# Обучение словаря с помощью Optimal Transport

Роли:

- Докладчик: Булатова Катя
- Рецензент: Смирнов Тимофей
- Практик-исследователь: Седашов Даня
- Хакер: Пахалко Илья

# План презентации

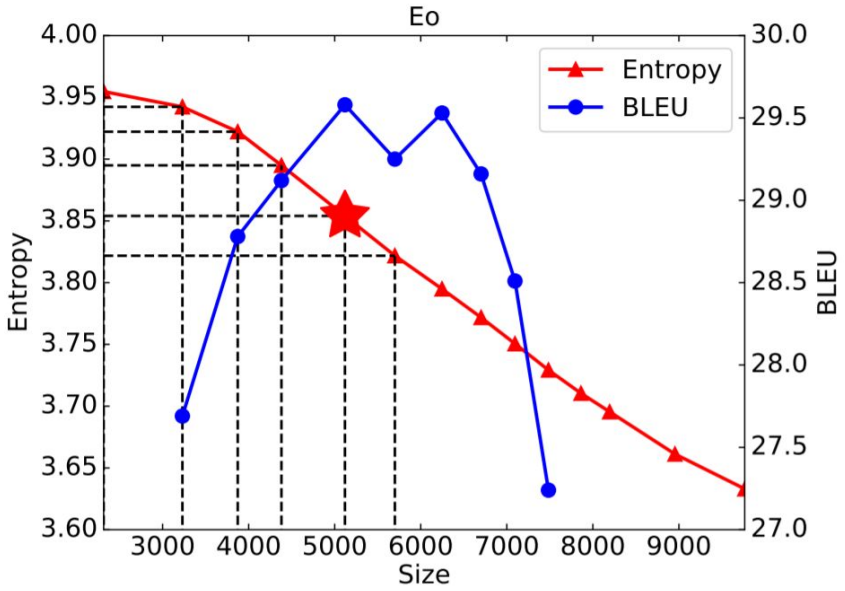
1. Докладчик
    - Постановка задачи
    - Общий план решения: VOLT
    - Детали предложенного метода
    - MUV
    - Перенос на нашу задачу
    - Уточнение формул
    - Результаты
2. Хакер
3. Рецензент
4. Практик-исследователь

# Постановка задачи

- Размер словаря важен
- Нельзя перебрать все размеры (используют стандартные)
- Хотим автоматический вычислительно недорогой способ поиска словаря
- Проблемы:
  - Энтропия ⇅ размер словаря
  - Слишком много токенов ⇒ разреженность данных, много параметров у моделей
  - Задача скорее всего будет экспоненциально сложна

# Предложенный метод: VOLT

- VOcabulary Learning approach via optimal Transport
- Измерение качества -- MUV
- Рассматривает конечное количество размеров словаря
- Учитывает и энтропию, и размер словаря
- Работает за полиномиальное время
- Можно рассматривать как задачу OT



# MUV

- Предельная полезность, или Marginal Utility (of Vocabularization)
- Оценивает выгоду (энтропию) ~ увеличения цены (размер словаря)
- Оценка производной

$$\mathcal{M}_{v(k+m)} = \frac{-(\mathcal{H}_{v(k+m)} - \mathcal{H}_{v(k)})}{m}$$

$v(k)$ -- словарь размера *k*

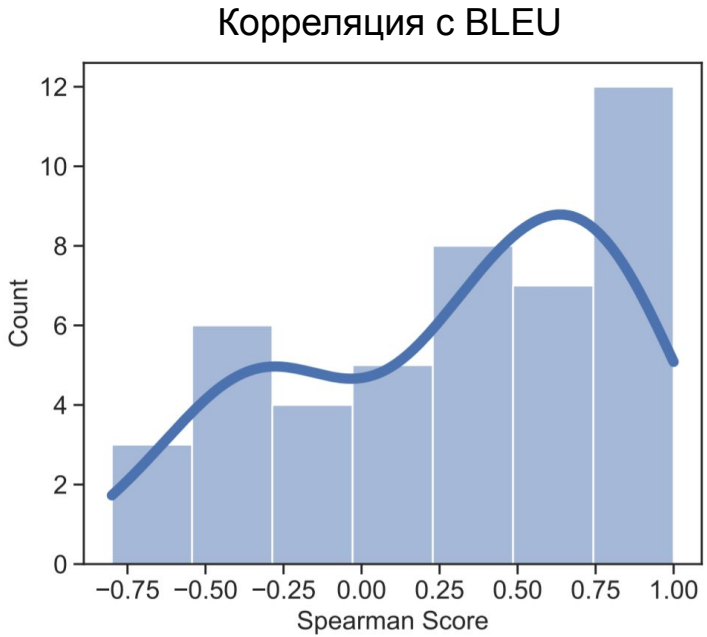
$\mathcal{H}_v$ -- энтропия корпуса *v*: $\mathcal{H}_v = -\frac{1}{l_v} \sum_{j \in v} P(j) \log P(j)$

