

ОТЗЫВ

официального оппонента кандидата физико-математических наук

Тутубалиной Елены Викторовны

на диссертационную работу

Потапенко Анны Александровны

“Семантические векторные представления текста

на основе вероятностного тематического моделирования”,

представленную на соискание ученой степени

кандидата физико-математических наук по специальности

05.13.17 – Теоретические основы информатики

Актуальность темы

В диссертационной работе А.А. Потапенко исследуется задача построения интерпретируемых разреженных векторных представлений текста. Последние достижения в области векторных представлений слов фактически превратили этот подход в подход по умолчанию в современной интеллектуальной обработке естественного языка. Однако главным недостатком большинства моделей, включая самые популярные модели семейства word2vec, является сложная интерпретируемость: координаты вектора очень трудно трактовать как определенные аспекты смысла без применения специальных методов. Альтернативный подход заключается в применении тематического моделирования, активно развивающегося с конца 90-х годов. Тематические модели позволяют характеризовать слова и документы удобным для понимания способом с помощью вероятностных распределений.

Диссертация А.А. Потапенко посвящена исследованию применимости подхода вероятностного тематического моделирования для получения семантических векторных представлений слов и текстов. Применимость данного подхода к задаче определения семантической близости слов является мало изученной. В диссертации предлагается аддитивно регуляризованная тематическая модель, позволяющая достичь высокой разреженности, различности и интерпретируемости предметных тем. Тема диссертации является актуальной, а модели и алгоритмы, предложенные в работе, являются ключевыми

Основные положения и результаты работы

Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы, содержащего 143 наименования. Общий объем диссертации — 147 страниц.

Во **введении** обосновывается актуальность исследований, формулируется цель, ставятся задачи работы, изложены научная новизна и практическая значимость представляемой работы.

В **первой главе** описываются основные принципы дистрибутивной семантики, типы семантической близости слов, ключевые этапы обработки текстов, представлен обзор математических моделей векторных представлений, описываются терминологические особенности в области векторных представлений слов. Сформулирована теорема о необходимых условиях локального экстремума задачи обучения модели PLSA (Probabilistic Latent Semantic Analysis), позволяющая интерпретировать традиционно используемый для нее ЕМ-алгоритм (Expectation-Maximization) как метод простой итерации.

Необходимо отметить, что А.А. Потапенко выполнила большой объем работы при подготовке обзора по моделям построения низкоразмерных векторов слов, излагая собранные сведения в едином формализме без разделения моделей на частотные и предсказывающие типы, что часто свойственно многим обзорным статьям.

Во **второй главе** рассматриваются две классические тематические модели PLSA и LDA, способы обучения моделей с помощью сэмплирования Гиббса и вариационного вывода, приводится описание ЕМ-алгоритма и его применение для максимизации правдоподобия в модели PLSA и максимизации апостериорной вероятности в модели LDA. Сформулировано и доказано более общее утверждение о необходимых условиях задачи обучения аддитивно регуляризованных тематических моделей ARTM. На основе данной теоремы предложена и реализована модификация ЕМ-алгоритма для обучения данных моделей. На основании проведенного анализа делается заключение о том, что недостатками существующих алгоритмов обучения моделей в рамках байесовского подхода являются сложность вывода алгоритмов для новых моделей и сложность комбинирования дополнительных требований.

В **третьей главе** исследуются применение семейства ЕМ-подобных методов для построения тематических моделей. Автором предложен обобщенный ЕМ-алгоритм, позволяющий комбинировать известные тематические модели, обеспечивая контроль

эквистиками сглаживания, сэмплирования, частого обновления параметров, робастности и разреживания на двух коллекциях на русском и английском языках. Проведенные эксперименты показывают, что робастные алгоритмы с разреживанием являются лучшими по внутреннему критерию качества модели (перплексии), при сравнении на больших коллекциях перплексии PLSA и LDA практически не различаются, а принудительное разреживание в робастных моделях PLSA позволяет получать до 99% нулей в распределениях без ухудшения контрольной перплексии. Это свидетельствует о практической необходимости сочетать разнородные требования в рамках одной математической модели.

