

Вероятностные тематические модели

Лекция 9.

Модели локальных контекстов

К. В. Воронцов
`k.v.vorontsov@phystech.edu`

Этот курс доступен на странице вики-ресурса
<http://www.MachineLearning.ru/wiki>
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 14 ноября 2024

- 1 Линейная тематизация текста**
 - Однопроходный E-шаг
 - Локализованный E-шаг
 - Двухнаправленные векторы контекста
- 2 Модели внимания и трансформеры**
 - Модели внимания
 - Трансформер кодировщик
 - Трансформер декодировщик
- 3 На пути к тематическим моделям внимания**
 - Локализованный E-шаг и модель внимания
 - Локализованный E-шаг и трансформер
 - Онлайн-алгоритм EM с локализованным E-шагом

Дано: коллекция текстовых документов $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$

Критерий: максимум **регуляризованного** log-правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); & n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \end{cases} \end{cases}$$

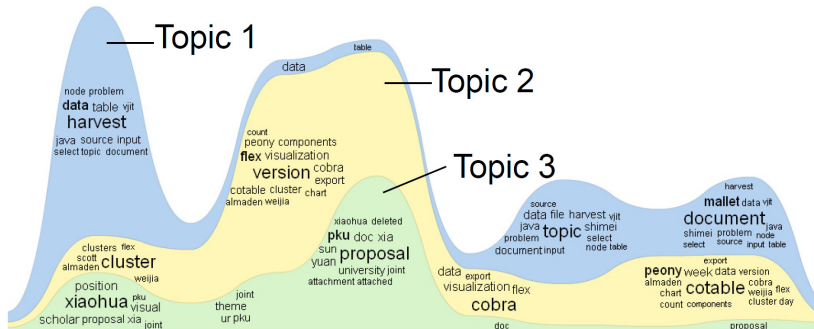
Задача тематизации фрагмента текста

Ситуации, в которых необходим тематический вектор $p(t|s)$ текстового фрагмента s , возможно, короткого:

- поиск в тексте фрагментов, наиболее релевантных запросу
- поиск наиболее тематичных предложений или фраз для
 - суммаризации документа (document summarization)
 - суммаризации темы (topic summarization)
 - автоматического именования темы (topic labeling)
- определение тематики нового слова/словосочетания по окружающему его локальному контексту s
- тематические модели предложений
 - когда разбиение текста на предложения/секции задано
- тематические модели сегментации текста
 - когда границы сегментов требуется найти
- отображение карты распределения тем внутри документа

Пример 1. Отображение распределения тем внутри документа

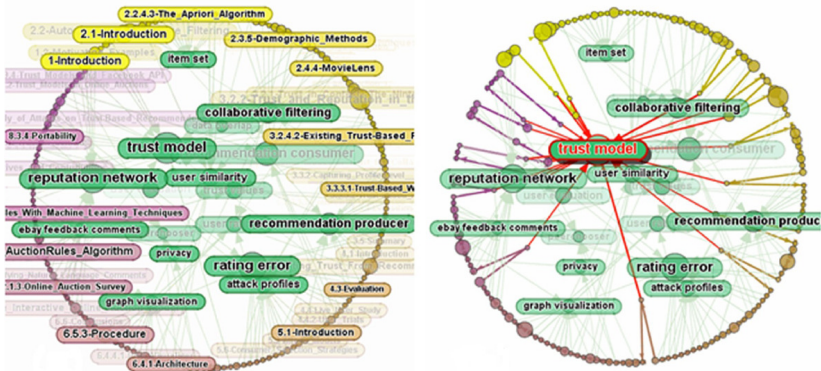
Визуализация динамики тем обычно применяется к новостным потокам, но может также применяться к длинным документам, если их разбивать на сегменты.



Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, Xiaoxiao Lian. TIARA: interactive, topic-based visual text summarization and analysis. 2012.

