Вероятностные тематические модели: от теории регуляризации к моделям внимания

Воронцов Константин Вячеславович
(ВМК МГУ, МФТИ, ФИЦ ИУ РАН)

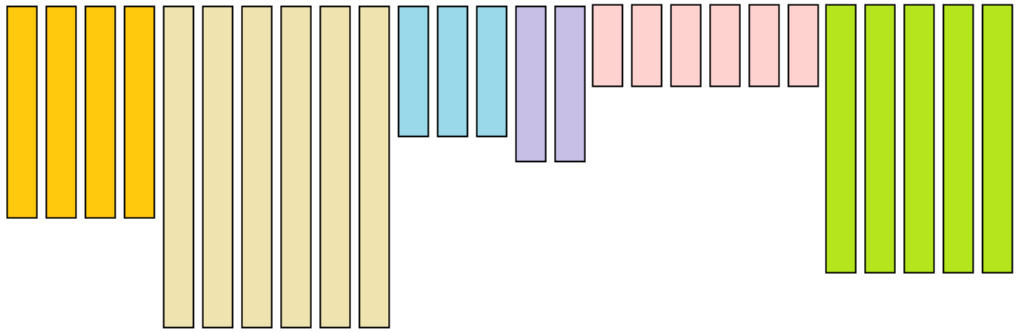
XII Международная молодёжная научно-практическая конференция с элементами научной школы «Прикладная математика и фундаментальная информатика»
Омский ГТУ • 16–22 мая 2022

Содержание

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Решения $\omega_j \equiv 0$ отбрасываются как вырожденные
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020.

Пусть

- W — конечное множество слов (термов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- каждое слово w в документе d связано с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Тематическая модель, по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d)$$

Постановка задачи тематического моделирования

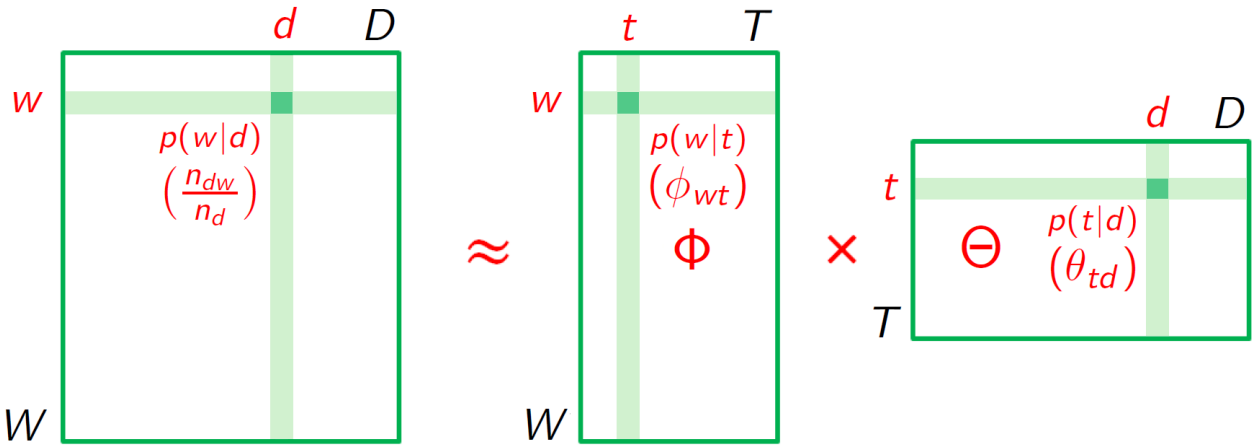
Дано: коллекция текстовых документов

- n_{dw} — частоты термов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

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

- $\phi_{wt} = p(w|t)$ — вероятности термов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \underbrace{\cancel{p(d)}}_{\text{const}} \to \max_{\Phi, \Theta}$$

приводит к задаче математического программирования:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Пример. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Интерпретируемость тематических векторов

Тематические векторные представления текста:

- $p(t|d) = \theta_{td}$ для каждого документа d
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ для каждого терма w
- $p(t|d,w)$ для каждого локального контекста (d,w)

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром*, частотным словарём слов $\left\{w\colon p(w|t) > \gamma p(w)\right\}$, встречающихся в данной теме в γ раз чаще обычного
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\left\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\right\}$

Цели и не-цели тематического моделирования

Цели:

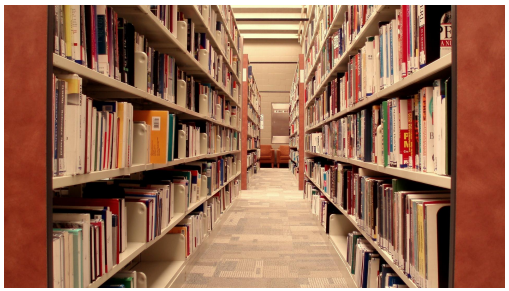
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и какие они
- Получать интерпретируемые тематические векторные представления документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать следующие слова (TM — слабые модели языка)
- Генерировать связный текст
- Понимать смысл текста

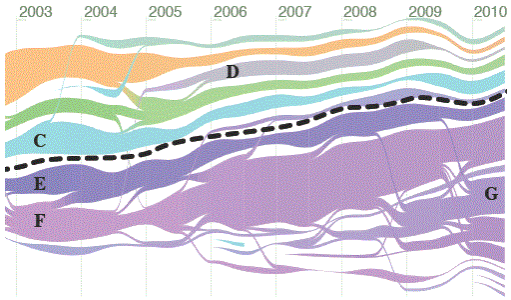
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

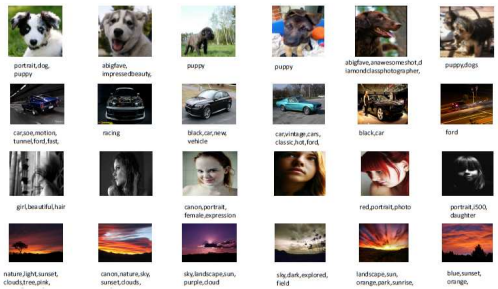


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

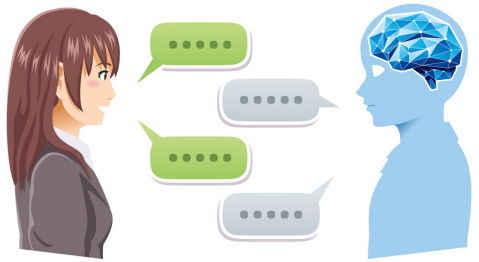


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$
- $L(\Phi', \Theta') \leqslant L(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw} \end{cases}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \underset{w\in W}{\operatorname{norm}}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

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

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \underset{t\in T}{\operatorname{norm}}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Два наиболее известных частных случая: модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td}\big).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

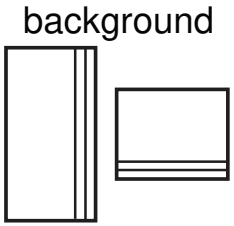
$$R(\Phi, \Theta) = \sum_{t,w}(\beta_w - 1)\ln\phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln\theta_{td}.$$

M-шаг — частотные оценки с поправками $\beta_w > 0$, $\alpha_t > 0$:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w - 1\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t - 1\big).$$

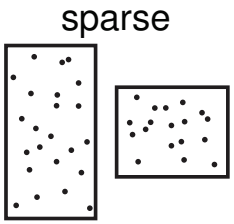
Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

Регуляризаторы для улучшения интерпретируемости тем



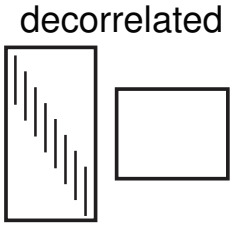
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$



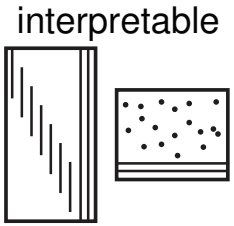
Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$



Декоррелирование для повышения различности тем:

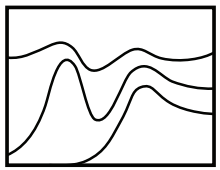
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

