# Вероятностные иерархические векторные представления слов

Петр Алексеевич Остроухов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

6 декабря 2018 г.

# Цель исследования

Построение иерархических тематических векторных представлений слов (эмбеддингов слов) для нескольких модальностей, где вместо матрицы [*слова*; *документы*] используется матрица [*слова*; *псевдодокументы*] (в качестве псевдокументов используются контексты слов некоторой длины).

## Задачи:

- Построение мультимодальной иерархической тематической модели над корпусом Википедии.
- Взяв в качестве эмбеддингов слов строки полученной матрицы [*слова*; *темы*], проверить качество их работы на задачах поиска синонимов, категоризации, кросс-язычного поиска и сравнить с нейросетевыми моделями на задаче семантической близости.

# План

- Тематическое моделирование
  - Дано Найти Критерий
  - ARTM
  - EM-алгоритм
  - Модальности
  - Иерархии
- Существующие подходы
- Преимущества вероятностных эмбеддингов
- Применимость к задачам
  - Категоризация
  - Синонимия
  - Схожесть документов
  - Кросс-язычный поиск
- Новизна

# Тематическое моделирование (ТМ)

- $W$ — конечное множество слов
- $D_p$ — конечное множество псевдодокументов
- $T$ — конечное множество тем (скрыто)
- $D_p \times W \times T$ — дискретная генеральная совокупность
- гипотеза условной независимости: $p(w|t,d) = p(w|t)$

Вероятностная тематическая модель порождения текста:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

# ТМ [ДНК]

- **Дано**: $n_{dw}$ — частоты слов в документах $\Rightarrow \hat{p}(w|d) = \frac{n_{dw}}{n_d}$
- **Найти**:
  - $\phi_{wt} = p(w|t)$ — вероятности слов в темах
  - $\theta_{td} = p(t|d)$ — вероятности тем в документах
- **Критерий** максимизации правдоподобия:

$$\mathcal{L}(\Phi;\Theta) = \ln p\big((d,w)_{d\in D_p, w\in W};\Phi,\Theta\big) =$$
$$= \ln \prod_{w\in W}\prod_{d\in D_p} \big(p(w|d)p(d)\big)^{n_{dw}} =$$
$$= \sum_{w\in W}\sum_{d\in D_p} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi;\Theta}$$

$$\sum_{w\in W}\phi_{wt} = 1,\ \phi_{wt} \geq 0,\ \sum_{t\in T}\theta_{td} = 1,\ \theta_{td} \geq 0.$$

# TM [ARTM]

## ММП + ARTM:

$$\sum_{w \in W} \sum_{d \in D_p} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta};$$

$$R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$$

$$\sum_{w \in W} \phi_{wt} = \{0, 1\}, \ \phi_{wt} \geq 0, \ \sum_{t \in T} \theta_{td} = \{0, 1\}, \ \theta_{td} \geq 0.$$

# ТМ [EM-алгоритм]

- **E-шаг**: условные вероятности $p(t|d,w)$ вычисляются по формуле Байеса из $\phi_{wt}$ и $\theta_{td}$:

$$p(t|d,w) = \frac{p(w|t,d)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}$$

- **M-шаг**: частотные оценки матриц $\Phi$ и $\Theta$ вычисляются на основе счетчика $n_{tdw} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad n_{wt} = \sum_{d\in D_p} n_{tdw}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \quad n_{td} = \sum_{w\in d} n_{tdw}, \qquad n_d = \sum_{t\in T} n_{td}.$$

# ТМ [модальности]

Пусть $M$ — множество модальностей (в нашем случае — тексты на разных языках, категории). Тогда
$\forall m \in M\ \exists W_m:\ \forall i \neq j\ W_i \bigcap W_j = \emptyset,\ \bigcup_{m \in M} W_m = W.$
Тематическая модель для каждой модальности остается без изменений.