Пример 2. Отображение распределения тем внутри документа

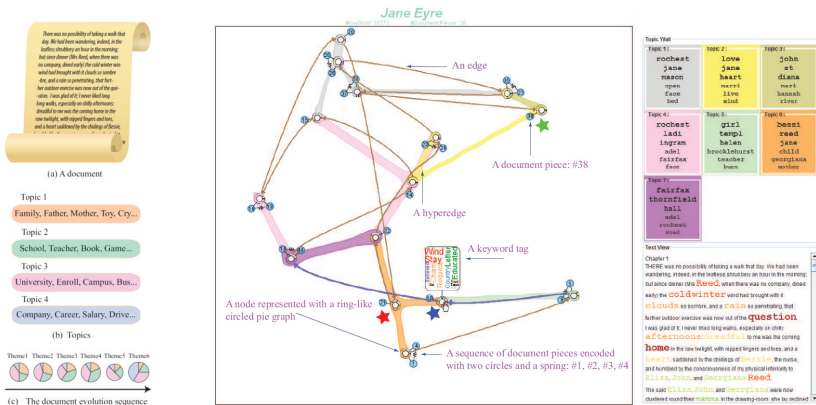
Динамика всех тем и одной выделенной темы в документе



Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. 2012.

Пример 3. Выделение сюжетных линий в длинных текстах

Иерархическая модель выделяет сюжетные линии (topics) и их сочетания (themes) в тексте художественного произведения



Guizhen Wang, Chaokai Wen, Binghui Yan, Jing Xia, Zhen Liu, Wei Chen. Topic hypergraph: hierarchical visualization of thematic structures in long documents. 2012.

Идея тематизации текста за один проход

Дано: s — фрагмент текста d , Φ — тематическая модель

Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- как не переобучить вектор $p(t|s)$, если текст короткий?
- как согласовать $p(t|s)$ с объемлющим контекстом $p(t|d)$?
- как согласовать $p(t|s)$ с $p(t|w) = \phi_{wt} \frac{p(t)}{p(w)}$ термов $w \in s$?

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right) = \sum_{w \in d} \frac{n_{dw}}{n_d} \operatorname{norm}_{t \in T} (\phi_{wt} \theta_{td}^0)$$

- формула полной вероятности:

$$\theta_{td}(\Phi) = \sum_{w \in d} p(w|d) p(t|w, d) = \sum_{w \in d} \frac{n_{dw}}{n_d} \operatorname{norm}_{t \in T} (\phi_{wt} p_t)$$

EM-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} \equiv p(t|d, w) = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}}$$

$$\phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по Лемме о максимизации на симплексах)

Оптимизационная задача M-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d \in D} \sum_{u \in W} \sum_{s \in T} n_{du} p_{sdu} \ln(\phi_{us} \theta_{sd}(\Phi)) + R(\Phi, \Theta(\Phi)) \rightarrow \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q :

$$\begin{aligned} \phi_{wt} \frac{\partial Q}{\partial \phi_{wt}} &= \sum_{d \in D} n_{dw} p_{tdw} + \sum_{d,s,u} n_{du} p_{sdu} \frac{\phi_{wt}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \sum_{d,s} \frac{\partial R}{\partial \theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = \\ &= \sum_{d \in D} n_{dw} \left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{\phi_{wt}}{\theta_{sd}} \underbrace{\left(\sum_{u \in d} n_{du} p_{sdu} + \theta_{sd} \frac{\partial R}{\partial \theta_{sd}} \right)}_{n_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} = \\ &= \sum_{d \in D} n_{dw} \underbrace{\left(p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right)}_{p'_{tdw}} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}. \quad \blacksquare \end{aligned}$$