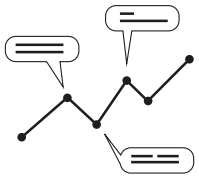
temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|$$

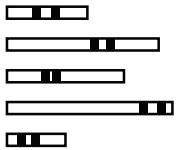
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

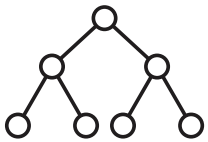
coherence



Модели сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

hierarchy



Связь родительских тем t с дочерними подтемами s:

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

Модульность аддитивной регуляризации

Набор регуляризаторов подбирается для каждой задачи

Выявление этнорелевантного дискурса в социальных сетях:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{n-gram}\right) + R\left(\text{seed words}\right) \to \max$$

Тематический поиск научных и научно-популярных статей:

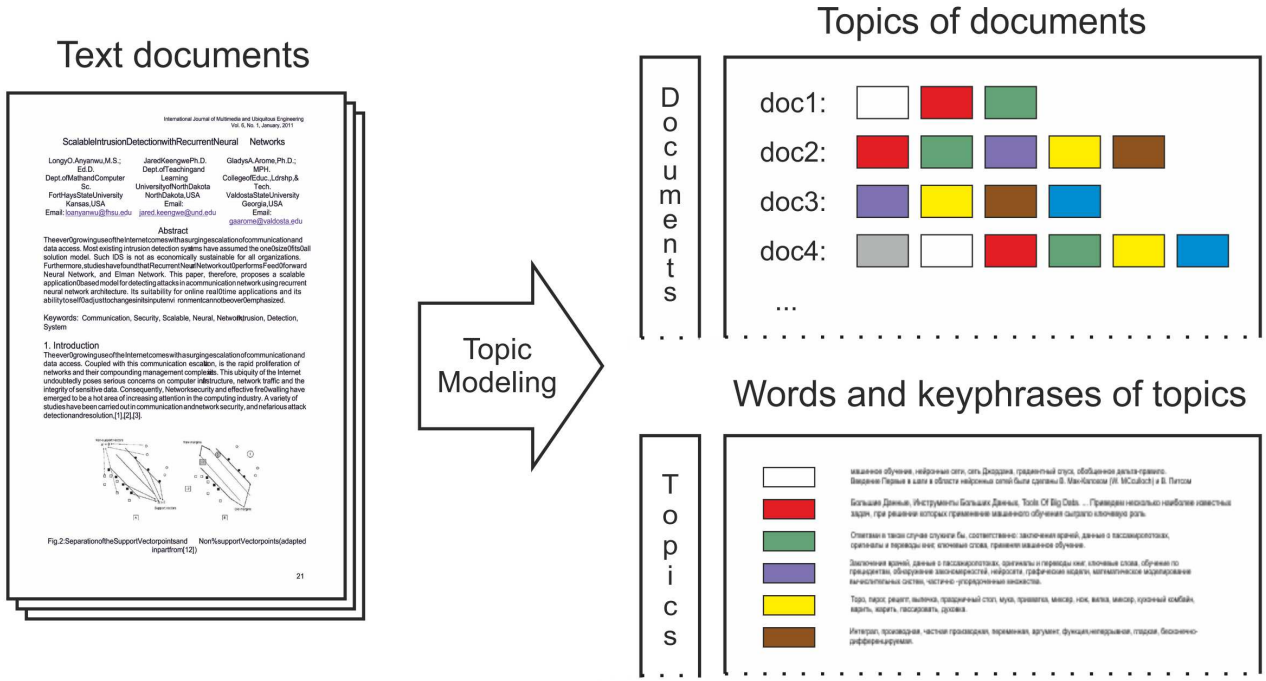
$$\mathscr{L}\left(\text{multimodal}\right) + R\left(\text{interpretable}\right) + R\left(\text{n-gram}\right) + R\left(\text{hierarchy}\right) \to \max$$

Выявление и прослеживание событий в новостном потоке:

$$\mathscr{L}\left(\text{multimodal}\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{sentiment}\right) \to \max$$

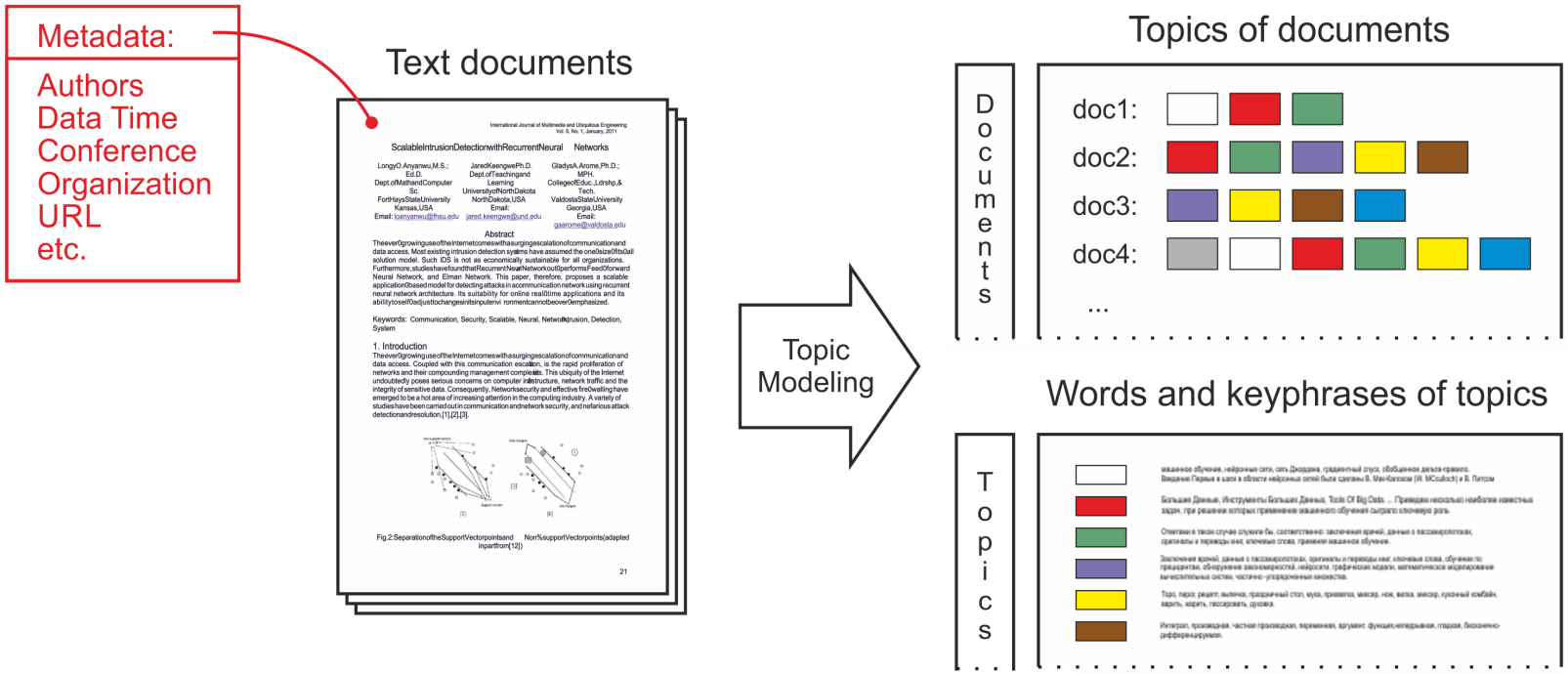
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$,