Четвертая глава посвящена аддитивной регуляризации тематических моделей, основанный на максимизации взвешенной суммы критериев регуляризации. Автором предложена тематическая модель фоновых и предметных тем, позволяющая увеличивать интерпретируемость тем с помощью комбинации трех регуляризаторов. Модель разбивает множество тем на два подмножества: предметные темы содержат термины предметных областей, а фоновые темы — слова общей лексики. Автором формулируется методология многокритериального оценивания качества модели и экспериментального подбора оптимальной траектории регуляризации. Проведена оценка качества построенной тематической модели на коллекции научных статей на английском языке. Результаты экспериментов подтверждают, что комбинирование регуляризаторов в предложенной модели позволяет улучшить чистоту и когерентность тем при сравнении со стандартными моделями PLSA и LDA.

В **пятой главе** описываются разработанные в рамках диссертационной работы модели вероятностных тематических представлений слов для задачи определения семантической близости. Сформулирована и доказана теорема, позволяющая формализовать гипотезу дистрибутивной семантики в рамках подхода аддитивной регуляризации тематических моделей. На ее основе обучена новая тематическая модель PWE векторных представлений слов. Модель PWE использует для обучения информацию о локальной со-встречаемости слов, как и самая популярная модель Skip-Gram (SGNS), предложенная Томашем Миколовым с соавторами. В отличие от модели SGNS параметры модели PWE удовлетворяют ограничениям неотрицительности и нормировки. Эксперименты на корпусе английского языка подтверждают, что предложенный алгоритм сохраняет информацию о семантической близости слов и обладающих интерпретируемыми компонентами. Для

хранение параметров модели в виде матрицы и требующая меньшего числа проходов по коллекции за счет более частого обновления параметров. Качество определения семантической близости оценивается на стандартных тестовых наборах пар слов на английском языке WordSim-353, SimLex-999, Bruni MEN. Приводятся экспериментальные результаты, которые подтверждают, что модель PWE превосходит по когерентности стандартную тематическую модель LDA и модель SGNS при сопоставимом качестве оценки близости слов при сравнении с моделью SGNS.

Для учета дополнительной мета-информации, представленной в виде временных меток документов и категорий, А.А. Потапенко обобщила разработанный алгоритм построения векторных представлений слов на случай мультимодальных данных с помощью подхода аддитивной регуляризации тематических моделей. Проведенные эксперименты показывают, что тематическое моделирование требует данные меньших объемов для получения качественных представлений, чем модели семейства word2vec. Это свидетельствует о практической применимости разработанных моделей и алгоритмов построения представлений слов.

В **заключении** описаны основные результаты диссертационной работы, даются рекомендации по использованию полученных результатов и рассматриваются направления дальнейших исследований. Сформулированные в диссертационной работе научные положения, выводы и рекомендации являются вполне обоснованными.

Научная новизна и практическая значимость

Научная новизна диссертационной работы А.А. Потапенко заключается в том, что предложена новая аддитивно регуляризованная тематическая модель, позволяющая достичь высокой разреженности, различности и интерпретируемости предметных тем, и алгоритмы построения единого векторного пространства для слов, сегментов текста и мета-данных документа, где компоненты векторов также приобретают интерпретируемость. Для предложенных моделей приведено теоретическое и экспериментальное обоснование. Особенно хочется отметить подробное теоретическое обоснование, излагаемое на высоком профессиональном уровне и включающее ряд математических доказательств теорем. Автором предлагается формализация дистрибутивной гипотезы в рамках подхода аддитивной регуляризации тематических моделей. Разработанные модели улучшают

предметной области текстовых коллекций и могут применяться для поиска по коллекциям научных статей, категоризации и суммаризации документов.

Практическая значимость диссертации А.А. Потапенко заключается в том, что на основе предложенных моделей, разработан программный комплекс с использованием библиотек NLTK, Gensim, BigARTM с учетом принципов воспроизводимости результатов.

