

Многоязычное тематическое моделирование

Марина Дударенко
ВМК МГУ

научный семинар • ШАД Яндекс • 30 сентября 2014

Содержание

- 1 **Задача многоязычного тематического моделирования**
 - Постановка задачи
 - Источники многоязычной информации
 - Связь с мультимодальными моделями
- 2 **Регуляризация многоязычных тематических моделей**
 - Параллельные коллекции
 - Сравнимые коллекции
 - Двужычные словари
- 3 **Эксперименты**
 - Кросс-язычный поиск
 - Тематический контекст переводов
 - Выводы

От однопозначности ...

Дано:

W — словарь, множество терминов

D — множество текстовых документов

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Найти:

$p(t|d)$ — к каким темам t относится каждый документ d

$p(w|t)$ — какими терминами w определяется каждая тема t

... К МНОГОЯЗЫЧНОСТИ

Дано:

L — множество языков

W_ℓ — словарь, множество терминов языка $\ell \in L$

D_ℓ — множество текстовых документов на языке $\ell \in L$

$D = \cup_{\ell \in L} D_\ell$ — многоязычная коллекция

n_{dw} — сколько раз термин $w \in W_\ell$ встретился в документе $d \in D_\ell$

Найти:

$p(t|d)$ — к каким темам t относится каждый документ d

$p(w|t, \ell)$ — какими терминами w определяется каждая тема t
в языке ℓ

Задача максимизации правдоподобия

Задача: найти максимум правдоподобия

$$L(\Phi, \Theta) = \sum_{\ell \in L} \sum_{d \in D_{\ell}} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}^{\ell} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt}^{\ell} \geq 0; \quad \sum_{w \in W_{\ell}} \phi_{wt}^{\ell} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1$$

Проблема:

без дополнительной информации о связи языков полученные многоязычные тематические распределения не согласованы

Решение:

аддитивная регуляризация логарифма правдоподобия
с использованием источников многоязычной информации

Дополнительная многоязычная информация

- позволяет учитывать наличие документов–переводов
- позволяет связывать слова–переводы из различных языков

Виды дополнительной многоязычной информации:

- параллельные коллекции, состоящие из документов–переводов
✓ **Europarl**
- сравнимые коллекции, состоящие из документов, содержащих схожие идеи
✓ **Wikipedia**
- базы знаний
✓ **WordNet, GermaNet, MENTA**
- двуязычные словари

Многоязычность и мультимодальность

В случае параллельной или сравнимой коллекции многоязычная модель эквивалентна мультимодальной модели:

- число модальностей равно количеству языков L ;
- модальность X^j представляется словарем W_j языка j ;
- вероятностное пространство $D \times T \times W$,
 $W = W_1 \times \dots \times W_L$;
- каждый документ d состоит из токенов $w_1, \dots, w_{n_d} \in W$.

Параллельная коллекция: равенство профилей документов

Гипотеза: у параллельных документов тематические профили θ_{td} равны между собой

Приравниваем покомпонентно профили соответствующих документов

$\Pi(d) \subseteq D$ — множество документов–переводов документа d

$$L(\Phi, \Theta) = \sum_{\ell \in L} \sum_{d \in D_\ell} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}^\ell \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

при ограничениях $\theta_{td} = \theta_{td'}$ для всех переводов $d' \in \Pi(d)$

Сравнимая коллекция: близость профилей документов

Гипотеза: у сравнимых документов тематические профили θ_{td} похожи, но не равны между собой

Приближаем покомпонентно профили соответствующих документов

$\Pi(d) \subseteq D$ — множество документов, связанных с документом d («пересказы»)

$$R(\Theta) = \tau \sum_{d \in D} \sum_{d' \in \Pi(d)} \sum_{t \in T} \theta_{td} \theta_{td'} \rightarrow \max_{\Phi, \Theta}$$

Словарь–1: близость тематических профилей слов

Гипотеза: у слова $w \in W_\ell$ и его перевода $u \in W_k$
тематические профили близки: $p(t|w, \ell) \simeq p(t|u, k)$

Минимизируем KL-дивергенцию между оценками тематических профилей слов–переводов и модельным тематическим профилем выбранного слова: $KL(\hat{p}(t|u, k) || p(t|w, \ell)) \rightarrow \min$

$\Pi_k(w) \subseteq W_k$ — множество переводов слова w в языке k

$$R(\Phi) = \tau \sum_{\ell \in L} \sum_{w \in W_\ell} \sum_{\substack{k \in L \\ k \neq \ell}} \sum_{u \in \Pi_k(w)} \sum_{t \in T} n_{ut} \ln \phi_{wt}^\ell \rightarrow \max_{\Phi, \Theta}$$

Словарь-2: настраиваемая матрица переводов для каждой темы

$\pi_{uwt}^{kl} = p(u|w, t)$, $u \in W_k$, $w \in W_\ell$ — матрица вероятностей перевода слов из языка ℓ на язык k в теме t

Гипотеза: тематическое распределение темы t в коллекции близко к распределению, которое получается путем перевода