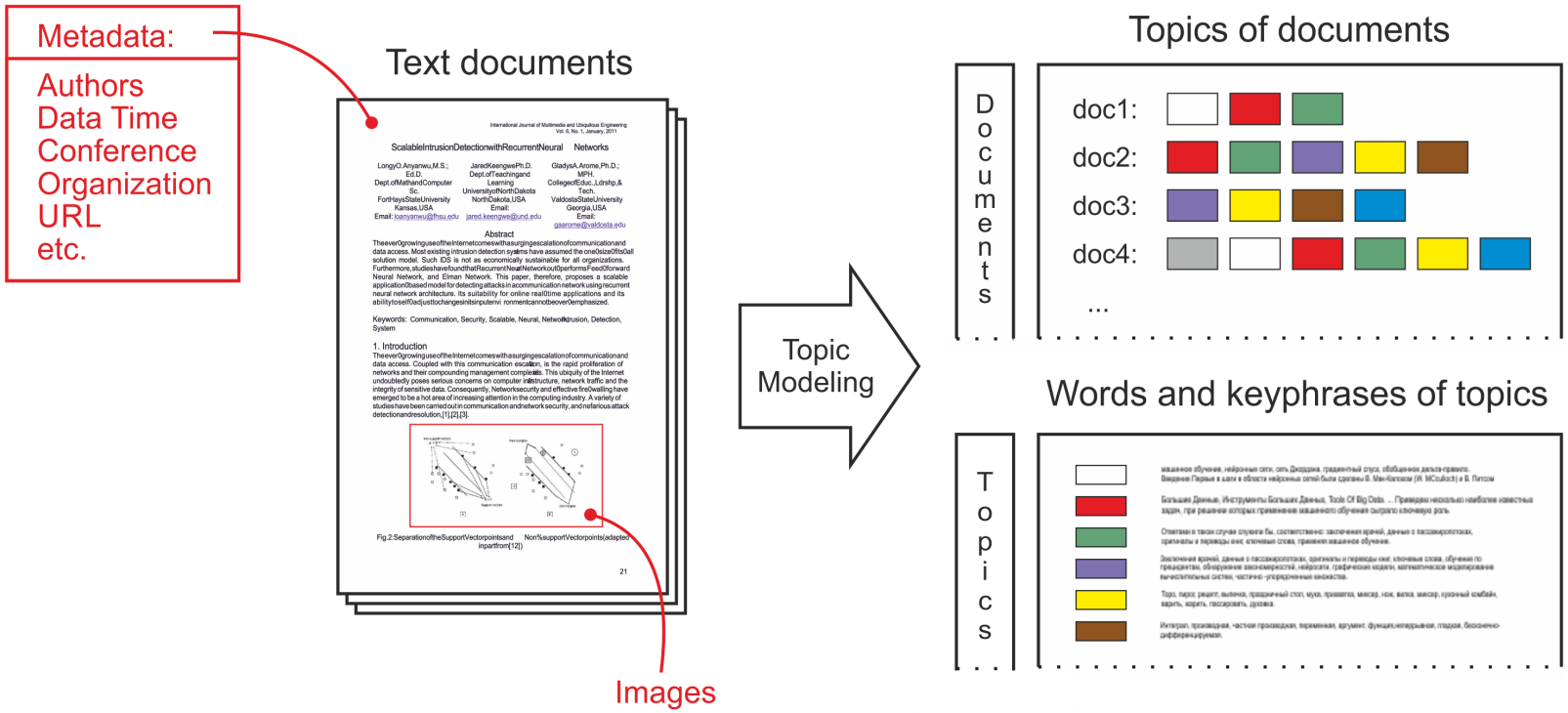
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,



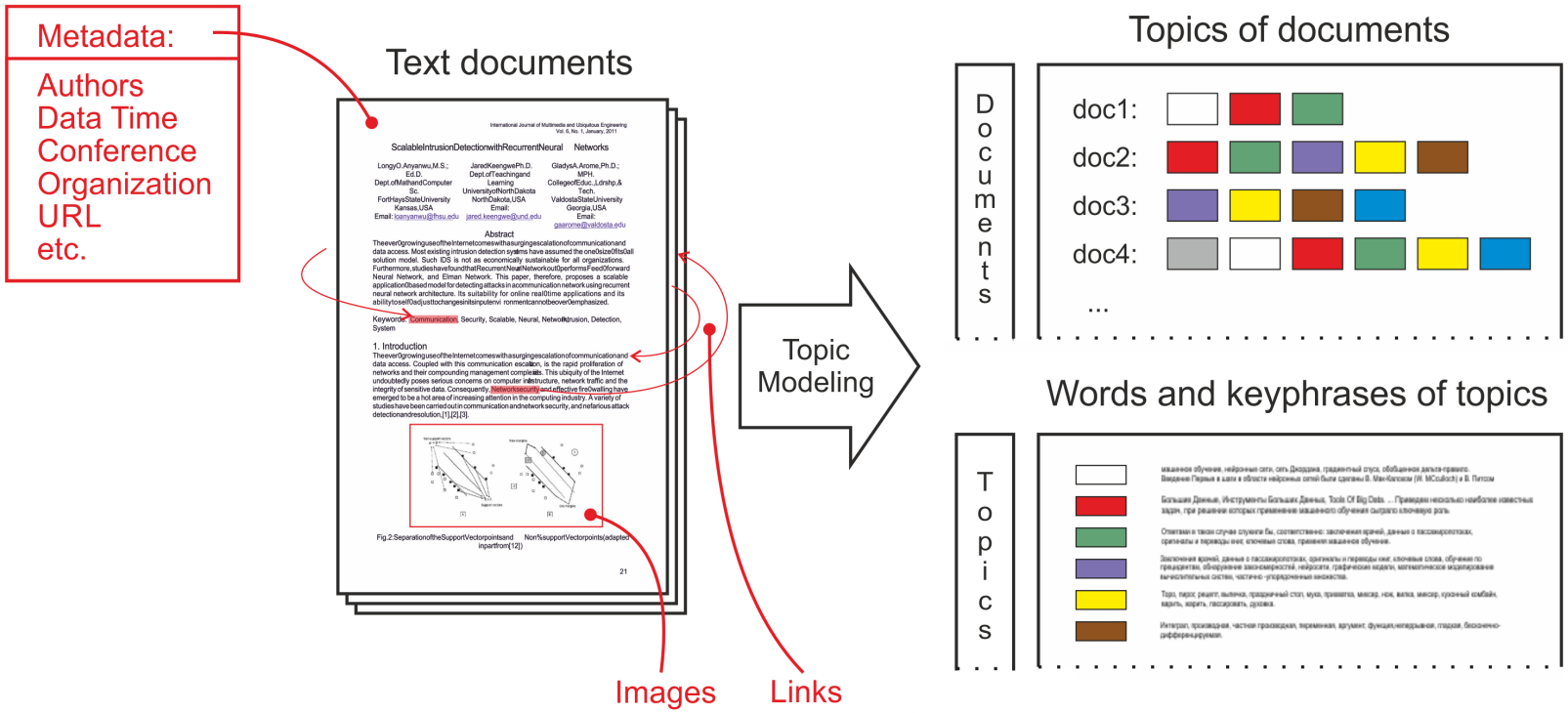
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$,