Замечания

К недостаткам, носящий оформительский характер, можно отнести следующие:

1. Основные положения, выносимые на защиту, и экспериментальные результаты являются личным вкладом автора, однако в тексте диссертации остаются ссылки на научный коллектив: “мы подробно рассмотрим возможные математические модели (стр. 19), “в наших экспериментах” (стр. 61), “эта гипотеза была подтверждена в нашем эксперименте” (стр. 63), “мы использовали в наших экспериментах” (стр. 90), “наше дальнейшее исследование” (стр. 107), “по нашим наблюдениям” (стр. 108) и др.

2. В тексте перечисляются популярные библиотеки для обработки текстов, однако отсутствуют ссылки на статьи о них. В частности, авторы библиотеки Gensim на официальном сайте рекомендуют использовать ссылку на статью [Rehurek R., Sojka P. Software framework for topic modelling with large corpora //In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010.]

3. На 5 странице диссертации упоминается метрика PMI, однако сокращение расшифровывается только через несколько страниц (на 18 странице).

4. На 57 странице диссертации сказано, что “на графиках данной главы приводится зависимость перплексии от номера итерации обучения”. Хотелось бы увидеть точные ссылки на конкретные рисунки.

По существу можно отметить следующие замечания:

1. В третьей главе не объясняется выбор количества тем для проведения экспериментов на коллекции авторефератов диссертаций RuDis на русском языке и коллекции текстов научных статей NIPS на английском языке. Указано, что “выбор числа тем дополнительно исследуется в четвертой главе и в работе [25]”. Однако в четвертой главе и в статье [25] проводятся эксперименты только на коллекции NIPS.

2. В пятой главе для экспериментов автор использует версию англоязычной Википедии; оценка качества моделей представлений слов проводится на стандартных наборе с экспертными оценки близости WordSim353. Для русского языка существуют схожие ресурсы: версия Википедии и русскоязычная версия WordSim353, описанная в статье 2015го года [Panchenko et al. RUSSE: The first workshop on Russian semantic similarity (2015) *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2 (14), pp. 89-105.]. Хотелось бы увидеть дополнительные эксперименты, чтобы убедиться в том, что предложенная модель тематических представлений слов не зависит от особенностей языка.

3. В третьей главе качество вероятностных моделей оцениваются по перплексции, а в четвертой главе – по когерентности, перплексии и ряду дополнительных метрик, предложенных автором. Чем объясняется достаточность оценки моделей по одному критерию в третьей главе?

Указанные недостатки не являются принципиальными и не умаляют достоинств диссертации.

Заключение

Всего по теме диссертации опубликовано 12 печатных работ: 7 из них в изданиях, включенных ВАК РФ в перечень изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание доктора и кандидата наук. 6 печатных работ опубликованы в изданиях, индексируемых в Scopus. Опубликованные работы полностью отражают результаты диссертации.

Автореферат соответствует диссертации, отражает её содержание и дает представление об актуальности темы, целях, задачах, методах исследования, научной новизне, практической ценности, реализации, апробации, объеме, кратком содержании и результатах работы.

Исходя из вышеизложенного, можно утверждать, что диссертация Потапенко А.А. на соискание ученой степени кандидата физико-математических наук является законченной научно-квалификационной работой, в которой предложено решение задачи построения семантических векторных представлений текста на основе вероятностного тематического моделирования, что имеет существенное значение в области технологий интеллектуальной обработки текстов.

Считаю, что диссертационная работа Потапенко Анны Александровны

моделирования” удовлетворяет формуле специальности и всем требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а её автор заслуживает присуждения ей учёной степени кандидата физико-математических наук по специальности 05.13.17 – Теоретические основы информатики.

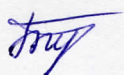
Кандидат физико-математических наук,
старший научный сотрудник
НИИ “Хемоинформатика и молекулярное моделирование”
Химического института им. А.М. Бутлерова
ФГАОУ ВО “Казанский (Приволжский)
федеральный университет”

Контактные данные:

Email: elvtutubalina@kpfu.ru

Тел.: +79600318025

Адрес: Казань, ул. Лобачевского, д. 1/29


29.01.2019

Тутубалина Елена Викторовна