$l_v$ -- средняя длина токенов в *v*

$P(j)$ -- частота токена j в обучающем корпусе

# MUV

- Предельная полезность, или Marginal Utility (of Vocabularization)
- Оценивает выгоду (энтропию) ~ увеличения цены (размер словаря)
- Оценка производной



# VOLT

- $\boldsymbol{S} = \{k, 2 \cdot k, ..., (t-1) \cdot k, \cdots\}$-- набор испытуемых размеров, т.е. значения, для которых вычисляется MUV
- Функционал -- значение MUV:

$$\arg\max_{t} \quad \arg\max_{v(t-1)\in\mathbb{V}_{\boldsymbol{S}[t-1]}, v(t)\in\mathbb{V}_{\boldsymbol{S}[t]}} \quad -\frac{1}{k}\left[\mathcal{H}_{v(t)} - \mathcal{H}_{v(t-1)}\right]$$

*v*(t) имеет размер, не превышающий S[t]: $\mathbb{V}_{\boldsymbol{S}[t]}$ содержит все словари, размер которых <= S[t]

- Появятся запрещенные пары: размер *v*(t-1) > размера *v*(t). Их можно не учитывать

# VOLT: алгоритм

Нашу задачу можно переформулировать в виде (смотрим на верхнюю оценку):

$$\arg\max_t \frac{1}{k}\Big[\arg\max_{v(t)\in\mathbb{V}_{\boldsymbol{S}[t]}} \mathcal{H}_{v(t)} - \arg\max_{v(t-1)\in\mathbb{V}_{\boldsymbol{S}[t-1]}} \mathcal{H}_{v(t-1)}\Big]$$

то есть для каждого t нужно найти модель с максимальной энтропией:

$$\arg\max_{v(t)\in\mathbb{V}_{\boldsymbol{S}[t]}} -\frac{1}{l_{v(t)}}\sum_{j\in v(t)} P(j)\log P(j)$$

Сложно решить, будем оценивать верхнюю границу и решать задачу известным алгоритмом

# VOLT: переход к задаче OT

Рассмотрим $\mathbb{T} \in \mathbb{V}_{\boldsymbol{S}[t]}$, содержащий S[t] самых частых токенов, $\mathbb{C}$ -- буквы, тогда:

$$\min_{v \in \mathbb{V}_{\boldsymbol{S}[t]}} \frac{1}{l_v} \sum_{j \in v} P(j) \log P(j) \leq \frac{1}{l_{\mathbb{T}}} \sum_{j \in \mathbb{T}} P(j) \log P(j) \;=$$

$$= \frac{1}{l_{\mathbb{T}}} \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i) \log P(j,i)}_{\mathcal{L}_1} + \frac{1}{l_{\mathbb{T}}} \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i)(-\log P(i|j))}_{\mathcal{L}_2}$$

Здесь L1 -- отрицательная энтропия распределения *P(j, i)*, обозначим *-H(P)*;

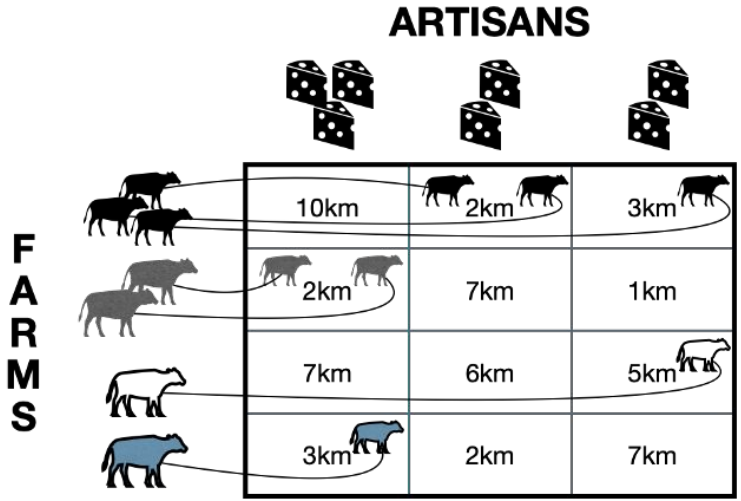
L2 можно переписать через матричное умножение <*P, D*>, где *D* -- матрица размера $|\mathbb{C}| \times |\mathbb{T}|$ с $D[i, j] = -\log P(i \mid j)$.