Минимизируем KL-дивергенцию между оценками тематических распределений и их выражениями через матрицу переводов:

$$KL(\hat{p}(u|t, k) || \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{kl} \phi_{wt}^\ell)$$

$$R(\Phi) = \tau \sum_{\ell \in L} \sum_{\substack{k \in L \\ k \neq \ell}} \sum_{t \in T} \sum_{u \in W_k} n_{ut} \ln \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{kl} \phi_{wt}^\ell \rightarrow \max_{\Phi, \Theta}$$

Данные

Коллекции:

- **Math** — параллельная коллекция математических статей на русском и английском языках
 - $|D_{ru}| = |D_{en}| = 154$ статей;
 - словарь $|W_{ru}| = 4574$, $|W_{en}| = 6245$.
- **Wiki** — сравнимая коллекция статей из категории «Математика» и связанных с ней категорий русской и английской Википедии, имеющих ссылки интервики друг на друга
 - $|D_{ru}| = |D_{en}| = 586$ статей;
 - словарь $|W_{ru}| = 19305$, $|W_{en}| = 23413$.

Двуязычный словарь: получен из русско–английского электронного словаря путем извлечения всех однословных переводов, включает около 82 000 пар переводов.

Оценка качества модели

Кросс-язычный поиск — запросом является документ на одном языке, а поиск производится среди документов другого языка.

Требуется найти документы на другом языке, максимально похожие на запрос по тематическому профилю.

Мера качества:

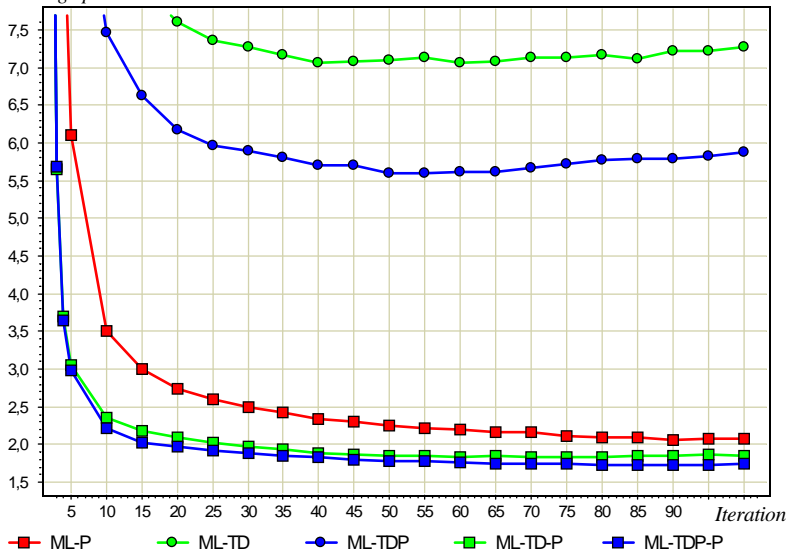
AP (Average Position) — позиция настоящего документа–перевода в ранжированной поисковой выдаче для запроса q , усредненная по множеству запросов Q :

$$AP(Q) = \frac{1}{|Q|} \sum_{q \in Q} rank(q).$$

Сравнение методов

Коллекция	Модель	25 тем	50 тем	100 тем
Math	ML-P	1.051	1.002	1.009
	ML-TD	5.762	4.848	1.318
	ML-TDP	1.646	1.091	1.006
	ML-TD-P	1.024	1.018	1.000
	ML-TDP-P	1.028	1.006	1.000
	Mallet	1.066	1.041	1.006
	Vector	1.000		
Wiki	ML-P	4.117	2.826	2.082
	ML-TD	41.018	13.290	7.269
	ML-TDP	32.160	9.945	5.876
	ML-TD-P	4.191	2.659	1.860
	ML-TDP-P	3.578	2.265	1.739
	Mallet	4.596	3.489	2.693
	Vector	4.898		

Average position



Комбинирование методов

Модель	25 тем	50 тем	100 тем
ML-P	4.117	2.826	2.082
ML-TDP	32.160	9.945	5.876
ML-TDP-P0.1	12.315	8.045	4.850
ML-TDP-P0.2	9.737	6.254	3.794
ML-TDP-P0.3	7.487	4.895	3.567
ML-TDP-P0.4	6.098	4.379	3.030
ML-TDP-P0.5	5.461	3.508	2.675
ML-TDP-P0.6	4.767	3.349	2.637
ML-TDP-P0.7	4.052	2.977	2.150
ML-TDP-P0.8	4.027	2.674	2.058
ML-TDP-P0.9	3.701	2.555	1.886
ML-TDP-P	3.578	2.265	1.739

Различие переводов «сумма»–«sum» и «сумма»–«total»

Тема 6		Тема 12		Тема 20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer
Тема 5		Тема 19		Тема 22	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Основные выводы

- 1 Наилучшие результаты достигаются при комбинировании методов.
- 2 В методах со словарем использование матрицы переводов лучше, чем сближение тематических профилей слов-переводов.
- 3 Использование словаря позволяет уменьшить число парных документов на обучении без ухудшения качества поиска.

Спасибо за внимание.