], этап генерации словаря и токенизации на данный момент является неотъемлемой частью любой задачи обработки естественного языка, поэтому хотелось бы узнать, коррелирует ли MUV с метриками других задач и можно ли таким образом подбирать оптимальный словарь. Уверен, предложенный метод будет широко применяться в индустриальных приложениях, ведь он позволяет сократить нагрузку на вычислительные ресурсы, а также может сохранить большое количество времени, которое могло бы уйти на подбор гиперпараметра размера словаря.

Статьи:
[1] Paul Samuelson - «A Note on Measurement of Utility» - 1937.
[2] Marco Cuturi - «Sinkhorn distances: Lightspeed Computation of Optimal Transport» - 2013.
[3] Gabriel Peyre, Marco Cuturi - «Computational Optimal Transport» - 2019.
[4] Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, Sarath Chandar - «Demystifying Neural Language Models' Insensitivity to Word-Order» - 2021.
[5] Jonathan H. Clark, Dan Garrette, Iulia Turc, John Wieting - «CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation» - 2021.
[6] Lihua Qian, Yi Zhou, Zaixiang Zheng, Yaoming Zhu, Zehui Lin, Jiangtao Feng, Shanbo Cheng, Lei Li, Mingxuan Wang, Hao Zhou - «The Volctrans GLAT System: Non-autoregressive Translation Meets WMT21» - 2021.
[7] Christian Bentz, Dimitrios Alikaniotis - «The Word Entropy of Natural Languages» - 2016.