Тогда нашу задачу можно переписать как оптимизацию:

$$\min_{\boldsymbol{P} \in \mathbb{R}^{m \times n}} \langle \boldsymbol{P}, \boldsymbol{D} \rangle - \gamma H(\boldsymbol{P})$$

# VOLT: Optimal Transport

Задача OT: как “перекинуть” одно распределение в другое, учитывая веса, нужные для “перемещения” элементов.




В нашем случае, одним из распределений будут символы, а вторым -- токены; P -- transport matrix, а D -- distance matrix.

$$\min_{\boldsymbol{P} \in \mathbb{R}^{m \times n}} \langle \boldsymbol{P}, \boldsymbol{D} \rangle - \gamma H(\boldsymbol{P})$$

# VOLT: финальная формулировка

$$\min_{\boldsymbol{P}\in\mathbb{R}^{m\times n}} \langle \boldsymbol{P}, \boldsymbol{D}\rangle - \gamma H(\boldsymbol{P}), \quad \boldsymbol{D}(j,i) = \begin{cases} -\log P(i|j) = +\infty, & \text{if } i \notin j \\ -\log P(i|j) = -\log \frac{1}{len(j)}, & \text{otherwise} \end{cases}$$

Transport Matrix P

| | | | |
|---|---|---|---|
| $P_a$ | $P_{a,ab}$ | $P_{a,ab}$ | $P_{a,a}$ |
| $P_b$ | $P_{b,ab}$ | $P_{b,bc}$ | $P_{b,a}$ |
| $P_c$ | $P_{c,ab}$ | $P_{c,bc}$ | $P_{c,a}$ |
| | $P_{ab}$ | $P_{bc}$ | $P_a$ |

Distance Matrix D

| | | | |
|---|---|---|---|
| $a$ | 1/2 | ∞ | 1/1 |
| $b$ | 1/2 | 1/2 | ∞ |
| $c$ | ∞ | 1/2 | ∞ |
| | $ab$ | $bc$ | $a$ |

Constraints

$$\forall j \in \{a,b,c\}, \sum_{i\in\{ab,bc,a\}} P_{i,j} = P_j$$

$$\forall i \in \{ab,bc,a\}, \sum_{j\in\{a,b,c\}} P_{i,j} - P_i \le \epsilon$$

Problem

$$\min_{\text{all } P} C(P)$$

Cost Function

$$C(P) = -H(P) + \sum_{\substack{j\in\{a,b,c\},\\ i\in\{ab,bc,a\}}} P_{i,j} D_{i,j}$$