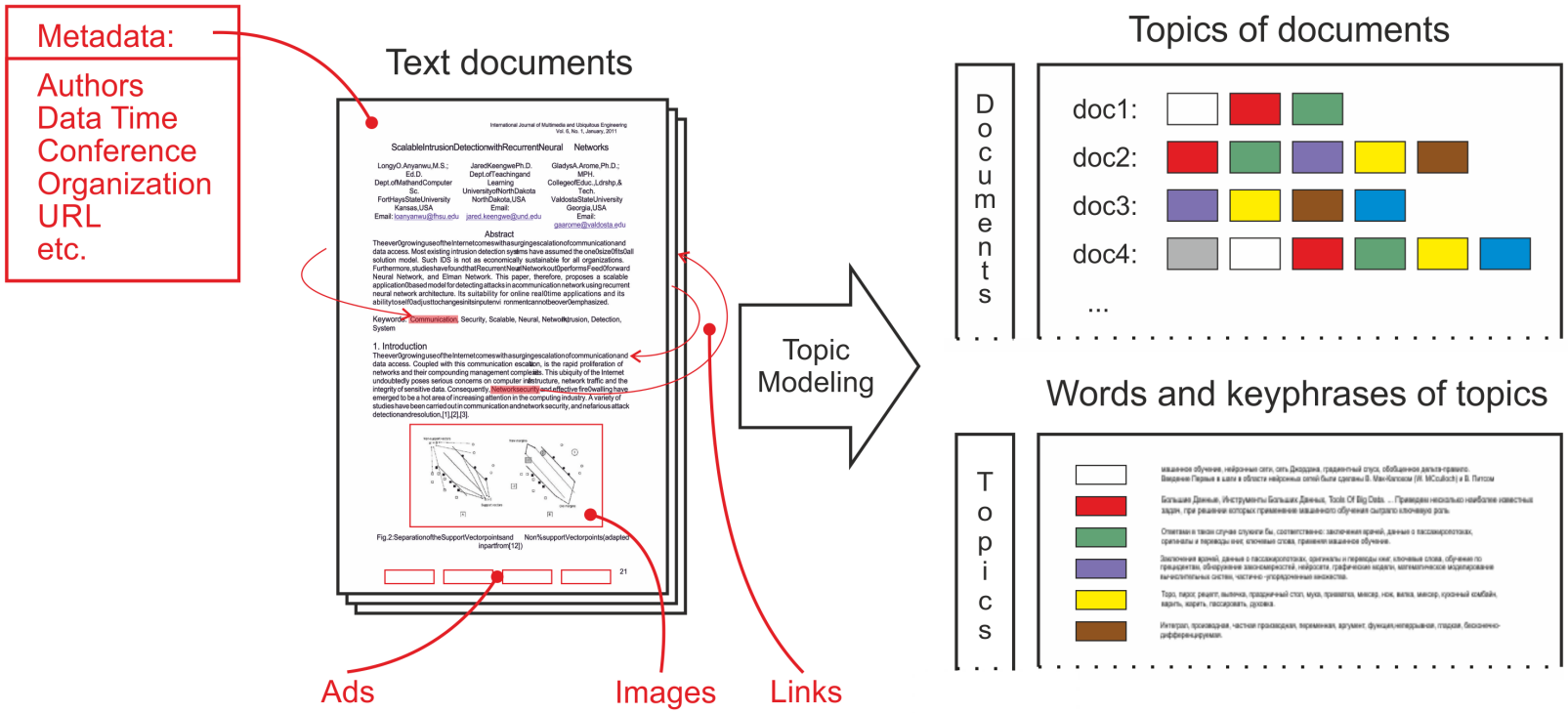
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



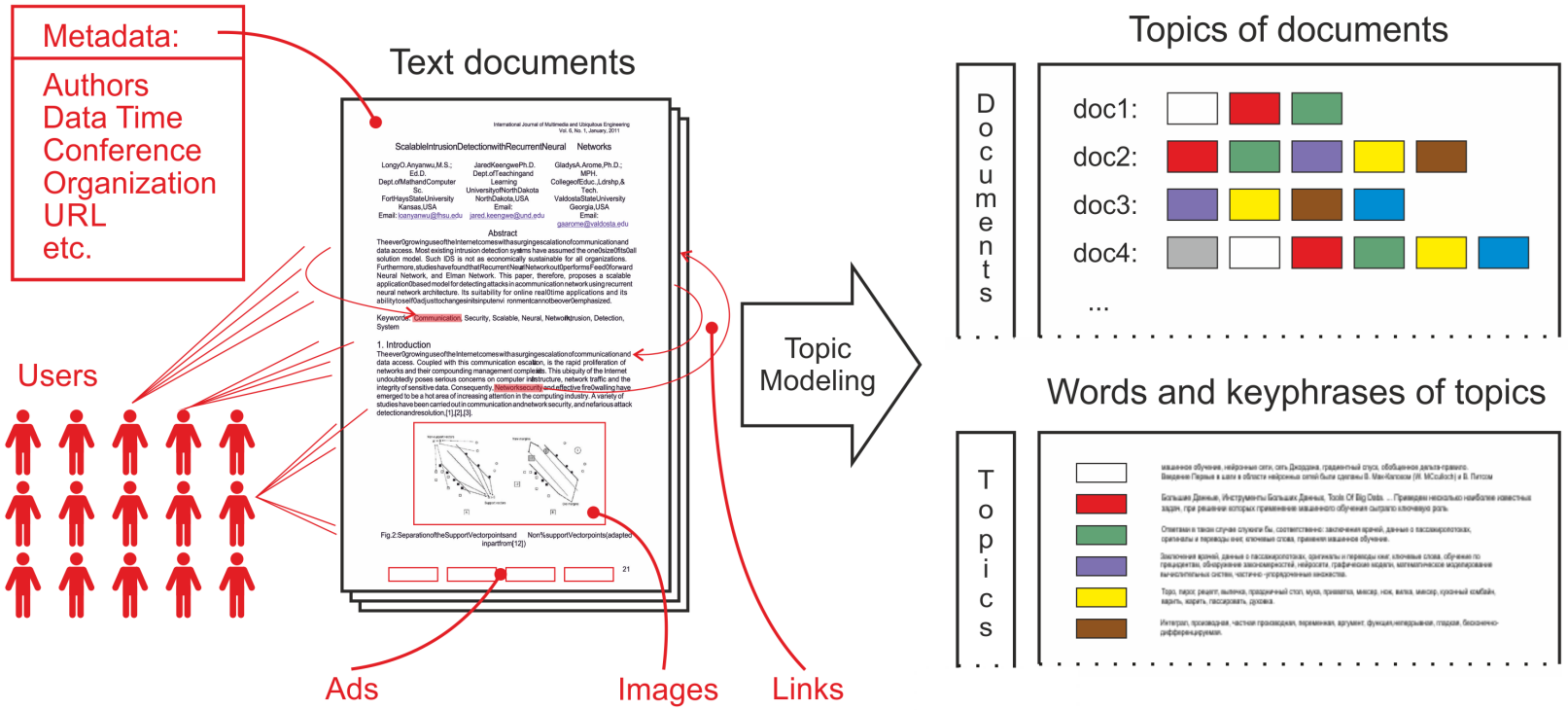
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



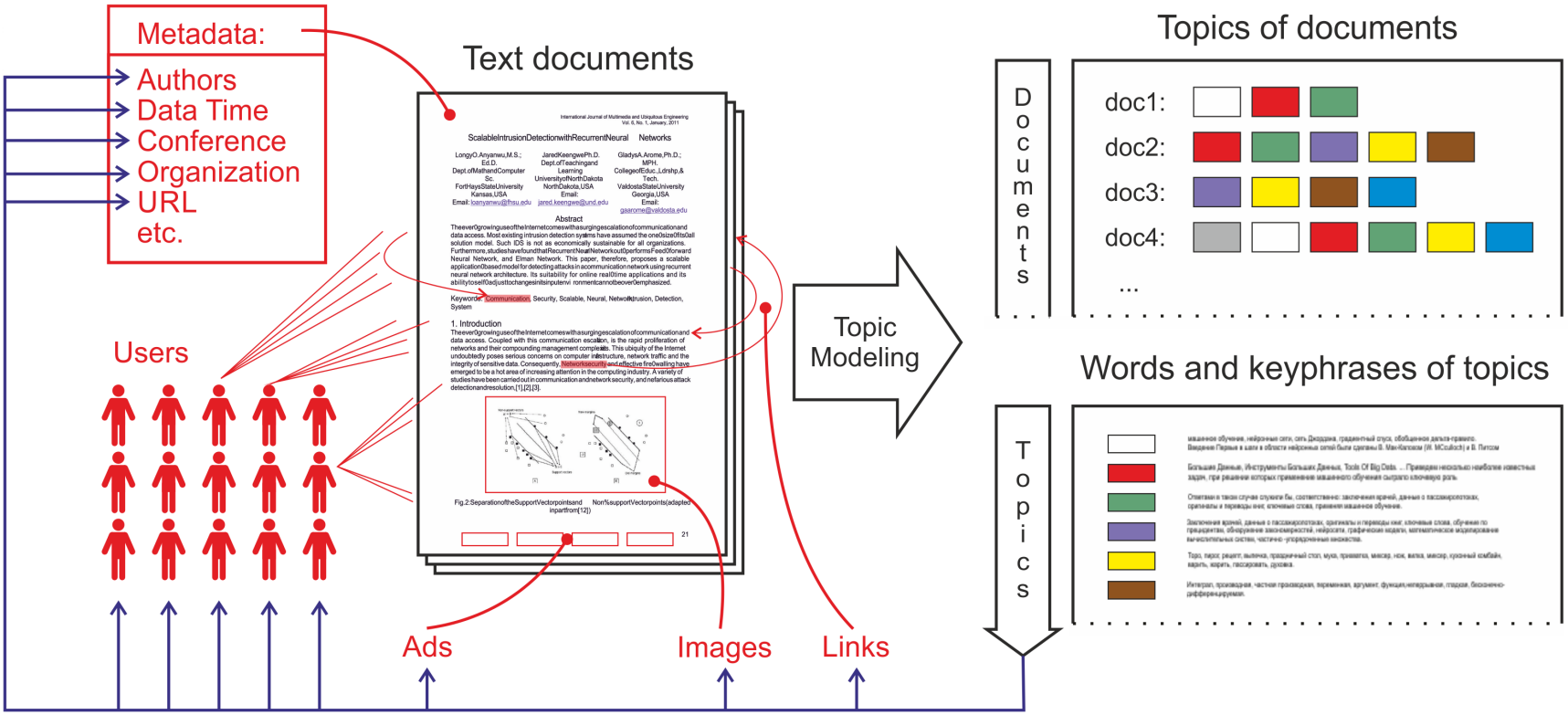
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(\sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases}$$