EM-алгоритм для ARTM с линейной тематизацией документов

$$\theta_{td}(\Phi) = \sum_{w \in D} \frac{n_{dw}}{n_d} \operatorname{norm}_{t \in T}(\phi_{wt} p_t) \Rightarrow \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = \frac{n_{dw}}{n_d} \phi'_{tw} (\delta_{st} - \phi'_{sw})$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \operatorname{norm}_{t \in T}(\phi_{wt} n_t); \quad \theta_{td} = \sum_{w \in D} \frac{n_{dw}}{n_d} \phi'_{tw}$$

$$p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad n_t = \sum_{d \in D} \sum_{w \in D} n_{dw} p_{tdw}$$

$$n_{td} = \sum_{w \in D} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

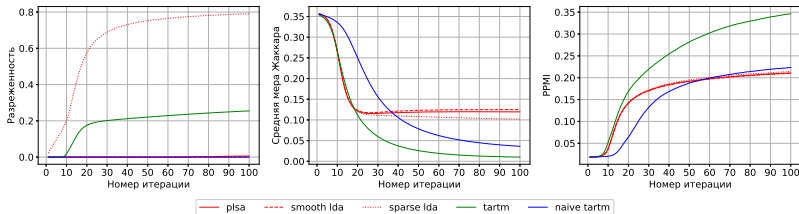
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θ less ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

https://github.com/ilirhin/python_artm

Упрощение EM-алгоритма для линейной тематизации

- Нет регуляризации по Θ , следовательно, $\frac{\partial R}{\partial \theta_{td}} = 0$
- Значение отношения $\frac{n_{td}}{\theta_{td}} \approx n_d$ не зависит от t , подстановка в формулу M-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \operatorname{norm}_{t \in T}(\phi_{wt} n_t); \quad \theta_{td} = \sum_{w \in D} \frac{n_{dw}}{n_d} \phi'_{tw};$$

$$p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad n_t = \sum_{d \in D} \sum_{w \in D} n_{dw} p_{tdw};$$

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

Это обычный EM-алгоритм, только с однопроходным E-шагом!
ОГО! И ТАК МОЖНО БЫЛО?!

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \dots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\dots, w_i, \dots)$ термина w_i :

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Тематизация локального контекста с распределением весов:

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i), \quad \sum_{u \in C_i} \alpha(u|i) = 1, \quad \alpha(u|i) \geq 0$$

Локализованная тематическая модель (похожа на BitermTM):

$$p(w|C_i) = \sum_{t \in T} p(w|t) p(t|C_i) = \sum_{t \in T} \phi_{wt} \sum_{u \in C_i} \phi'_{tu} \alpha(u|i)$$

EM-алгоритм с локализованным E-шагом

w_1, \dots, w_n — сквозная нумерация термов во всей коллекции

C_i — локальный контекст (окружение) термина w_i

$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для термина w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение коллекции на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \operatorname{norm}_{t \in T}(\phi_{wt} p_t); \quad \theta_{ti} \equiv p(t|C_i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i);$$

$$p_{ti} \equiv p(t|C_i, w_i) = \operatorname{norm}_{t \in T}(\phi_{w_i t} \theta_{ti}); \quad p_t \equiv p(t) = \frac{1}{n} \sum_{i=1}^n p_{ti};$$

$$\phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{i=1}^n [w_i = w] p_{ti} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

Быстрое вычисление двунаправленных векторов контекста

Два прохода по тексту — «слева направо» и «справа налево» для вычисления экспоненциальных скользящих средних (ЭСС):

$$\vec{p}(t|i) = \vec{\gamma}_i p(t|w_i) + (1 - \vec{\gamma}_i) \vec{p}(t|i-1), \quad i = 1, \dots, n, \quad \vec{\gamma}_1 = 1$$

$$\vec{p}(t|i) = \vec{\gamma}_i p(t|w_i) + (1 - \vec{\gamma}_i) \vec{p}(t|i+1), \quad i = n, \dots, 1, \quad \vec{\gamma}_n = 1$$

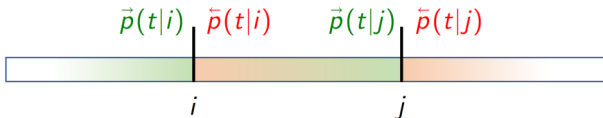
где $\vec{\gamma}_i$, $\vec{\gamma}_i$ — коэффициенты сглаживания в позиции i

Основное свойство: если $\gamma_i = \gamma$, то $\alpha(w_k|i) = \gamma(1 - \gamma)^{|i-k|}$

Несколько соображений, как распоряжаться выбором $\vec{\gamma}_i$, $\vec{\gamma}_i$:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм
- γ_i можно умножать на оценку важности термина

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{p}(t|i)$ — тематика левого контекста термина w_i
- $\vec{p}(t|i)$ — тематика правого контекста термина w_i
- $\frac{1}{2}(\vec{p}(t|i) + \vec{p}(t|i))$ — тематика двустороннего контекста w_i
- $p(t|i \dots j) = \frac{1}{2}(\vec{p}(t|i) + \vec{p}(t|j))$ — тематика сегмента $[i \dots j]$
- $\vec{p}(t|i) \approx \vec{p}(t|j)$ — однородность тематики сегмента $[i \dots j]$
- $\max_i \|\vec{p}(t|i) - \vec{p}(t|i)\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты

Гипотеза: нет ли полезной аналогии с моделями внимания?

Онлайновый EM-алгоритм с локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры $\beta, \vec{\gamma}_i, \tilde{\gamma}_i, \alpha, \delta$;

Выход: матрица Φ , векторы термов документов p_{ti} ;

инициализация: $n_{wt} := 0; \tilde{n}_{wt} := 0; n_t := 1; \phi_{wt} := \text{random};$

для всех документов $d \in D$

$$p_{ti} := \text{norm}_t(\phi_{w_i t} n_t), \quad i = 1, \dots, n_d, \quad t \in T;$$

$$\vec{\theta}_{ti} := \vec{\gamma}_i p_{ti} + (1 - \vec{\gamma}_i) \vec{\theta}_{t, i-1}, \quad i = 1, \dots, n_d, \quad \vec{\gamma}_1 = 1, \quad t \in T;$$

$$\tilde{\theta}_{ti} := \tilde{\gamma}_i p_{ti} + (1 - \tilde{\gamma}_i) \tilde{\theta}_{t, i+1}, \quad i = n_d, \dots, 1, \quad \tilde{\gamma}_{n_d} = 1, \quad t \in T;$$

$$p_{ti} := \text{norm}_t(\phi_{w_i t} (\beta \vec{\theta}_{ti} + (1 - \beta) \tilde{\theta}_{ti})), \quad i = 1, \dots, n_d, \quad t \in T;$$

$$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti}, \quad i = 1, \dots, n_d, \quad t \in T;$$

если пора обновить матрицу Φ **то**

$$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}; \quad \tilde{n}_{wt} := 0;$$

$$\phi_{wt} := \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

Напоминание. Модель внимания Query–Key–Value

q — вектор-запрос, трансформируемый в контекстный вектор z .

Контекст задаётся последовательностью n значений с ключами:

$V = (v_1, \dots, v_n)$ — векторы-значения;

$K = (k_1, \dots, k_n)$ — векторы-ключи.

Модель внимания — это выпуклая комбинация векторов v_i , взвешенных по сходству их ключей k_i с запросом q :

$$z = \text{Attn}(q, K, V) = \sum_{i=1}^n v_i \text{SoftMax}_i \langle k_i, q \rangle$$

Модель само-внимания (self-attention) трансформирует

$X = (x_1, \dots, x_n)$ — входные бесконтекстные векторы в

$Z = (z_1, \dots, z_n)$ — выходные контекстные векторы:

$$z_i = \text{Attn}(W_q x_i, W_k X, W_v X),$$

где W_q, W_k, W_v — обучаемые матрицы параметров.

Vaswani et al. Attention is all you need. 2017.

BERT — Bidirectional Encoder Representations from Transformers

Трансформер BERT — двунаправленный кодировщик текста, предобучаемый для решения различных задач NLP

Схема преобразования данных:

- $S = (w_1, \dots, w_n)$ — токены входного текста
↓ обучение векторов (эмбедингов) токенов
- $X = (x_1, \dots, x_n)$ — бесконтекстные векторы токенов
↓ многократная трансформация через само-внимание
- $Z = (z_1, \dots, z_n)$ — контекстные векторы токенов
↓ дообучение на конкретную задачу
- Y — разметка текста / классификация и т.п.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Dichao Hu. An introductory survey on attention mechanisms in NLP problems. 2018.

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i :

$$h_i = x_i + p_i, \quad H = (h_1, \dots, h_n) \quad \begin{array}{l} d = \dim x_i, p_i, h_i = 512 \\ \dim H = 512 \times n \end{array}$$

2. Многомерное само-внимание: $j = 1, \dots, J = 8$

$$h_i^j = \text{Attn}(W_q^j h_i, W_k^j H, W_v^j H) \quad \begin{array}{l} \dim h_i^j = 64 \\ \dim W_q^j, W_k^j, W_v^j = 64 \times 512 \end{array}$$

3. Конкатенация (multi-head attention):

$$h_i' = \text{MH}_J(h_i^j) \equiv [h_i^{j_1} \dots h_i^{j_J}] \quad \dim h_i' = 512$$

4. Сквозная связь + нормировка уровня:

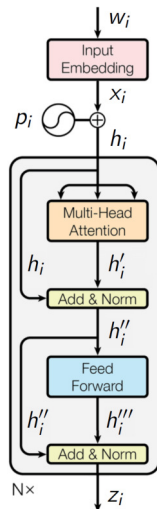
$$h_i'' = \text{LN}(h_i' + h_i; \mu_1, \sigma_1) \quad \dim h_i'', \mu_1, \sigma_1 = 512$$

5. Полносвязная 2х-слойная сеть FFN:

$$h_i''' = W_2 \text{ReLU}(W_1 h_i'' + b_1) + b_2 \quad \begin{array}{l} \dim W_1 = 2048 \times 512 \\ \dim W_2 = 512 \times 2048 \end{array}$$

6. Сквозная связь + нормировка уровня:

$$z_i = \text{LN}(h_i''' + h_i''; \mu_2, \sigma_2) \quad \dim z_i, \mu_2, \sigma_2 = 512$$



Критерий обучения MLM (Masked Language Modeling)

Критерий маскированного языкового моделирования MLM, строится автоматически по текстам (self-supervised learning):

$$\sum_S \sum_{i \in M(S)} \ln p(w_i | i, S, W) \rightarrow \max_W,$$

где $M(S)$ — подмножество (15%) маскированных токенов из S ,

$$p(w | i, S, W) = \underset{w}{\text{SoftMax}}(W_z z_i(S, W_T) + b_z)$$

— языковая модель, предсказывающая i -й токен в тексте S ;

$z_i(S, W_T)$ — контекстный вектор i -го токена текста S на выходе Трансформера с параметрами W_T ;

$W = (W_T, W_z, b_z)$ — все параметры языковой модели

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Трасформер для генерации текста

Добавляется трасформер-декодировщик схожей архитектуры

Схема преобразования данных:

- $S = (w_1, \dots, w_n)$ — токены входного текста
↓ обучаемая или пред-обученная векторизация
- $X = (x_1, \dots, x_n)$ — бесконтекстные векторы токенов
↓ трансформер-кодировщик
- $Z = (z_1, \dots, z_n)$ — контекстные векторы входных токенов
↓ трансформер-декодировщик
- $Y = (y_1, \dots, y_m)$ — контекстные векторы выходных токенов
↓ генерация токенов по контекстным векторам
- $\tilde{S} = (\tilde{w}_1, \dots, \tilde{w}_m)$ — токены выходного текста

Vaswani et al. (Google) Attention is all you need. 2017.

Tom B. Brown et al. (OpenAI) Language models are few-shot learners. 2020.

Архитектура трансформера декодировщика

$y_0 = \langle \text{BOS} \rangle$ — эмбединг символа начала;

для всех $t = 1, 2, \dots$:

1. Маскирование «данных из будущего»:

$$h_t = y_{t-1} + p_t; \quad H_t = (h_1, \dots, h_t)$$
2. Многомерное само-внимание:

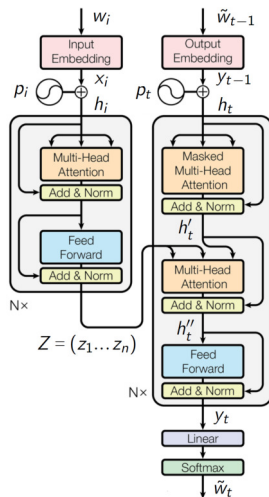
$$h'_t = \text{LN} \circ \text{MH}_J \circ \text{Attn}(W_q^j h_t, W_k^j H_t, W_v^j H_t)$$
3. Многомерное внимание на кодировку Z :

$$h''_t = \text{LN} \circ \text{MH}_J \circ \text{Attn}(\tilde{W}_q^j h'_t, \tilde{W}_k^j Z, \tilde{W}_v^j Z)$$
4. Двухслойная полносвязная сеть:

$$y_t = \text{LN} \circ \text{FFN}(h''_t)$$
5. Линейный предсказывающий слой:

$$p(\tilde{w}|t) = \text{SoftMax}(W_y y_t + b_y)$$

генерация $\tilde{w}_t = \arg \max_{\tilde{w}} p(\tilde{w}|t)$ пока $\tilde{w}_t \neq \langle \text{EOS} \rangle$



Vaswani et al. (Google) Attention is all you need. 2017.

Аналогия локализованного E-шага с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \text{norm}_{t \in T}(\phi_{w_i t} \theta_{ti}) = \text{norm}_{t \in T} \left(\sum_{u \in C_i} \phi'_{tu} \phi_{w_i t} \alpha(u|i) \right)$$

Контекстный вектор на выходе модели само-внимания:

$$z_i = \sum_{u \in C_i} W_v x_u \alpha(u|i) = \sum_{u \in C_i} W_v x_u \text{SoftMax}_{u \in C_i}(W_q x_i, W_k x_u)$$

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ'_u из контекста терма w_i ,
- наиболее (семантически) схожих с вектором терма w_i .

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i} ;
- нет обучаемых матриц W_q, W_k, W_v как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Аналогия локализованного E-шага с моделью трансформера

Один проход документа аналогичен модели внимания:

— для каждого $d \in D$, для каждой позиции $i = 1, \dots, n_d$
вычисляются 5 тематических векторов, связанных с термом w_i :

$\phi'_{tw_i} = \text{norm}_t(\phi_{w_i t} p_t)$ — бесконтекстный вектор термина $p(t|w_i)$

$\vec{p}(t|i) = \vec{\theta}_{ti}$, $\bar{p}(t|i) = \bar{\theta}_{ti}$ — векторы левого и правого контекста

$\theta_{ti} = \beta \vec{\theta}_{ti} + (1 - \beta) \bar{\theta}_{ti}$ — вектор двустороннего контекста

$p_{ti} = \text{norm}_t(\phi_{w_i t} \theta_{ti})$ — контекстный вектор термина $p(t|C_i, w_i)$

Несколько таких проходов аналогичны трансформеру:

контекстный вектор термина $p_{ti} = p(t|C_i, w_i)$ с предыдущего прохода
используется вместо его бесконтекстного вектора $\phi'_{tw_i} = p(t|w_i)$

L таких итераций аналогичны проходу L блоков внимания

Онлайновый EM с многопроходным локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры $L, \beta, \vec{\gamma}_i, \overleftarrow{\gamma}_i, \alpha, \delta$;

Выход: матрица Φ , векторы термов документов p_{ti} ;

инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} := \text{random}$;

для всех документов $d \in D$

$$p_{ti} := \text{norm}_t(\phi_{wt} n_t);$$

для всех $l = 1, \dots, L$ (аналог L блоков внимания)

$$\vec{\theta}_{ti} := \vec{\gamma}_i p_{ti} + (1 - \vec{\gamma}_i) \vec{\theta}_{t,i-1}, \quad i = 1, \dots, n_d, \quad \vec{\gamma}_1 = 1;$$

$$\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i p_{ti} + (1 - \overleftarrow{\gamma}_i) \overleftarrow{\theta}_{t,i+1}, \quad i = n_d, \dots, 1, \quad \overleftarrow{\gamma}_{n_d} = 1;$$

$$p_{ti} := \text{norm}_t((\beta \vec{\theta}_{ti} + (1 - \beta) \overleftarrow{\theta}_{ti}) p_{ti} / n_t);$$

$$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}; \quad n_t := n_t + p_{ti};$$

если пора обновить матрицу Φ **то**

$$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}; \quad \tilde{n}_{wt} := 0;$$

$$\phi_{wt} := \text{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right);$$

Открытые проблемы и постановки задач

- надо ли исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}, \bar{\theta}_{ti}$?
 - какие другие варианты $\alpha(u|i)$ кроме скользящих средних?
 - как выбрать вес β левого контекста?
 - правильно ли подставлять p_{ti}/n_t вместо $\phi_{w_{it}}$ на E-шаге?
 - имеет ли смысл увеличивать число проходов L ?
-
- как (и нужно ли) параметризовать модель внимания?
 - как обучать её параметры, разные для разных проходов?
 - как (и нужно ли) ввести аналог многих голов внимания?
-
- слишком много эвристических преобразований сделано... мы всё ещё решаем исходную оптимизационную задачу?
 - действительно ли на E-шаге можно подвергать $p(t|d, w_i)$ всяким модификациям, почему и в каких пределах?

- Работает ли такая тематическая модель внимания?
- Возможны ли другие тематические модели внимания? (есть много попыток объединять тематические модели с нейросетевыми моделями языка, attention, transformer)
- Механизмы учёта порядка слов в ARTM:
 - модели n -грамм, коллокаций, словосочетаний
 - модели сочетаемости пар слов: BitermTM, WNTM
 - линейная однопроходная тематизация документов
 - многопроходная тематизация (аналог трансформера)
 - регуляризация E-шага
 - модели предложений или сегментов
 - тематическая сегментация: модель TopicTiling

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. 2022

He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021

Xiaobao Wu, Thong Nguyen, Anh Tuan Luu. A Survey on Neural Topic Models: Methods, Applications, and Challenges. 2023

Tian Tian et al. Attention-based Autoencoder Topic Model for Short Texts. 2019

Shuangyin Li et al. Recurrent Attentional Topic Model. 2017

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	5X
обзор по NeuralTM	5X
интеграция ARTM в pyTorch	5X
участие в одном из проектов	10X
работа над открытой проблемой	10X

где X — оценка за вид деятельности по 5-балльной шкале.

Итоговая оценка: $\min(10, \lfloor \text{score}/5 \rfloor)$ по 10-балльной шкале.

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайн-алгоритма EM
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_{it}}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}, \overleftarrow{\theta}_{ti}$ »

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

Найти параметры модели ξ_{dw} .

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

Найти параметры модели ξ_w .

Подсказка: применить условия ККТ или основную лемму.

3. (более творческое задание)

Предложите модель, определяющую роли слов в текстах:

— тематические слова

— специфичные слова документа (шум)

— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т, ш, ф}\}$.

Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

4. Заменяем \log другой монотонно возрастающей функцией μ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \mu \left(\sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Как изменится EM-алгоритм? Возможно ли подобрать функцию μ так, чтобы сократился объём вычислений?

5. Заменяем \log монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \mu(\phi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_t \mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

6. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \text{norm}_w(n_{wt} [n_{wt} > \gamma n_t])$$

Аналитик построил тематическую модель Φ^0, Θ^0 и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

7. Предложите регуляризаторы для этого.

8. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi_{wt}^0$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

9. Предложите способ инициализации Φ для новой модели.

10. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = (p(s|t))$, гарантирующий, что

- 1) у каждой родительской темы будет хотя бы одна дочерняя;
- 2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу M-шага для матрицы Ψ .

11. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s , то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$.

12. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

- 15.** Выведите EM-алгоритм с локализованным E-шагом (слайд 15) для локализованной тематической модели. Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw} ?
- 16.** Предложите параметризацию для тематической модели внимания (слайд 25). Используя «основную лемму», получите уравнения для новых параметров модели.