# Результаты

Двуязычные модели: меньше памяти, выше BLEU

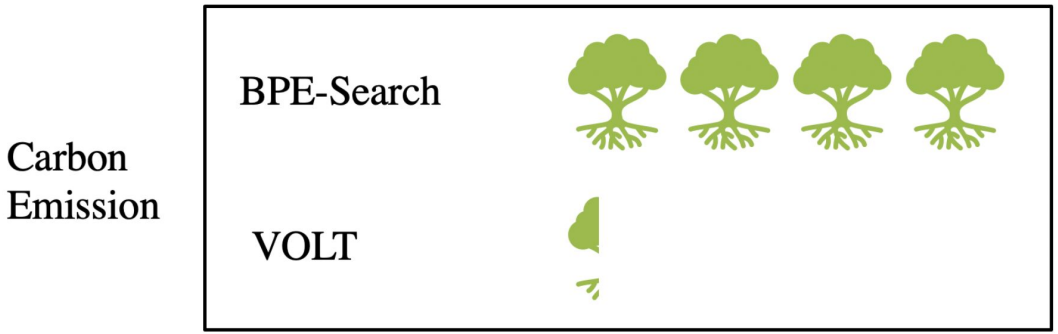
| Bilingual | WMT-14 | TED | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| En-X | De | Es | PTbr | Fr | Ru | He | Ar | It | Nl | Ro | Tr | De | Vi |
| BPE-30K | 29.31 | 39.57 | 39.95 | 40.11 | 19.79 | 26.52 | 16.27 | 34.61 | 32.48 | 27.65 | 15.15 | 29.37 | 28.20 |
| **VOLT** | **29.80** | **39.97** | **40.47** | **40.42** | **20.36** | **27.98** | **16.96** | **34.64** | **32.59** | **28.08** | **16.17** | **29.98** | **28.52** |
| X-En | De | Es | PTbr | Fr | Ru | He | Ar | It | Nl | Ro | Tr | De | Vi |
| BPE-30K | **32.60** | **42.59** | 45.12 | **40.72** | 24.95 | 37.49 | 31.45 | 38.79 | 37.01 | 35.60 | 25.70 | 36.36 | 27.48 |
| **VOLT** | 32.30 | 42.34 | **45.93** | **40.72** | **25.33** | **38.70** | **32.97** | **39.09** | **37.31** | **36.53** | **26.75** | **36.68** | **27.39** |
| Vocab Size (K) | De | Es | PTbr | Fr | Ru | He | Ar | It | Nl | Ro | Tr | De | Vi |
| BPE-30K | 33.6 | 29.9 | 29.8 | 29.8 | 30.1 | 30.0 | 30.3 | 33.5 | 29.8 | 29.8 | 29.9 | 30.0 | 29.9 |
| **VOLT** | **11.6** | **5.3** | **5.2** | **9.2** | **3.3** | **7.3** | **9.4** | **3.2** | **2.4** | **3.2** | **7.2** | **8.2** | **8.4** |

VOLT не требует серий экспериментов, быстрый. Показывает хорошие результаты на multilingual задачах. Проведено сравнение с MUV-Search.

| En-De | BLEU | Size | Cost |
|---|---|---|---|
| BPE-Search | **29.9** | 12.6K | 384 GH |
| MUV-Search | 29.7 | **9.70K** | 5.4 CH + 30 GH |
| **VOLT** | 29.8 | 11.6K | **0.5 CH + 30 GH** |

# Ссылки

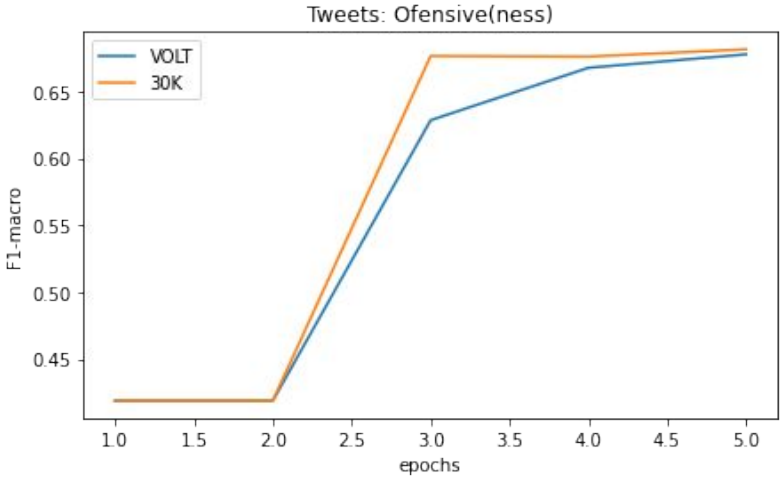
- https://arxiv.org/pdf/2012.15671.pdf
- https://www.idiap.ch/webarchives/sites/www.idiap.ch/talk/otml/
- https://indico.cern.ch/event/845380/attachments/1915103/3241592/Dvurechenskiy_lectures.pdf
- https://aclanthology.org/W17-3204.pdf
- https://towardsdatascience.com/optimal-transport-a-hidden-gem-that-empowers-todays-machine-learning-2609bbf67e59

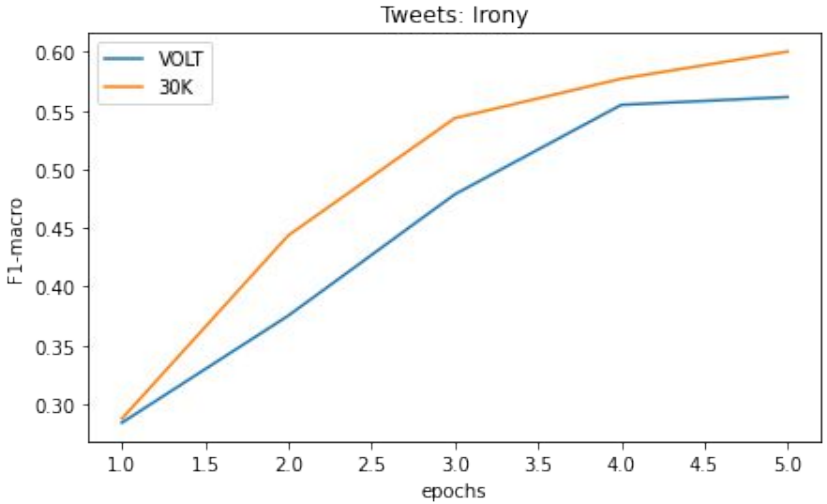


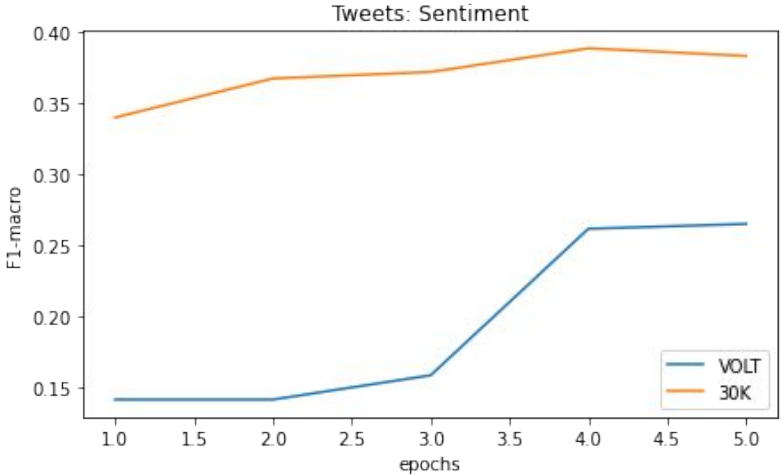
# Два утверждения из раздела “Результаты”

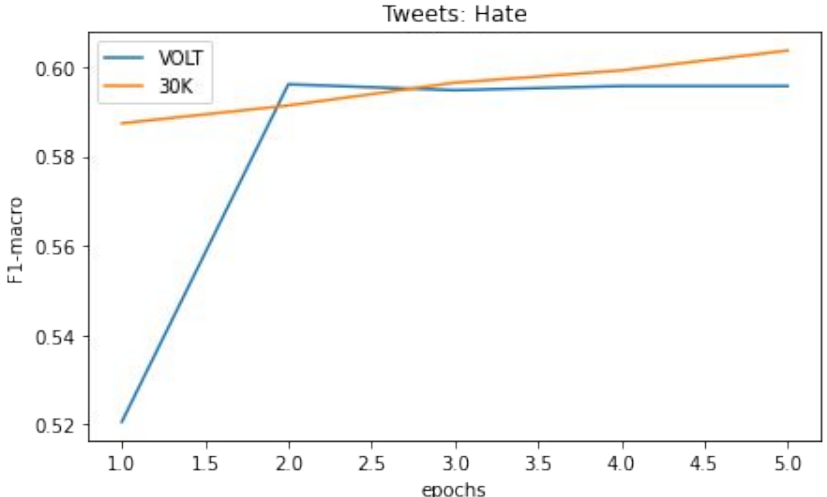
- “A Simple Baseline with a VOLT-generated Vocabulary Reaches SOTA Results”

- “VOLT Vocabularies and BPE Vocabularies are Highly Overlapped <...> They also have similar downstream performance”


Tweets: Ofensive(ness)
VOLT
30K
F1-macro
epochs


Tweets: Irony
VOLT
30K
F1-macro
epochs


Tweets: Sentiment
VOLT
30K
F1-macro
epochs


Tweets: Hate
VOLT
30K
F1-macro
epochs

| | VOLT | 30K |
|---|---|---|
| **hate** | 0.5961 | 0.6036 |
| **sentiment** | 0.2649 | 0.3889 |
| **offensive** | 0.6776 | 0.6813 |
| **irony** | 0.5616 | 0.6004 |
| **imdb** | 0.8687 | 0.8658 |


#hacked

# Плюсы

- Актуальная проблема подбора размера словаря для задачи машинного перевода
- Предложен способ и метрика для оптимизации размера словаря
- Уникальная работа, до этого никто не пробовал придумывать метрику описывающую оптимальность размера словаря для данной задачи

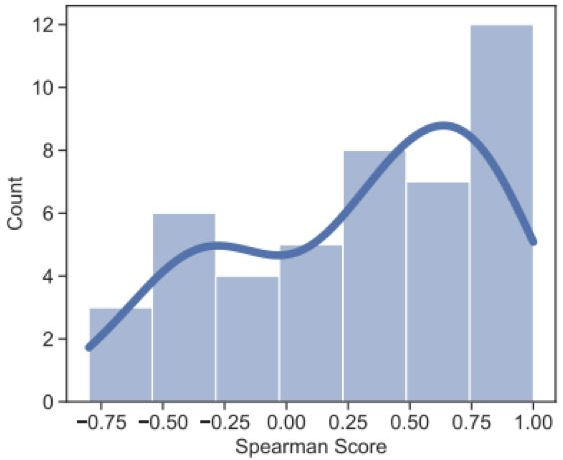
# Минусы

- Нет обоснования почему MUV хорошая метрика
- Кажется нечестным выбор BPE-30k и BPE-1k в качестве бейзлайна
- Нечестное сравнение VOLT с MUV-Search

| En-De | BLEU | Size | Cost |
|---|---|---|---|
| BPE-Search | **29.9** | 12.6K | 384 GH |
| MUV-Search | 29.7 | **9.70K** | 5.4 CH + 30 GH |
| **VOLT** | 29.8 | 11.6K | **0.5 CH + 30 GH** |

# Минусы от ревьюверов

- Необоснованность зависимости качества от MUV
- Все датасеты довольно маленькие(кроме WMT14 En-De)
- Оптимальные параметры BPE для данных датасетов 0.5-2к, а сетка перебора 1-30к, все бенчмарки 1к, 30к и 60к
- Все приросты в 3-м знаке BLUE




Корреляция MUV с качеством. Медиана 0.4

# Минусы от ревьюверов

## Поправленные

- Не рассмотрена предшествующая работа по теме
- Непонятна постановка экспериментов, глубина сети, выбранная архитектура и тд
- Воспроизводимость
- Изначально неправильно были подсвечены лучшие эксперименты
- Много опечаток и неточностей в математической части

# Итоги

- Показали что подбор размера словаря можно делать не только через trial-search, до этого это никто не делал
- Теоретически метод может выдавать неплохое значение размера словаря по умолчанию и сэкономить время на подборе параметров
- VOLT в некоторых случаях лучше конструирует сам словарь и позволяет достичь прирост качества

# Кто авторы?

- ByteDance AI Lab, Висконский университет в Мадисоне и Наньцинский университет
- Преимущественно занимаются исследованиями в NLP
- Один из авторов занимается только задачами NMT
- Тёзка одного из авторов — профессор экономики

# История публикаций

- Submission на ICLR 2021

  Сырая статья (мало экспериментов, плохие обоснования и пр.)
  Получили отказ от 4/4 ревьюеров на openreview
  31 декабря 2020 отозвали сабмишн

# История публикаций

- Submission на ICLR 2021

  Сырая статья (мало экспериментов, плохие обоснования и пр.)
  Получили отказ от 4/4 ревьюеров на openreview
  31 декабря 2020 отозвали сабмишн

- ACL 2021 (дедлайн подачи заявок — 1 февраля)

  Best paper (а ещё поменяли название метода с Info-VOT на VOLT)

# Повлиявшие статьи

Идея метрики MUV:

- Christian Bentz and Dimitrios Alikaniotis. 2016. The word entropy of natural languages
- Ben Allison, David Guthrie, and Louise Guthrie. 2006. Another look at the data sparsity problem
- Paul A Samuelson. 1937. A note on measurement of utility

Как быстро решать задачу оптимального транспорта:

- Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport

# “Конкуренты” статьи

Вся предшествующая работа относилась к оптимальному подбору гранулярности деления слов на подслова:

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

Таким образом, статья обладает новизной: авторы исследовали область, которой раньше никто не занимался, и показали, что это тоже важно!

# Дальнейшие исследования

- В статье рассмотрена оптимизация размера словаря применимо к задаче NMT => можно рассмотреть эффективность метода применимо к другим задачам NLP