K.Vorontsov, O.Frei, M.Apishev et al. Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

Пример. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Тематические модели битермов или сочетаемости слов

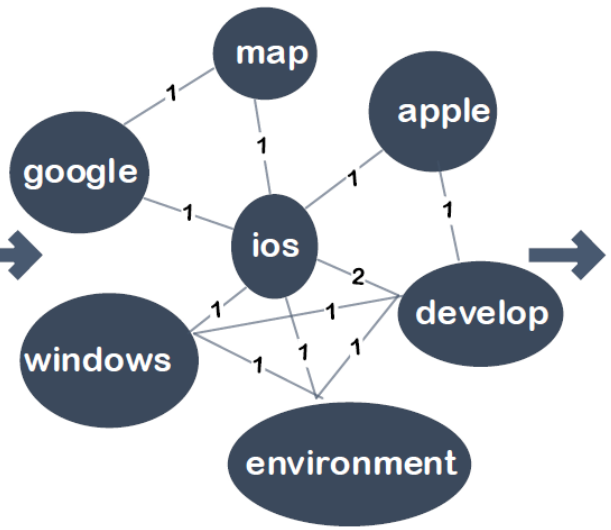
Идея: моделировать не документы, а контексты слов.

d_u — псевдо-документ, объединение всех контекстов слова u.

n_{uw} — число вхождений слова w в псевдо-документ d_u.

Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM (Word Network) и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

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

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} + R(\Phi, \Theta) \to \max_{\Phi,\Theta},$$

где n_{uw} — частота сочетания пары слов (w, u).

В таких моделях интерпретируемо вычитание векторов $p(t|w)$, аналогично word2vec: «король − королева = муж − жена»

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. 2009.

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d\in D} \sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Тематическая модель гиперграфа

V^m — словарь термов модальности $m \in M$
$V = V^1 \sqcup \cdots \sqcup V^M$ — словарь термов всех модальностей
$\Gamma = \langle V, E \rangle$ — гиперграф, система конечных подмножеств V
(d, x) — ребро, $d \in V$ — вершина-контейнер, $x \subset V$

Дано:
E_k — наблюдаемая выборка рёбер (транзакций) типа k,
n_{kdx} — число вхождений ребра (d, x) в выборку E_k.

Найти: тематическую модель рёбер типа k

$$p(x|d) = \sum_{t \in T} \underbrace{p(t|d)}_{\theta_{td}} \prod_{v \in x} \underbrace{p(v|t)}_{\phi_{vt}}$$

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in E_k} n_{kdx} \ln \sum_{t\in T}\theta_{td}\prod_{v\in x}\phi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

E-шаг:
$$p_{tdx} = \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big)$$

M-шаг:
$$\begin{cases} \phi_{vt} = \underset{v\in V^m}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k}\big[v\in x\big] n_{kdx}p_{tdx} + \phi_{vt}\dfrac{\partial R}{\partial \phi_{vt}}\Big) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx}p_{tdx} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Доказательство (по лемме о максимизации на симплексах)

Применим Лемму к log-правдоподобию с регуляризатором R:

$$\phi_{vt} = \underset{v \in V_m}{\text{norm}} \left(\phi_{vt} \sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \frac{\theta_{td}}{p(x|d)} \frac{\partial}{\partial \phi_{vt}} \prod_{u \in x} \phi_{ut} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) =$$

$$= \underset{v \in V_m}{\text{norm}} \left(\sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{kdx} [v \in x] p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(\theta_{td} \sum_{k \in K} \tau_k \sum_{x \in d} n_{kdx} \frac{1}{p(x|d)} \prod_{v \in x} \phi_{vt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) =$$

$$= \underset{t \in T}{\text{norm}} \left(\sum_{k \in K} \sum_{x \in d} \tau_k n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

Примеры задач с транзакционными данными

Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

Гиперграфовые тематические модели языка

Рёбрами гиперграфа могут быть любые подмножества термов, связанные по смыслу и порождаемые общей темой:

- предложение / фраза / синтагма
- ветка синтаксического дерева / именная группа
- факт «объект, субъект, действие»
- пары синонимов, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст комментария и его автор

Модель даёт интерпретируемые тематические эмбединги:

- $p(t|d)$ — каждого контейнера, в частности, документа
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ — каждого терма, в частности, слова
- $p(t|d,x)$ — каждой отдельной транзакции (фразы, факта)

Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Регуляризатор E-шага: $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad \begin{cases} p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \end{cases}$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три леммы

Лемма 1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

Лемма 2. Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw};\quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

Лемма 3. Формулы M-шага:

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

Гипотеза о пост-обработке E-шага

Между E- и M-шагом добавляется обработка матрицы (p_{tdw}) тематических векторов последовательности термов документа:

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \qquad (1)$$

Пост-обработка E-шага позволяет учитывать порядок термов в документе в обход гипотезы «мешка слов».

Гипотеза

Любое «разумное» преобразование $p_{tdw} \to \tilde{p}_{tdw}$ эквивалентно некоторому регуляризатору $R(\Pi(\Phi, \Theta))$.

Открытый вопрос: при каких условиях по заданным p_{tdw} и $\tilde{p}_{tdw}$ возможно подобрать функцию $R(\Pi)$ так, чтобы выполнялось уравнение пост-обработки (1)?

Возможна ли тематизация фрагмента за один проход?

Дано: q — фрагмент текста, Φ — готовая тематическая модель
Найти: $p(t|q)$ — тематический вектор фрагмента текста

Проблемы:

- если текст короткий, то определение $p(t|q)$ не надёжно
- согласование $p(t|q)$ с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ отдельных слов
- согласование $p(t|q)$ с более широким контекстом $d \supset q$

Наводящие соображения:

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

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

- формула полной вероятности:

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

EM-алгоритм для ARTM без матрицы Θ

Максимизация log-правдоподобия при ограничении $\Theta = \Theta(\Phi)$:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big);$$

$$n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw};$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} + \phi_{wt} \sum_{d \in D} \sum_{s \in T} \left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}} \right) \frac{\partial \theta_{sd}}{\partial \phi_{wt}} \right)$$

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 $\Theta(\Phi)$:

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

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} +$$

$$+ \sum_{d,s,u} n_{du} p_{sdu} \frac{\phi_{wt}}{\theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt} \sum_{d,s} \frac{\partial R}{\partial \theta_{sd}} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} =$$

$$= n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \phi_{wt} \sum_{d,s} \left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}} \right) \frac{\partial \theta_{sd}}{\partial \phi_{wt}}$$

