Митрофанова О.А.
Mitrofanova O.A.

ПОИСК И РАНЖИРОВАНИЕ ТЕКСТОВ В СПЕЦИАЛЬНОМ КОРПУСЕ НА ОСНОВЕ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ[1]

SEARCH AND RANKING OF TEXTS IN A SPECIALIZED CORPUS BASED ON TOPIC MODELING

Аннотация. В докладе рассматриваются результаты адаптации процедуры мультимодального тематического моделирования в решении задачи поиска и ранжирования документов в русскоязычном корпусе текстов по корпусной и компьютерной лингвистике кафедры математической лингвистики СПбГУ.

Abstract. The report discusses results of adapting the procedure of multimodal topic modeling in solving the problem of search and ranking documents in the Russian corpus of texts on Corpus and Computational Linguistics, Department of Mathematical Linguistics, SPbSU.

Ключевые слова. тематическое моделирование, специальный корпус текстов, русский язык

Keywords. Topic modeling, specialized text corpus, Russian language

Введение

В области применения машинного обучения для решения лингвистических задач особое место занимает тематическое моделирование (далее ТМ), рассматриваемое как разновидность нечеткой кластеризации [Daud et al. 2010; Воронцов 2023]. Тематическая модель – способ представления корпуса текстов в виде набора тем, объединяющих дистрибутивно близкие слова-

[1] Исследование проводится в рамках НИП СПбГУ № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта» и гранта РНФ № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики».

тематизаторы, при этом и темы, и слова-тематизаторы соотнесены с документами корпуса. В алгоритмическом аспекте ТМ представляет собой снижение размерности исходных данных. В вычислительном аспекте ТМ опирается на преобразования матрично-векторного представления корпуса текстов, восстановление компонент смеси вероятностных распределений, описывающих процесс порождения текста. Лингвистический аспект ТМ состоит в интерпретации скрытых переменных, выявляемых в результате обучения модели. Цель данного исследования – доказать целесообразность применения ТМ для организации поиска и ранжирования информации в развивающихся корпусных ресурсах.

ТМ как процедура анализа лингвистических данных

Задача ТМ может быть выполнена с применением различных подходов, обычно выбор производится из групп алгебраических методов (LSA, NMF и т.д.) и вероятностных (pLSA, LDA, LDP, HMM-TM и т.д.). В ходе ТМ не важен ни порядок терминов в документе, ни порядок документов в коллекции: используется подход «мешок слов». Матрица [*документы* x *слова*] представляется в виде произведения матриц меньшей размерности [*слова* x *темы*] x [*темы* x *документы*]. Специфика алгебраических моделей определяется матричными преобразованиями, тогда как в случае построения вероятностных моделей производится переход от матриц к распределениям, производится стохастическое матричное разложение. Тема задается дискретным распределением на множестве слов, документ – дискретным распределением на множестве тем. Коллекция документов представляется как набор терминов, выбранных случайно и независимо из смеси распределений. Тем самым, ТМ состоит в восстановлении компонент смеси по выборке. Вероятностная модель порождения данных формально задается в следующем виде: $p(w|d) = \Sigma p(t|d)\ p(w|t)$, где $p(w|d)$ –

наблюдаемая вероятность появления термина *w* в документе *d*; *p(w|t)* – неизвестная вероятность появления термина *w* в теме *t*; *p(t|d)* – неизвестная вероятность появления темы *t* в документе *d*.

Вероятностное ТМ относится к числу методов машинного обучения с привлечением байесовских методов, ср. EM-алгоритм, предположение о порождении тем из распределения Дирихле и т.д. Мультимодальность ТМ обеспечивается введением дополнительных параметров в тематическую модель: учет коллокаций или конструкций для *n*-граммных моделей, обобщение тем с помощью меток для размеченных моделей, учет авторства документов для автор-тематических моделей, учет времени создания документа или описываемого события для динамических моделей, выявление иерархии тем для иерархических моделей, управляемый процесс для моделей, строящихся по заранее заданным ключевым словам и т.д. Помимо стандартных методов ТМ развиваются комбинированные, которые предполагают интеграцию с нейросетевыми дистрибутивно-семантическими моделями: со статическими моделями Word2Vec (LDA2Vec, Top2Vec и т.д.) и контекстуализированными моделями ELMo и BERT (T-BERT, KITTY и т.д.).

