Исследование статьи **Vocabulary Learning via Optimal Transport**

(Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, Lei Li)

Автор исследования: **Седашов Данила**

Авторы данной статьи — исследователи из ByteDance AI Lab (компания, владеющая TikTok), Висконского университета в Мадисоне и Наньцзинского университета. Все авторы работают преимущественно над задачами NLP. Из интересного — один из авторов работает исключительно над задачами машинного перевода (названия почти всех его статей из google scholar содержат слова “machine translation”). Также любопытно, что тёзка одного из авторов (Hao Zhou) — экономист, особенно учитывая, что в статье предлагается метрика MUV — аналог marginal utility из экономики.

Данная статья была опубликована летом 2021 на конференции ACL (насколько я понял, в формате 2-3х минутного видеоролика). Интересно, что с первой версией этой статьи те же авторы подавались на ICLR 2021 и получили там reject. Предлагаемый метод тогда назывался Info-VOT (а не VOLT), а статья была довольно сырой:

- Не было кодов эксперимента
- Мало экспериментов (маленькие датасеты, сложно назвать результаты значимыми)
- Плохо обоснована важность метрики MUV, в целом слабая математическая часть
- Плохой обзор аналогов

В итоге 31 декабря 2020 года авторы сняли свой сабмишн, однако на ACL сумели завоевать награду за лучшую статью конференции.

Как мне кажется, статья зиждется на двух идеях:

1) Метрика MUV, согласно которой подбирается словарь
2) Эффективное решение задачи дискретной оптимизации, помогающее найти MUV

Таким образом, на мой взгляд, можно выделить 4 наиболее повлиявших статьи:

1) Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport — про то как быстро решать задачу оптимального транспорта
2) Christian Bentz and Dimitrios Alikaniotis. 2016. The word entropy of natural languages — почему высокая энтропия в корпусе выгодна для обучения модели
3) Ben Allison, David Guthrie, and Louise Guthrie. 2006. Another look at the data sparsity problem — "больше данных — всегда лучше" в задачах NLP
4) Paul A Samuelson. 1937. A note on measurement of utility — идея с комбинированием метрик, одна из отвечает за что-то полезное, а вторая — за что-то вредящее

При этом у статьи есть “конкуренты”:

- Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2020. Optimizing segmentation granularity for neural machine transation
- Colin Cherry, George F. Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression.

Однако вся предшествующая работа относилась к оптимальному подбору гранулярности деления слов на подслова, а размер оптимизация размера словаря не рассматривалась. Таким образом, статья обладает новизной: авторы исследовали область, которой раньше никто не занимался, и показали, что это тоже важно. Возможно, это одна из причин того, что статья стала лучшей на ACL 2021.

В статье была рассмотрена оптимизация размера словаря применительно к задаче машинного перевода. По моему мнению, хорошей идеей будет рассмотреть метод VOLT применимо к другим задачам NLP — генерирование текста, классификация и проч.

При этом, учитывая, что предлагается метод, позволяющий быстрее и более экономно по памяти обучать модели машинного перевода, можно использовать его в индустрии, например, при проверке гипотез о моделях (сокращая время и расход ресурсов на обучение).