■

Частный случай $\theta_{td}(\Phi) = \sum_w p_{wd} \operatorname{norm}_t(\phi_{wt})$

Частные производные: $\frac{\partial \theta_{sd}}{\partial \phi_{wt}} = p_{wd} h_w \left(\delta_{st} - \phi_{ws} h_w\right)$

EM-алгоритм: метод простой итерации для системы уравнений

$$\theta_{td} = \sum_{w \in d} p_{wd} \phi_{wt} h_w; \qquad h_w = \left(\sum_t \phi_{wt}\right)^{-1};$$

$$p_{tdw} = \operatorname*{norm}_{t \in T}\left(\phi_{wt}\theta_{td}\right); \qquad c_{td} = \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}};$$

$$n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \qquad \gamma_{dw} = \sum_{t \in T} \phi_{wt} c_{td};$$

$$p'_{tdw} = p_{tdw} + n_d^{-1} \phi_{wt} h_w (c_{td} - h_w \gamma_{dw});$$

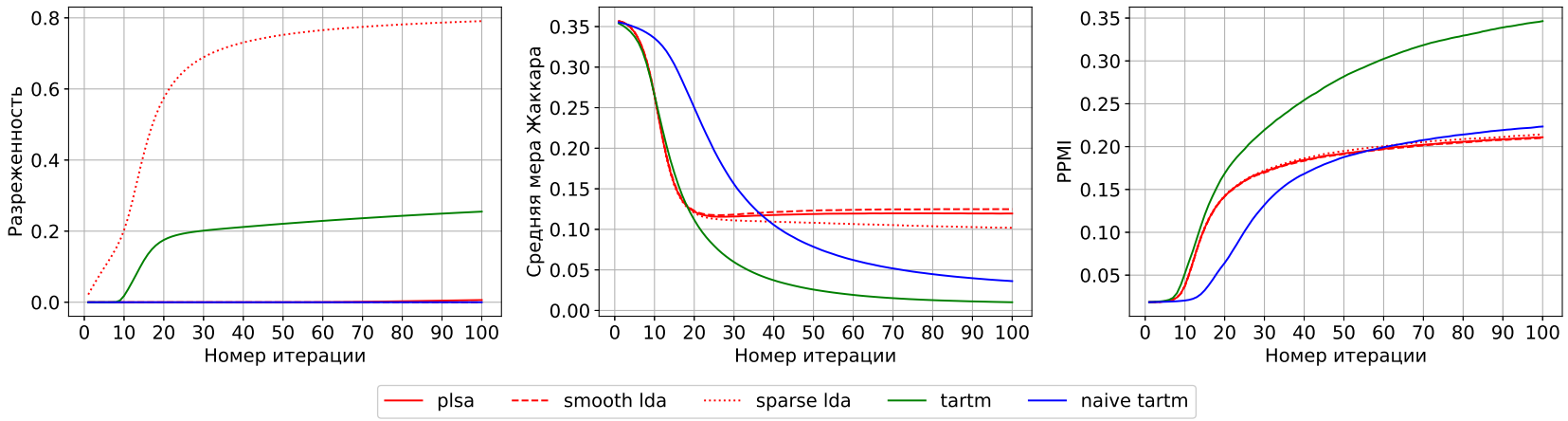
$$\phi_{wt} = \operatorname*{norm}_{w \in W}\left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

E-шаг по-прежнему занимает $O(n_d |T|)$ операций для каждого d

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless 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.

Быстрая векторизация текста за линейное время

Тематический вектор текста $p(t|d)$ вычисляется за один проход усреднением тематических векторов $p(t|w)$ всех слов текста:

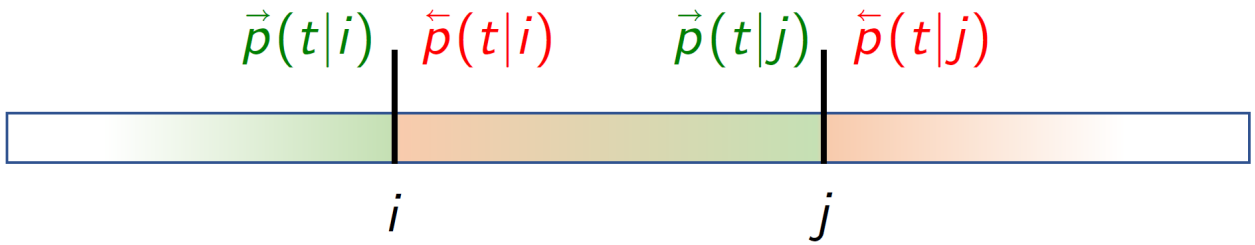
$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i)$$

Тематические векторы локального контекста $p(t|i)$ вычисляются для всех $i = 1, \ldots, n_d$ экспоненциальным скользящим средним за два прохода «слева направо» и «справа налево»:

$$\vec{p}(t|i) = \alpha_i \cdot p(t|w) + (1 - \alpha_i) \cdot \vec{p}(t|i-1)$$
$$\overleftarrow{p}(t|i) = \alpha_i \cdot p(t|w) + (1 - \alpha_i) \cdot \overleftarrow{p}(t|i+1)$$

α_i — коэффициент сглаживания в позиции i;
$\alpha_i \approx \frac{1}{m}$, где m — число усредняемых позиций;
α_i можно умножать на вес (важность, TF-IDF) слова в тексте,
α_i можно увеличивать до 1, если надо забыть контекст.

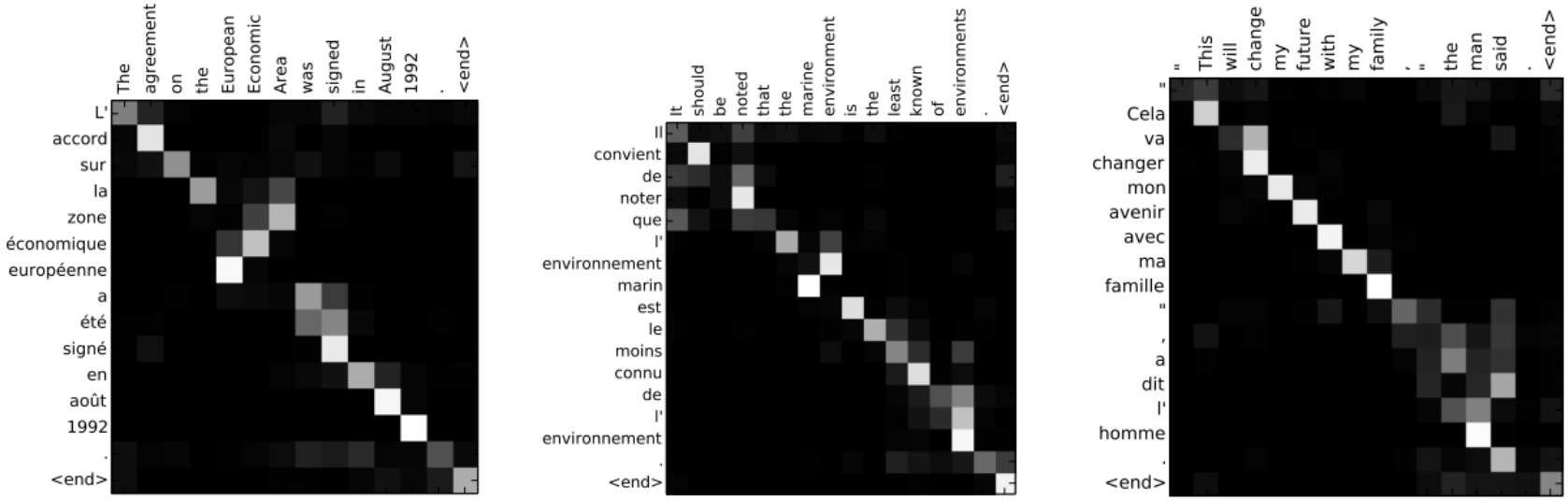
Тематические векторы локального контекста



Двунаправленные тематические векторы определяют:

- $\vec{p}(t|i)$ — тематику левого контекста слова w_i
- $\overleftarrow{p}(t|i)$ — тематику правого контекста слова w_i
- $\frac{1}{2}\left(\vec{p}(t|i) + \overleftarrow{p}(t|i)\right)$ — тематику двустороннего контекста w_i
- $p(t|i\ldots j) = \frac{1}{2}\left(\overleftarrow{p}(t|i) + \vec{p}(t|j)\right)$ — тематику сегмента $[i\ldots j]$
- тематическую однородность сегмента $[i\ldots j]$:
  насколько распределения $\overleftarrow{p}(t|i)$ и $\vec{p}(t|j)$ схожи
- позиции i границ между сегментами:
  насколько распределения $\vec{p}(t|i)$ и $\overleftarrow{p}(t|i)$ не схожи
- короткие и длинные контексты при различных α_i

Модели внимания в машинном переводе



Интерпретируемость моделей внимания:
матрица семантического сходства $A = (\alpha_{ti})$ показывает, на какие слова x_i из входной последовательности модель обращает внимание, когда генерирует слово перевода y_t

---

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015.