В нашем исследовании были применены алгоритмы LSA, nmf, LDA, BERTopic в библиотеках для ТМ MALLET[2], Scikit-learn[3], genism[4], tomotopy[5], BERTopic[6], в результате чего было осуществлено построение серии стандартных и мультимодальных моделей с разными параметрами, использование дистрибутивных методов для разметки ключевых

[2] URL: https://github.com/senderle/topic-modeling-tool
[3] URL: https://scikit-learn.org/stable/index.html
[4] URL: https://radimrehurek.com/gensim/
[5] URL: https://github.com/bab2min/tomotopy?ysclid=lapldlogwp49637756
[6] URL: https://github.com/MaartenGr/BERTopic ; URL: https://pypi.org/project/bertopic/

выражений и назначения меток тем, апробация процедуры в работе с неструктурированным корпусом текстов, находящимся в процессе разработки.

ТМ для поиска и ранжирования документов в корпусе

Тематическая модель отражает основные темы корпуса, группирует слова и документы корпуса внутри тем, тем самым, формирует внутреннюю структуру корпуса и представляет ее сжато в виде единиц текста, которые могут использоваться как информационный портрет. Сама по себе тема может быть охарактеризована как нечеткая переменная: нечетким является не только распределение слов по темам и тем по документам, но и отношения между словами-тематизаторами внутри тем, где наличествуют и парадигматическое, и синтагматическое, и эпидигматическое измерения.

Набор тем, соотносящийся с нечеткими кластерами документов корпуса, представляет собой результат семантической компрессии и в этом смысле может рассматриваться наряду с классификаторами, рубрикаторами, индексами, словарями, наборами ключевых слов, рефератами, аннотациями и т.д. как вторичный текст [Леонтьева 2006; Ягунова 2008]. Наряду с ключевыми выражениями, которые можно считать смысловыми опорами [Филиппов 2003], текстами-примитивами [Сахарный, Сибирский, Штерн 1984], темы корпуса создают тот семантический каркас, на котором можно восстановить содержание исходного текста в результате процессов перифразирования. Набор ключевых слов как текст-примитив отражает основное содержание исходного текста с сохранением цельности, но с потерей связности. Тематическая модель – также результат

векторов в векторной модели корпуса. Различия между процедурами ТМ и выделения ключевых выражений заключается в степени полноты представления содержания текста и в уровне его обобщения.

Особый интерес для исследователей в области компьютерной и корпусной лингвистики представляет специальный текст – это «текст, основное содержание которого составляет то или иное профессиональное знание…» [Герд 2011: 21]. «СКТ даёт специалисту самое главное – термины в их профессиональном конкретном окружении, что тот или иной автор имеет в виду под данным термином, какое понятие за ним стоит.» [Герд 2006: 92]. «Его [СКТ] основная задача – чисто информационная, например, смотрите: слово афазия зафиксировано в 150-ти документах, в 300-х контекстах, в таких-то тематических рубриках. Смотрите, анализируйте, делайте выводы.» [Герд 2006: 93].

Адаптация процедуры ТМ к решению задачи поиска и ранжирования документов в специальных корпусах текстов требует учета следующих факторов. Специальный корпус связан с особой предметной областью, его терминологичность накладывает ограничения на использование лексико-грамматических конструкций (структура терминосочетаний), парадигматических (приоритет

дополнительными источником информации о терминологической лексике, о значимости терминов в системе и об их ранжировании при составлении логико-понятийных схем предметных областей и формальных онтологий.

ТМ корпуса текстов по корпусной и компьютерной лингвистике кафедры математической лингвистики СПбГУ

С 2002 года на кафедре математической лингвистики СПбГУ под руководством В.П.Захарова ведется разработка корпусного ресурса, содержащего материалы конференций и тексты научных изданий кафедры. Проведены эксперименты по автоматическому выделению терминологической лексики, кластеризации терминов и построению формальной онтологии специального корпуса [Митрофанова, Захаров 2009; Виноградова, Митрофанова 2008, Вионградова, Митрофанова, Паничева 2007]. Корпус

построение модели BERTopic. Были выбраны следующие гиперпараметры обучения моделей: фильтрация словаря min_dif=1,3,5, max_dif=0.8…0.99, число итераций 100…400, число тем 10…30, объем тем для выдачи 10…20, выбор оптимального числа тем по когерентности (max c_v ≈ 0,39, topic_num = 10), визуализация в библиотеках *pyLDAvis*[7] (t-SNE) и *matplotlib*[8]. Пример выдачи представлен в таблицах 1−3. Применение BERTopic позволило дифференцировать две группы тем – специальные и фоновые. Специальные темы касаются собственно корпусной и компью-

al. 2018]. Ключевые выражения включают уни-, би- и триграммы, которые могут быть вложены друг в друга, например:

	информационный_поиск проект возможность ресурс учебный_корпус **создание** являться материал	деятельность и ее инструменты. Создание и использование *информационных ресурсов* в *научных проектах*. Образовательные *ресурсы* и *научное* творчество.
5	электронный **государственный информационный** орган развитие власть правительство услуга **открытый** гражданин	*Электронное правительство* и информационные технологии. Государственные органы и их роль в обеспечении открытости и доступности информации. Развитие *электронных государственных* услуг и участие *граждан* в этом процессе.
6	**поиск** запрос данные **результат** система информация поисковая_система база **пользователь** веб	Компоненты *поисковой системы*. *Информационные* ресурсы. *Пользователи* и интернет.
7	**социальная_сеть** интернет сеть новый **политический** являться **пользователь** человек пространство современный	*Социальные сети и интернет. Политические* аспекты. Современное *пространство*.
8	текст слово корпус работа **результат** словарь значение русский_язык **анализ метод**	Лингвистический анализ *текста. Работа* с данными. Лексикография.
9	предлог слово глагол семантический словосочетание **анализ** предложная_конструкция являться **синтаксический значение**	Лексический *анализ. Синтаксический анализ. Семантический анализ*.
10	**пользователь система электронный** свободный лицензия голосование дать использование работа являться	*Электронная система. Пользовательская работа***.** *Свободная лицензия.*

С каждой из тем ассоциируется ранжированный список статей, см. таблицу 2. Например, для темы 6 это 46 статей с весом тематизаторов свыше 100, 52 статьи с весом тематизаторов от 10

до 100, 36 статей от 1 до 10, в остальных статьях корпуса тематизаторы темы 6 не встретились.

По каждому из документов возможен дальнейший поиск и оценка вклада тем модели в содержание статьи. В таблице 3 приведено распределение тем в документе *Mochalova_IMS_2018_rus_lem.txt.*

Таблица 2. *Ранжированный список статей, относящихся к теме 6 ««поиск запрос данные результат система информация поисковая_система база пользователь веб».*

№	Число тематизаторов	Тексты
1.	1924	Mochalova_IMS_2018_rus_lem.txt
2.	1788	Lyapin_IMS_2013_rus_lem.txt
3.	1627	Soms_IMS_2014_rus_lem.txt
4.	1576	Lyapin_IMS_2014_rus_lem.txt
5.	1429	Pijavskij_IMS_2013_rus_lem.txt
6.	1187	Salin_IMS_2015_rus_lem.txt
7.	1030	Serebriakov_IMS_2014_rus_lem.txt
8.	995	Korablinov_IMS_2020_rus_lem.txt
9.	829	Smirnov_IMS_2014_rus_lem.txt
10.	827	Voroshilov_IMS_2013_rus_lem.txt
…	…	…

Таблица 3. Распределение тем в документе *Mochalova_IMS_2018_rus_lem.txt.*

№	Вклад темы в документ (%)	Темы
1.	79	поиск запрос данные результат система информация поисковая_система база пользователь веб
2.	14	текст слово корпус работа результат словарь значение русский_язык анализ метод
3.	5	понятие являться модель связь отношение знание система определение область семантический
…	0	(остальные темы)

Заключение

В данном докладе предложена адаптация процедуры ТМ к решению задачи поиска и ранжирования документов в неструктурированных корпусах текстов. Представление текста в виде набора тем, которые учитывают как униграммы, так и би- и триграммы, а также допускают обобщение с помощью меток, следует рассматривать как результат семантической компрессии, позволяющий регулировать полноту представления содержания текста и степень его обобщения с помощью гиперпараметров тематического моделирования. Тематическая модель текста занимает промежуточное положение между вариантами семантической компрессии (набор ключевых выражений, аннотация и т.д.) и логико-понятийными моделями (например, формальная онтология). Отличие прежде всего в нечеткой природе тематических моделей, это модели с допустимой неопределенностью. Скрытыми переменными могут быть как сами темы, так и семантические отношения внутри корпуса.

Дальнейшее развитие проекта связано с пополнением корпуса и корректировкой построенных моделей, это обеспечит достаточные условия для проведения экспертной оценки результатов ТМ в дополнение к количественным оценкам перплексии и когерентности. С увеличением числа авторов и расширением хронологических рамок корпуса станет возможно построение автор-тематической и динамической моделей, которые позволят автоматически формировать тематические портреты исследователей и отслеживать хронологические изменения в тематике их статей.

Литература

1. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // Proceedings of Frontiers of Computer Science in China, 2010. P. 280−301.