## ММП для мультимодальной модели:

$$\sum_{m \in M} \tau_m \sum_{d \in D_p} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta},$$

$$\sum_{w \in W} \phi_{wt} = \{0, 1\},\ \phi_{wt} \geq 0,\ \sum_{t \in T} \theta_{td} = \{0, 1\},\ \theta_{td} \geq 0,$$

где $\tau_m$ — веса модальностей.

# ТМ [иерархии]

- **Шаг** 1. Строим модель с небольшим числом тем.
- ...
- **Шаг** $k$. Пусть модель с числом тем $T$ уже построена. Строим множество дочерних тем $S$ : $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_t KL_w\big(\phi_{wt}||\sum_{s\in S} p(w|s)p(s|t)\big) \to \min_{\Phi,\Psi},$$

где $p(w|s) = \phi_{ws}$, $p(s|t) = \psi_{st}$, $\Psi = (\psi_{st})_{S\times T}$ — матрица связей.
Родительская $\Phi_p \approx \Phi\Psi$, отсюда регуляризатор матрицы $\Phi$:

$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{t\in T} \phi_{ws}\psi_{st} \to \max_{\Phi,\Psi}$$

# Существующие подходы

- Нейросетевые модели:
  - **word2vec**: эмбеддинги слов
    T. Mikolov et al. *Efficient Estimation of Word Representations in Vector Space.* 2013
  - **FastText**: эмбеддинги символьных *n*-грамм
    P. Bojanowski, E. Grave, A. Joulin, T. Mikolov *Enriching Word Vectors with Subword Information.* 2017
  - **StarSpace**: эмбеддинги чего угодно (слова, предложения, документы).
    L. Wu, A.Fisch, S. Chopra, K. Adams, A.B.J. Weston *StarSpace: embed all the things!* 2018
  - ...
- **Вероятностные эмбеддинги слов**
  A. Potapenko, A. Popov, K. Vorontsov *Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks.* 2017

# Преимущества вероятностных эмбеддингов

- **Интерпретируемость**: каждый вектор является компонентом матрицы [*слова*; *темы*]
- **Иерархичность**: категоризация происходит автоматически благодаря иерархии тем
- **Мультимодальность**:
  - Модальность категорий повысит интерпретируемость тем
  - Модальности разных языков позволят осуществлять кросс-язычный поиск

# Применимость

## Категоризация

- **Дано**: множество документов $D$.
- **Найти**: отображение $f : D \to C$, где $C$ — латентное множество категорий с задаваемой размерностью.
- **Критерий**: Accuracy, Precision, Recall.

## Синонимия

- **Дано**: множество слов $W$.
- **Найти**: для некоторого $w \in W$ такое слово $w' \in W \setminus \{w\}$, что

$$w' = \arg \min_{w_s \in W \setminus \{w\}} \rho(f(w), f(w_s)),$$

  где $f : W \to \mathbb{R}^n$, а $\rho$ — некоторая метрика в $\mathbb{R}^n$.
- **Критерий**: корреляция с человеческими оценками.

# Применимость [2]

## Схожесть документов

- **Дано**: множество троек документов $\{q, d_1, d_2\}$, где
  - $q$ — рассматриваемый документ,
  - $d_1,\ d_2$ — два документа, один из которых схожей тематики с рассматриваемым, другой — нет.
- **Найти**: отображение $f : (q, d_1, d_2) \to \{(0, 1), (1, 0)\}$, определяющее, какой из документов является похожим, а какой — нет.
- **Критерий**: Accuracy, Precision, Recall.

# Применимость [3]

## Кросс-язычный поиск

- **Дано**: множества корпусов текстов параллельных переводов (EuroParl, TED) или пересказов (Википедия) на разных языках: $W^l,\ l \in L$.
- **Найти**: для некоторого запроса $W' \in W^l$ наиболее релевантный результат на всех языках.
- **Критерий**: среднее и медиана соответствующего документа в ранжированном списке документов.

# Новизна

- Тематические иерархические эмбеддинги слов
- Интерпретируемые и разреженные компоненты
- Предобученные на большой коллекции текстов с несколькими модальностями
- Способные эффективно решать несколько задач (Multi-task learning)