Модели внимания на изображениях для генерации описаний

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Засветка показывает, на какие области изображения модель обращает внимание, генерируя слово в описании изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Применения моделей внимания

Преобразование одной последовательности в другую, seq2seq:

- машинный перевод (machine translation)
- ответы на вопросы (question answering)
- ведение диалога (conversational agents)
- суммаризация текста (text summarization)
- описание изображений, аудио, видео (multimedia description)
- распознавание и синтез речи (speech recognition/synthesis)

Обработка последовательности:

- классификация текстовых документов
- выделение и классификация фрагментов текста
- анализ тональности документа / предложений / аспектов

Модель внимания Запрос–Ключ–Значение (Query–Key–Value)

q — *вектор-запрос*, для которого хотим вычислить контекст
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*, сравниваемые с запросом
$V = (v_i, \ldots, v_n)$ — *векторы-значения*, образующие контекст
$a(k_i, q)$ — *оценка релевантности* (сходства) ключа k_i запросу q
c — искомый вектор контекста, релевантный запросу

Модель внимания — 3х-слойная нейросеть, вычисляющая выпуклую комбинацию векторов v_i, релевантных запросу q:

$$c = \text{Attn}(q, K, V) = \sum_i v_i \operatorname*{SoftMax}_i a(k_i, q)$$

$c_t = \text{Attn}(W_q y_{t-1}, W_k X, W_v X)$ — в машинном переводе, где
$X = (x_1, \ldots, x_n)$ — векторы слов входного предложения,
y_{t-1} — предшествующий выходной вектор

Внутреннее внимание или «самовнимание» (self-attention):
$c_i = \text{Attn}(W_q x_i, W_k X, W_v X)$ — частный случай, когда $x_i \in X$

Разновидности функций сходства векторов

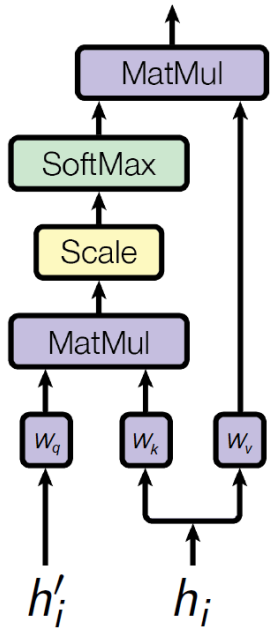
$a(h, h') = h^{\mathsf{T}} h'$ — скалярное произведение
$a(h, h') = h^{\mathsf{T}} W h'$ — с матрицей обучаемых параметров W
$a(h, h') = w^{\mathsf{T}} \operatorname{th}(Uh + Vh')$ — аддитивное внимание с w, U, V

Линейные преобразования векторов query, key, value:

$$a(h_i, h'_{t-1}) = (W_k h_i)^{\mathsf{T}} (W_q h'_{t-1}) / \sqrt{d}$$
$$\alpha_{ti} = \operatorname*{SoftMax}_i a(h_i, h'_{t-1})$$
$$c_t = \sum_i \alpha_{ti} W_v h_i$$

$W_q{}_{d \times \dim(h')}$, $W_k{}_{d \times \dim(h)}$, $W_v{}_{d \times \dim(h)}$ — матрицы коэффициентов обучаемых линейных преобразований в пространство размерности d

Возможно упрощение модели: $W_k \equiv W_v$



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

Многомерное внимание (multi-head attention)

Идея: J разных моделей внимания совместно обучаются выделять различные аспекты входной информации (например, части речи, синтаксис, фразеологизмы):

$$c^j = \text{Attn}(W_q^j q, W_k^j H, W_v^j H), \quad j = 1, \ldots, J$$

Варианты агрегирования выходного вектора:

$c = \frac{1}{J}\sum_{j=1}^{J} c^j$ — усреднение

$c = \left[c^1 \cdots c^J\right]$ — конкатенация

$c = \left[c^1 \cdots c^J\right] W$ — чтобы вернуться к нужной размерности

Регуляризация: чтобы аспекты внимания были максимально различны, строки $J \times n$ матриц A, $\alpha_{ji} = \text{SoftMax}_i\, a(W_k^j h_i, W_q^j q)$, декоррелируются $(\alpha_s^{\mathsf{T}}\alpha_j \to 0)$ и разреживаются $(\alpha_j^{\mathsf{T}}\alpha_j \to 1)$:

$$\|AA^{\mathsf{T}} - I\|^2 \to \min_{\{W_k^j, W_q^j\}}$$

Zhouhan Lin, Y.Bengio et al. A structured self-attentive sentence embedding. 2017.

Трасформер для машинного перевода

Трасформер (transformer) — это нейросетевая архитектура на основе моделей внимания и полносвязных слоёв

Схема преобразований данных в машинном переводе:

- $S = (w_1, \ldots, w_n)$ — слова предложения на входном языке
  $\downarrow$ обучаемая или пред-обученная векторизация слов
- $X = (x_1, \ldots, x_n)$ — эмбединги слов входного предложения
  $\downarrow$ трансформер-кодировщик
- $Z = (z_1, \ldots, z_n)$ — контекстные эмбединги слов
  $\downarrow$ трансформер-декодировщик, похож на кодировщика
- $Y = (y_1, \ldots, y_m)$ — эмбединги слов выходного предложения
  $\downarrow$ генерация слов из построенной языковой модели
- $\tilde{S} = (\tilde{w}_1, \ldots, \tilde{w}_m)$ — слова предложения на выходном языке

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

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \ldots, h_n)$
$d = \dim x_i, p_i, h_i = 512$
$\dim H = 512 \times n$

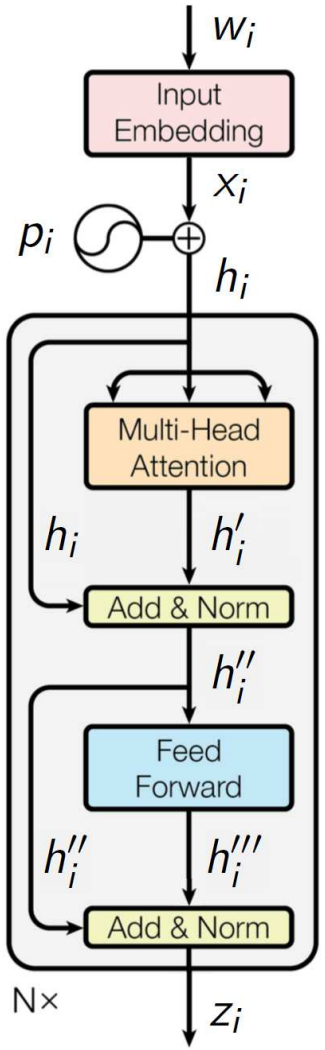
2. Многомерное самовнимание:
$h_i^j = \mathrm{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$
$j = 1, \ldots, J = 8$
$\dim h_i^j = 64$
$\dim W_q^j, W_k^j, W_v^j = 64 \times 512$

