

УТВЕРЖДАЮ

Проректор Национального
исследовательского университета
«Высшая школа экономики»

к.э.н., доцент Рощин Сергей Юрьевич

«июль» 2018 г.

ЗАКЛЮЧЕНИЕ

**федерального государственного автономного
образовательного учреждения высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»**

Диссертация Потапенко Анны Александровны на тему: «Семантические векторные представления текста на основе вероятностного тематического моделирования» выполнена на базовой кафедре Яндекс департамента больших данных и информационного поиска факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики».

В период подготовки диссертации соискатель ученой степени Потапенко Анна Александровна работала в компании Общество с ограниченной ответственностью «Яндекс» в группе предиктивной аналитики в период с 01.08.2016 по 28.02.2017 в должности младшего аналитика, в период с 03.2017 по настоящее время не работала.

В 2014 г. окончила федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Московский государственный университет имени М.В. Ломоносова» по специальности «Прикладная математика и информатика».

В 2018 г. окончила очную аспирантуру федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» по направлению 09.06.01 «Информатика и вычислительная техника», направленность (образовательная программа) 05.13.17 «Теоретические основы информатики».

Научный руководитель – д. физ.-мат. н., профессор РАН Воронцов Константин Вячеславович – работает в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» в должности профессора базовой кафедры Яндекс департамента больших данных и информационного поиска факультета компьютерных наук.

По итогам обсуждения принято следующее заключение.

Актуальность темы

Векторные представления текста используются как признаковое описание данных в широком круге задач анализа текстов, в частности, в задачах классификации, машинном переводе, диалоговых системах, информационном поиске, рекомендательных системах. Несмотря на активное развитие моделей векторной семантики, большинство из них остаются не интерпретируемыми: компоненты построенных векторов не отвечают каким-либо аспектам смысла. Это не соответствует гипотезам из когнитивных наук о представлении понятий в мозге человека, а также затрудняет использование моделей в системах, требующих определения конкретных аспектов, обеспечивающих семантическую близость текстов. В данной работе развивается аппарат вероятностного тематического моделирования, позволяющий строить интерпретируемые векторные представления текста. Исследуются проблемы неединственности решения, различности и разреженности тем, определения числа тем в модели, устойчивости к шуму в данных.

Обоснованность научных положений

В работе соискателя приводится подробный обзор существующих методов в области семантического анализа текстов. Предложенные методы направлены на решение открытых проблем и согласуются с известными результатами. В теоретической части работы используются понятия из машинного обучения, оптимизации, линейной алгебры. Постановка экспериментов отвечает стандартам воспроизводимости, качество оценивается на наборах данных, известных в области семантического анализа текстов.

Личное участие соискателя ученой степени в получении результатов, изложенных в диссертации, заключается в развитии подхода аддитивной регуляризации тематических моделей, в частности, в формализации гипотезы дистрибутивной семантики в рамках данного подхода и включении в модель данных дополнительных модальностей. Соискателем были самостоятельно реализованы все предлагаемые методы и проведены эксперименты на реальных текстовых коллекциях, подтверждающие практическую значимость полученных результатов.

Степень достоверности результатов проведенных соискателем ученой степени исследований обусловлена математическим доказательством теорем, проведением воспроизводимых экспериментов на открытых данных, а также отзывами специалистов в области анализа текстов, проводивших рецензирование научных работ соискателя на российских и международных конференциях.

Научная новизна работы

В диссертации развивается подход аддитивной регуляризации тематических моделей, который является альтернативой байесовскому подходу, доминирующему в литературе по вероятностному тематическому моделированию. В рамках этого подхода становится возможным комбинирование требований к тематической модели в единой многокритериальной задаче.

Соискателем предлагается формализация понятия интерпретируемости предметных тем и разрабатывается методика ее оценивания. Предлагается новая тематическая модель, в которой предметные темы очищаются от слов общей лексики языка и фоновых терминов коллекции, повышается их разреженность и различность.

Демонстрируется применимость тематических векторных представлений к задачам определения семантической близости текстов. Предлагается новая тематическая модель со-встречаемости слов, которая на данных задачах существенно превосходит по качеству стандартные тематические модели коллекций текстовых документов.

Теоретическая значимость

Изложены основные подходы построения векторных представления текста в единой терминологии. Предложен обобщенный ЕМ-алгоритм, который приводит как к известным тематическим моделям, так и новым, в которых сокращается эффект переобучения и повышается разреженность настраиваемых параметров. В рамках подхода аддитивной регуляризации тематических моделей предложена комбинация из пяти регуляризаторов, позволяющая повышать интерпретируемость и различность предметных тем при сохранении основного критерия правдоподобия. В рамках того же подхода формализована гипотеза дистрибутивной семантики. Доказана теорема, на основе которой построен регуляризованный ЕМ-алгоритм для обучения тематической модели со-встречаемости слов в локальных контекстах. Данный алгоритм обобщен на случай сегментированного текста и мультимодальных данных. Показана доказана связь между двумя существующими тематическими моделями со-встречаемости слов, заключающаяся в наложении дополнительных условий на параметры моделей при инициализации.

Практическая значимость результатов проведенных соискателем ученой степени исследований

Предложенные алгоритмы вероятностных тематических представлений слов и документов могут применяться в системах информационного и разведочного поиска, категоризации документов, автоматического построения онтологий предметных областей и других задачах анализа текстов.

В проведенных экспериментах продемонстрировано высокое качество предложенных алгоритмов на задаче предсказания близости научных статей arXiv, при этом показано, что векторные представления документов обладают разреженностью и интерпретируемостью компонент. Совмещение данных

требований является ключевым для построения поисковых и рекомендательных систем научной литературы.

Апробация работы

1. 16-ая Всероссийская конференция «Математические методы распознавания образов», Казань, 6-12 октября 2013 г. – Доклад по теме: «Разреживание вероятностных тематических моделей».
2. The 35-th European Conference on Information Retrieval, Москва, 24-27 марта. – Постер по теме: “Robust PLSA Performs Better Than LDA”.
3. Международная конференция по компьютерной лингвистике «Диалог», Москва, 4-8 июня 2014 г. – Доклад по теме: «Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем».
4. XXI Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов», Москва, 7-11 апреля 2014 г. – Доклад по теме: «Регуляризация вероятностной тематической модели для выделения ядер тем».
5. The Third International Symposium on Learning and Data Sciences, Лондон, 20-23 апреля 2015. – Доклад по теме: “Additive Regularization of Topic Models for Topic Selection and Sparse Factorization”.
6. Artificial Intelligence and Natural Language, Санкт-Петербург, 20-23 сентября 2017. – Доклад по теме: “Interpretable Probabilistic Embeddings: Bridging the Gap Between Topic Models and Neural Networks”.
7. 7th International Conference - Analysis of Images, Social networks and Texts, Москва, 5-7 июля 2018. – Доклад по теме: “Probabilistic approach for embedding arbitrary features of text”.
8. Доклады на семинарах в компаниях: Google (июнь 2016 г., Цюрих, Швейцария), Microsoft Research (апрель 2015, Кэмбридж, Великобритания).
9. Доклады в научных группах: лаборатория профессора Томаса Хофманна, ETH Zurich (ноябрь 2017 г.), лаборатория профессора Криса Бимана, TU Darmstadt (июль 2016 г.).
10. Постеры на воркшопах международных конференций: ACL (август 2017 г., Ванкувер, Канада), EMNLP (октябрь 2018 г., Брюссель, Бельгия).

Полнота изложения материалов диссертации в публикациях

Основные результаты диссертационного исследования опубликованы в 12 работах общим объемом 8,5 п.л.; личный вклад автора составляет 4,1 п.л.

Научная публикация	Личны й вклад	Публикация входит в		
		Междуна родные	Список рекоме	Перечень рецензиру

		базы данных и системы цитирова ния (Web of Science/ Scopus/ MathSci Net...)	ндованн ых журнал ов НИУ ВШЭ	емых научных журналов, рекоменд ованных БАК
1. Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей/ Воронцов К. В., Потапенко А. А. // Компьютерные исследования и моделирование. 2012 Т. 4. №4. С 693–706. – 0,81 п.л.	в соавт., личный вклад – 0,3 п.л.	нет	да	да
2. Potapenko A. A. Additive Regularization of Topic Models/ Vorontsov K.V., Potapenko A. A. // Machine Learning, Volume 101, Issue 1, Page 303-323, Springer, 2015. DOI: 10.1007/s10994-014-5476-6. – 1,31 п.л.	в соавт., личный вклад – 0,6 п.л.	Web of Science – Q2, Scopus – Q1	да	да
3. Potapenko A. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. / Potapenko A. Popov A., Vorontsov K. // Artificial Intelligence and Natural Language. AINL-2017. Communications in Computer and Information Science, Springer, 2017. Vol. 789, pp. 167-180. 2017. – 0,8 п.л.	в соавт., личный вклад – 0,7 п.л.	Scopus – Q3	да	да
4. Potapenko A.A. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization./ Vorontsov K.V., Potapenko A.A., Plavin A.V. // The Third International Symposium On Learning And Data Sciences. April 20-22, 2015. Royal Holloway, University of London, UK. Springer International Publishing Switzerland 2015, A. Gammerman et al.	в соавт., личный вклад – 0,4 п.л.	Web of Science – Q4, Scopus – Q2	да	да

(Eds.): SLDS 2015, Lecture Notes in Computer Science, vol. 9047, pp. 193–202, 2015. – 0,6 п.л.				
5. Potapenko A. A. Robust PLSA Performs Better Than LDA/ Potapenko A. A., Vorontsov K. V. // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. Lecture Notes in Computer Science, Springer-Verlag Germany, 2013. Vol. 7814, pp. 784–787. – 0,25 п.л.	в соавт., личный вклад – 0,15 п.л.	Web of Science – Q4, Scopus – Q2	да	да
6. Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization./ Vorontsov K. V., Potapenko A. A. // Analysis of Images, Social Networks and Texts. AIST-2014. Communications in Computer and Information Science, 2014. Vol 436, pp 29-46. – 1,1 п.л.	в соавт., личный вклад – 0.2 п.л.	Scopus – Q4	да	да
7. Потапенко А. А. Многокритериальная регуляризация вероятностных тематических моделей для улучшения интерпретируемости тем и определения числа тем./ Воронцов К. В., Потапенко А. А. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог».— М.: Изд-во РГГУ, 2014, С. 707-719. – 0,8 п.л.	в соавт., личный вклад – 0.5 п.л.	Scopus	да	да
8. Potapenko A. Learning and Evaluating Sparse Interpretable Sentence Embeddings./ Trifonov V., Ganea O., Potapenko A., Hofmann T. // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.— Association for Computational Linguistics, 2018, pp 200-210. – 0,6 п.л.	в соавт., личный вклад – 0,2 п.л.	нет	нет	нет
9. Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования./ Воронцов К. В., Потапенко А. А. // Машинное обучение и анализ данных, 2013, Т. 1, № 6, С. 657-686. – 1,8 п.л.	в соавт., личный вклад – 0,8 п.л.	нет	нет	нет

10. Потапенко А. А. Разреживание вероятностных тематических моделей. // Математические методы распознавания образов: 16-ая Всеросс. конф. (ММРО-2013) — М.: МАКС Пресс, 2013, С. 89. — 0,05 п.л.	0,05 п.л.	нет	нет	нет
11. Потапенко А. А. Регуляризация вероятностной тематической модели для выделения ядер тем. // XXI Международная научная конференция студентов, аспирантов и молодых ученых "Ломоносов-2014".— М.: Издательский отдел факультета ВМК МГУ, 2014. С. 80-82. — 0,12 п.л.	0,1 п.л.	нет	нет	нет
12. Потапенко А. А. Робастные разреженные вероятностные тематические модели./ Воронцов К. В., Потапенко А. А. // Международная научная конференция Интеллектуализация обработки информации (ИОИ-2012), Черногория, г. Будва: Докл. — Москва: Торус Пресс, 2012. С. 605–608. — 0,25 п.л.	0,1 п.л.	нет	нет	нет

Публикации полностью соответствуют теме диссертационного исследования и раскрывают её основные положения.

Ценность научных работ соискателя ученой степени

Ценность научных работ соискателя заключается в развитии подхода вероятностного тематического моделирования и установлении связи тематических моделей с векторными моделями семантики. С теоретической точки зрения, предложены новые алгоритмы построения интерпретируемых разреженных тематических векторных представлений текста. С практической точки зрения, продемонстрировано улучшение ряда критериев качества при решении стандартных задач, в частности, определения семантической близости текстов. Результаты научных работ докладывались на ведущих конференциях в области анализа текстов и были положительно восприняты научным сообществом.

Диссертация Потапенко Анны Александровны на тему: «Семантические векторные представления текста на основе вероятностного тематического

моделирования» – это законченная научно-квалификационная работа, которая соответствует требованиям пунктов 9, 10, 14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842, а также Паспорту научной специальности 05.13.17 «Теоретические основы информатики» (физико-математические науки), в частности по следующим пунктам:

- пункт 5: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.
- пункт 6: Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке.

Диссертация Потапенко Анны Александровны на тему: «Семантические векторные представления текста на основе вероятностного тематического моделирования» рекомендуется к защите на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 «Теоретические основы информатики».

Заключение принято на заседании департамента больших данных и информационного поиска факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» 9 ноября 2018 г., протокол № 2.3.7-01/0911-1.

Присутствовало на заседании 10 человек.

Результаты голосования: «за» – 10 человек, «против» – 0 человек, «воздержалось» – 0 человек.

Руководитель департамента больших данных и информационного поиска
факультета компьютерных наук НИУ ВШЭ,
к. физ.-мат. н., доцент



В.В. Подольский

Подпись заверяю

19.11.2018
Специалист по кадровому делопроизводству
ОТДЕЛА КАДРОВОГО АДМИНИСТРИВАЦИОННОГО
УПРАВЛЕНИЯ ПЕРСОНАЛА
Т.В. Щегольская