3. Конкатенация:
$h'_i = \mathrm{MH}_j(h_i^j) \equiv \left[h_i^1 \cdots h_i^J\right]$
$\dim h'_i = 512$

4. Сквозная связь + нормировка уровня:
$h''_i = \mathrm{LN}(h'_i + h_i; \mu_1, \sigma_1)$
$\dim h''_i, \mu_1, \sigma_1 = 512$

5. Полносвязная 2х-слойная сеть FFN:
$h'''_i = W_2\,\mathrm{ReLU}(W_1 h''_i + b_1) + b_2$
$\dim W_1 = 2048 \times 512$
$\dim W_2 = 512 \times 2048$

6. Сквозная связь + нормировка уровня:
$z_i = \mathrm{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$
$\dim z_i, \mu_2, \sigma_2 = 512$

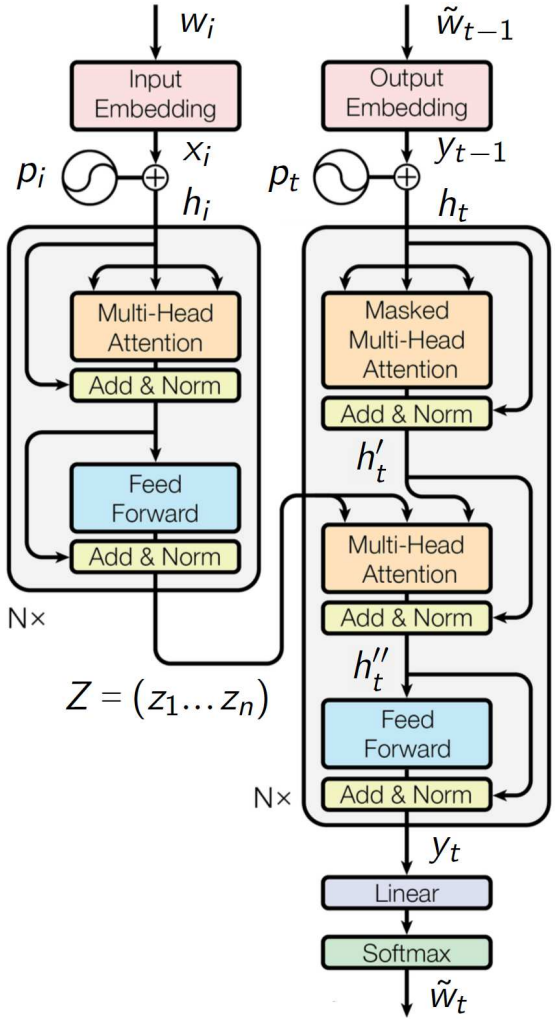


Архитектура трансформера декодировщика

Авторегрессионный синтез последовательности:
$y_0 = \langle \mathtt{BOS} \rangle$ — эмбеддинг символа начала;
для всех $t = 1, 2, \ldots$:

1. Маскирование «данных из будущего»:
   $h_t = y_{t-1} + p_t; \quad H_t = (h_1, \ldots, h_t)$
2. Многомерное самовнимание:
   $h'_t = \mathrm{LN} \circ \mathrm{MH}_j \circ \mathrm{Attn}(W^j_q h_t, W^j_k H_t, W^j_v H_t)$
3. Многомерное внимание на кодировку Z:
   $h''_t = \mathrm{LN} \circ \mathrm{MH}_j \circ \mathrm{Attn}(\tilde{W}^j_q h'_t, \tilde{W}^j_k Z, \tilde{W}^j_v Z)$
4. Двухслойная полносвязная сеть:
   $y_t = \mathrm{LN} \circ \mathrm{FFN}(h''_t)$
5. Линейный предсказывающий слой:
   $p(\tilde{w}|t) = \underset{\tilde{w}}{\mathrm{SoftMax}}(W_y y_t + b_y)$

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



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

BERT (Bidirectional Encoder Representations from Transformers)

Трансформер BERT — это кодировщик без декодировщика, предобучаемый на большой текстовой коллекции для решения широкого класса задач NLP

Схема преобразования данных в задачах NLP:

- $S = (w_1, \ldots, w_n)$ — токены предложения входного текста
  $\downarrow$ обучение эмбедингов вместе с трансформером
- $X = (x_1, \ldots, x_n)$ — эмбединги токенов входного предложения
  $\downarrow$ трансформер кодировщика
- $Z = (z_1, \ldots, z_n)$ — трансформированные эмбединги
  $\downarrow$ дообучение на конкретную задачу
- Y — выходной текст / разметка / 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.

Критерии обучения трансформеров

- **Машинный перевод:** максимизация правдоподобия слов перевода $\tilde{w}_t$ по выборке пар предложений «S, перевод $\tilde{S}$»:

$$\sum_{(S,\tilde{S})} \sum_{\tilde{w}_t \in \tilde{S}} \ln p(\tilde{w}_t | t, S, W) \to \max_W$$

- **BERT MLM (masked language modeling):** предсказание пропущенных слов по локальному контексту
- **BERT NSP (next sentence prediction):** предсказание, следуют ли два предложения друг за другом
- **Fine-tuning:** дообучение трансформера $Z(S, W)$ на задаче с моделью $f(Z(S, W), W_f)$, выборкой $\{S\}$ и $\mathscr{L}(S, f) \to \max$
- **Multi-task learning:** дообучение на наборе задач $\{t\}$ с моделями $f_t(Z(S, W), W_t)$, выборками $\{S\}_t$, по сумме критериев $\sum_t \lambda_t \sum_S \mathscr{L}_t(S, f_t) \to \max$

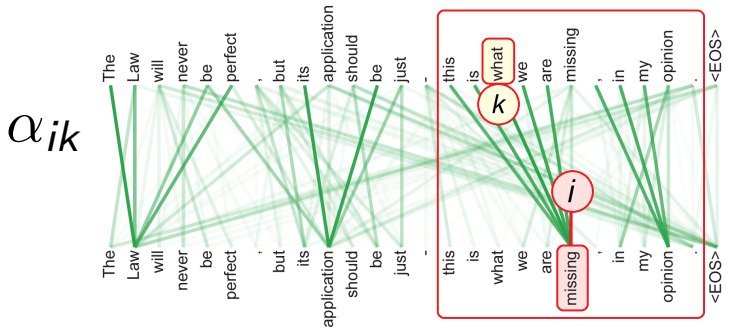
Тематическая модель внимания (self-attention)

Внимание в нейросетевых моделях языка:

x_i — эмбеддинги (размерности T) термов w_i, $i = 1, \ldots, n$

$\alpha_{ik} = \text{norm}_k \langle x_i, x_k \rangle$ — важность терма w_k в контексте терма w_i

$c_i = \sum_k V x_k \alpha_{ik}$ — эмбеддинг контекста терма w_i с обучаемой $\underset{T \times T}{V}$



Аналогичная конструкция в тематической модели:

$$\underbrace{p(t|i)}_{c_i} = \sum_k \sum_{t' \in T} \underbrace{p(t|t')}_{v_{tt'}} \underbrace{p(t'|w_k)}_{x_k} \underset{k}{\text{norm}} \Big\langle \underbrace{p(t''|w_k)}_{x_k}, \underbrace{p(t''|w_i)}_{x_i} \Big\rangle$$

Vaswani et al. Attention is all you need. 2017.
Dichao Hu. An Introductory Survey on Attention Mechanisms in NLP Problems. 2018.

Резюме

- Открытая проблема — «объединить лучшее от двух миров»:
  — покоординатную интерпретируемость BTM
  — глубину и выразительность нейросетевых моделей языка
- Что для этого уже есть:
  — тематические векторы слов-в-контексте $p(t|d, w_i)$
  — лемма о максимизации на симплексах
  — возможность вычислять градиенты методом BackProp
  — одно(двух)проходные алгоритмы тематизации текста
  — реализация ARTM в библиотеке BigARTM
- Чего не хватает:
  — уверенности, что смысл определяется тематикой
  — реализации и экспериментов

К.Воронцов. Вероятностное тематическое моделирование: теория регуляризации и проект BigARTM. 2022.
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf