Министерство науки и высшего образования
Российской Федерации

УНИВЕРСИТЕТ ИТМО

Некоммерческое партнерство ПРИОР Северо-Запад

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ВЫЧИСЛИТЕЛЬНЫЕ ОНТОЛОГИИ

Выпуск 8

Труды XXVII Международной объединённой научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г.

Сборник научных трудов

Санкт-Петербург

2024

УДК 81’33
ББК 81.1
К63

К63 **Компьютерная лингвистика и вычислительные онтологии.** Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г.) Сборник научных трудов. — СПб.: Университет ИТМО, 2024. — 83 с.

ISSN 2541-9781
ISBN 978-5-7577-0728-0

В сборник включены тексты научных статей, представленные на XXVII Международной объединенной научной конференции «Интернет и современное общество» (Internet and Modern Society – IMS). Работы прошли рецензирование и отобраны в результате конкурсной процедуры. Сборник снабжен авторским указателем.

Издание адресовано научным работникам, преподавателям, аспирантам и магистрантам, изучающих междисциплинарные проблемы влияния информационно-коммуникационных технологий на трансформацию социальных и политических отношений в современном обществе.

Информация о конференции «Интернет и современное общество» представлена на сайте объединенной конференции (http://ims.itmo.ru).

Все статьи и тезисы докладов конференции IMS публикуются в открытом доступе (лицензия Creative Commons — CC-BY 3.0 Unported). Сборники научных статей, издаваемые в рамках конференции IMS с 2011 года, размещаются в Научной электронной библиотеке (http://elibrary.ru/) и Российском индексе научного цитирования (РИНЦ).

Подготовка конференции осуществлялась при поддержке Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, Комитета информатизации и связи и Комитета по науке и высшей школе Санкт-Петербурга.

УДК 81’33
ББК 81.1

ИТМО (Санкт-Петербург) — национальный исследовательский университет, научно-образовательная корпорация. Альма-матер победителей международных соревнований по программированию, один из ведущих вузов России по подготовке кадров для цифровой экономики. Приоритетные направления: IT и искусственный интеллект, фотоника, робототехника, квантовые коммуникации, трансляционная медицина, Life Sciences, Art&Science, Science Communication.

Лидер федеральных программ «Приоритет-2030» и «Передовые инженерные школы». С 2022 года ИТМО работает в рамках новой модели развития — научно-образовательной корпорации. В её основе академическая свобода, поддержка начинаний студентов и сотрудников, распределенная система управления, приверженность открытому коду, бизнес-подходы к организации работы. Образование в университете основано на выборе индивидуальной траектории для каждого студента.

По версии SuperJob, ИТМО занимает первое место в Санкт-Петербурге и второе в России по уровню зарплат выпускников в сфере IT. Университет в топе международных рейтингов среди российских вузов. Входит в топ-5 российских университетов по качеству приема на бюджетные места. Рекордсмен по поступлению олимпиадников в Санкт-Петербурге. С 2019 года ИТМО самостоятельно присуждает ученые степени кандидата и доктора наук.

ISBN 978-5-7577-0728-0

XXVII Международная объединённая научная конференция **«Интернет и современное общество»** (IMS-2024)

Санкт-Петербург, 24–26 июня 2024 г.
http://ims.itmo.ru

Конференция «Интернет и современное общество» (Internet and Modern Society – IMS) проводится в Санкт-Петербурге ежегодно с 1998 г. С 2014 г. конференция проводится в международном формате.

Объединённая конференция «Интернет и современное общество» в 2024 г. была проведена при поддержке Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, Комитета по науке и высшей школе и Комитета по информатизации и связи Санкт-Петербурга. Отдельные специализированные мероприятия проводились в сотрудничестве с проектами, реализуемыми при поддержке Российского научного фонда и Санкт-Петербургского научного фонда.

Конференция названа объединённой, так как научная программа конференции консолидирует серию специализированных международных и российских научных конференций, симпозиумов, семинаров, круглых столов и других мероприятий, посвящённых специальным вопросам развития технологий информационного общества. Отдельные специализированные и проблемно-ориентированные мероприятия проводятся в сотрудничестве с партнёрскими организациями.

Основу научной программы конференции 2024 г. составили международные компоненты, включающие сессии на русском и английском языках:

- **VII Международная конференция по электронному управлению** (Digital Transformation in Governance and Society — DTGS-2024);
- **международный семинар «Компьютерная лингвистика»** (Computational Linguistics — CompLing-2024);
- **международный семинар «Искусство и инновации в музеях»** (International Art and Innovation in Museums Seminar — AIMs-2024).

Традиционно в программу конференции были включены сессии научных докладов:

- **Электронное обучение и дистанционные образовательные технологии;**
- **Культурология киберпространства;**
- **Киберпсихология;**
- **Этико-правовые аспекты цифровой трансформации.**

Программу объединённой конференции расширили специализированные мероприятия, ориентированные не только на исследователей, но и на экспертное сообщество, и молодых ученых:

- международный симпозиум **«Interactive Systems & Information Society Technologies»** (InterSys-2024), организованный пятью университетами: Университетом ИТМО (Санкт-Петербург, Россия), Новосибирским государственным техническим университетом (Новосибирск, Россия), Институтом технологий и науки Бирла (Birla Institute of Technology & Science; кампус в Дубае, ОАЭ), Цзинаньским институтом суперкомпьютерных технологий (Jinan Institute of Supercomputing Technology; Шаньдун, Китай) и Федеральным университетом Параны (Federal University of Paraná; Куритиба, Бразилия);

- международный научно-практический симпозиум **«Цифровизация как инструмент отложенного старения / Digital Health and Active Aging Development»,** организованный в сотрудничестве с Хуачжунским университетом науки и технологии, Ухань, Китай (Huazhong University of Science and Technology, Wuhan, China) и при поддержке проекта РНФ № 22-18-00461 «Отложенное старение или поздняя взрослость в России: как цифровое развитие меняет статус пожилых в эпоху COVID-19 и неопределенности»;
- межрегиональный научно-практический семинар **«Электронное участие в регионах России 2020–2024 гг.»** (при поддержке проекта РНФ № 22-18-00364 «Институциональная трансформация управления электронным участием в России: исследование региональной специфики» и в сотрудничестве с Министерством цифрового развития, связи и массовых коммуникаций Российской Федерации и АНО «Диалог Регионы»);
- семинар и круглый стол **«Цифровые экосистемы в государственном и муниципальном управлении»** (при поддержке проекта РНФ и СПбНФ № 23-18-20079 «Исследование социальной результативности электронного взаимодействия граждан и власти в Санкт-Петербурге на примере городских цифровых сервисов», в сотрудничестве с СПб ИАЦ и Комитетом цифрового развития Ленинградской области);
- специализированный научно-практический семинар **«Цифровое здравоохранение: развитие пациентоориентированности»** (при поддержке компании «Нетрика Медицина»);
- Young Scholars' Poster Session **«Digital Transformation in Governance and Society»** (Young DTGS-2024).

На конференцию IMS-2024 было подано 228 заявок авторами из России, Объединённых Арабских Эмиратов, Индии, Китая, Италии, Испании, Эфиопии, Нигерии, Сербии, Египта и других стран. В научную программу конференции вошло 137 докладов.

Отбор докладов на конференцию и текстов для публикации производится по результатам двойного слепого рецензирования членами программного комитета с использованием международной системы сопровождения научных конференций EasyChair.org. В 2024 г. в рецензировании научных текстов приняли участие более 90 членов программного комитета и приглашённых рецензентов со всего мира, сформировавших около 400 рецензий.

Общее количество зарегистрированных участников (докладчиков, слушателей, исследователей и экспертов-практиков), посетивших сессии научных докладов, научно-практические семинары и круглые столы конференции, составило более 400 человек.

Благодаря информационной и организационной поддержке, которую оказали органы власти Санкт-Петербурга и Ленинградской области, в 2024 г. в научно-практических мероприятиях и круглых столах конференции IMS-2024 приняли участие более 70 сотрудников исполнительных органов государственной власти, органов местного самоуправления и подведомственных учреждений.

В 2024 г. международный симпозиум «Interactive Systems & Information Society Technologies» прошёл в формате двух сессий. Первая сессия предваряла основные треки конференции IMS и состоялась 16–17 мая в Дубае в Институте технологий и науки Бирла (Birla Institute of Technology & Science). Научная программа первой сессии симпозиума включила в себя 13 докладов, подготовленных авторскими коллективами из России, Китая, Объединённых Арабских Эмиратов и Индии.

По результатам объединенной конференции IMS-2024 традиционно издаются три сборника научных трудов (сериальные издания) и сборник тезисов на русском языке:
- **Государство и граждане в электронной среде** (ISSN 2541-979X), вып. 8;
- **Информационное общество: образование, наука, культура и технологии будущего** (ISSN 2587-8557), вып. 8;
- **Компьютерная лингвистика и вычислительные онтологии** (ISSN 2541-9781), вып. 8;

- **Интернет и современное общество**: сборник тезисов докладов IMS-2024.

Статьи, представленные для докладов на английском языке и прошедшие рецензирование, включены в сборники, подготовленные совместно с зарубежными партнерами конференции. Сборники публикуется в издательстве Springer (индексация в базе Scopus). Также в сборники включены научные статьи, отобранные на конкурсной основе за авторством молодых учёных — участников Young DTGS-2024.

Оргкомитет конференции сотрудничает с профильными научными журналами и использует возможность рекомендации лучших докладов, заслушанных и обсужденных на конференции, для публикации в журналах в доработанном виде с представлением более подробной информации о проведенных исследованиях:

- С 2017 г. конференция сотрудничает с научным журналом **«International Journal of Open Information Technologies»** (http://injoit.org, ВАК, РИНЦ), издаваемым в МГУ, по формированию специального номера. В 2024 г. такой номер планируется к изданию.
- Международный научный электронный журнал **«Культура и технологии»** (http://cat.ifmo.ru/) регулярно публикует лучшие статьи авторов IMS по своей тематике.
- С 2022 г. началось партнерство с научным журналом **«Journal on Interactive Systems»** (https://sol.sbc.org.br/journals/index.php/jis), Бразилия. В 2024 г. ряд докладов, представленных на английском языке, рекомендован для публикации в доработанном виде в этом журнале.

Электронные версии сборников конференции размещаются в свободном доступе (лицензия Creative Commons – CC-BY 3.0 Unported) на сайте материалов конференции «Интернет и современное общество» (http://ojs.itmo.ru). С 2017 г. всем статьям присваивается международный идентификатор DOI, а информация на уровне метаданных размещается в информационной системе CrossRef (https://search.crossref.org). Метаданные сборников размещаются в Научной электронной библиотеке (https://elibrary.ru), а все статьи и тезисы индексируются в Российском индексе научного цитирования (РИНЦ).

Информация обо всех сборниках и специальных номерах журналов, опубликованных с 2011 г., представлена на сайте конференции со ссылками на первоисточники — https://ims.itmo.ru/proceedings.html.

ПРОГРАММНЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Председатель Программного комитета:

Васильев В. Н., д-р техн. наук, чл.-корр. РАН, ректор Университета ИТМО

Заместители председателя Программного комитета:

Борисов Н. В., д-р физ.-мат. наук, заведующий кафедрой информационных систем в искусстве и гуманитарных науках СПбГУ, председатель оргкомитета конференции

Чугунов А. В., канд. полит. наук, директор Центра технологий электронного правительства ИДУ Университета ИТМО, генеральный директор НП ПРИОР Северо-Запад, ученый секретарь конференции

Члены Программного комитета:

Алексейцев С. А., канд. техн. наук, Новосибирский государственный технический университет

Бабина О. И., канд. филол. наук, Южно-Уральский государственный университет

Бакаев М. А., канд. техн. наук, Новосибирский государственный технический университет

Балаян А. А., канд. полит. наук, НИУ «Высшая школа экономики» — Санкт-Петербург

Беляева Л. Н., д-р филол. наук, Санкт-Петербургский государственный университет

Блинова О. В., канд. филол. наук, Санкт-Петербургский государственный университет

Богачева Н. В., канд. психол. наук, Первый Московский государственный медицинский университет им. И. М. Сеченова

Бодрунова С. С., д-р полит. наук, Санкт-Петербургский государственный университет

Болгов Р. В., канд. полит. наук, Санкт-Петербургский государственный университет

Борисов Н. В., д-р физ.-мат. наук, Санкт-Петербургский государственный университет

Бундин М. В., канд. юрид. наук, Нижегородский государственный университет им. Н. И. Лобачевского

Видясова Л. А., канд. социол. наук, Университет ИТМО

Галиева А. М., канд. филос. наук, Казанский федеральный университет

Галкин К. А., канд. социол. наук, Социологический институт РАН — филиал ФНИСЦ РАН

Глазкова А. В., канд. техн. наук, Тюменский государственный университет

Григорьева И. А., д-р социол. наук, Социологический институт РАН — филиал ФНИСЦ РАН

Демарева В. А., канд. психол. наук, Нижегородский государственный университет им. Н. И. Лобачевского

Иванов С. Е., канд. физ.-мат. наук, Университет ИТМО

Игнатьев А. В., д-р техн. наук, Волгоградский государственный технический университет

Игнатьева О. А., канд. социол. наук, Санкт-Петербургский государственный университет

Кабанов Ю. А., НИУ «Высшая школа экономики» — Санкт-Петербург

Камшилова О. Н., канд. филол. наук, РГПУ им. А. И. Герцена

Карачай В. А., канд. полит. наук, Университет ИТМО

Коган М. С., канд. техн. наук, Санкт-Петербургский политехнический университет Петра Великого

Кольцова О. Ю., канд. социол. наук, НИУ «Высшая школа экономики» — Санкт-Петербург

Конюховский П. В., д-р экон. наук, РГПУ им. А. И. Герцена

Королева Н. Н., д-р психол. наук, РГПУ им. А. И. Герцена

Кузьмич П. А., Университет ИТМО

Куприенко И. В., Университет ИТМО

Курчеева Г. И., канд. экон. наук, Новосибирский государственный технический университет

Лапошина А. Н., канд. пед. наук, Государственный институт русского языка им. А. С. Пушкина

Литвинова Т. А., д-р филол. наук, Воронежский государственный педагогический университет

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Председатель оргкомитета:

Борисов Н. В., д-р физ.-мат. наук, заведующий кафедрой информационных систем в искусстве и гуманитарных науках Санкт-Петербургского государственного университета

Заместитель председателя оргкомитета:

Прокудин Д. Е., д-р филос. наук, доцент Санкт-Петербургского государственного университета, аналитик Центра юзабилити и смешанной реальности Университета ИТМО

Члены оргкомитета:

Бакаев М. А., канд. техн. наук, Новосибирский государственный технический университет

Болгов Р. В., канд. полит. наук, Санкт-Петербургский государственный университет

Видясова Л. А., канд. социол. наук, Университет ИТМО

Григорьева И. А., д-р социол. наук, Социологический институт РАН — филиал ФНИСЦ РАН

Кабанов Ю. А., НИУ «Высшая школа экономики» — Санкт-Петербург

Метелева А. С., Университет ИТМО (информационный менеджер конференции)

Низомутдинов Б. А., Университет ИТМО, НП ПРИОР Северо-Запад

Орлов Г. М., канд. физ.-мат. наук, Северо-западный окружной научно-клинический центр им. Л. Г. Соколова ФМБА России

Толстикова И. И., канд. филос. наук, Университет ИТМО

Чижик А. В., канд. культурологии, Санкт-Петербургский государственный университет, Университет ИТМО

Чугунов А. В., канд. полит. наук, Университет ИТМО, НП ПРИОР Северо-Запад (ученый секретарь конференции)

Elakkiya R, PhD, Birla Institute of Technology and Science Pilani, UAE

Aleksandr RAIKOV, PhD, Jinan Institute of Supercomputing Technology, China

От редколлегии

Современная наука переживает стремительное развитие технологий, что требует новых подходов к исследованию текстовых данных. Компьютерная лингвистика и вычислительные онтологии играют в этой области ключевую роль, предоставляя мощные инструменты для автоматизированного анализа и обработки больших массивов информации.

Статьи, публикуемые в данном сборнике, являются изложением докладов русскоязычной секции семинара «Компьютерная лингвистика», состоявшегося 24–26 июня 2024 г. в рамках XXVII Международной объединенной конференции «Интернет и современное общество» — IMS-2024. Представленные материалы помогут углубить понимание современных методов обработки текстов и предложат практические решения для различных областей науки и бизнеса.

Сборник посвящается памяти нашего друга и коллеги Виктора Павловича Захарова, руководителя семинара «Компьютерная лингвистика», специалиста в области компьютерной, прикладной, корпусной лингвистики и информационного поиска. Виктор Павлович был не только ведущим специалистом в своей области, но и человеком, который на протяжении многих лет вдохновлял сотрудников и студентов преданностью науке. Именно благодаря его усилиям на конференции IMS-2007 впервые была выделена секция «Компьютерная и прикладная лингвистика», ставшая основой для создания в 2008 г. семинара «Лингвистические информационные технологии в Интернете». Этот семинар завершился публикацией первого отдельного сборника и положил начало развитию идеи объединения специалистов, занимающихся обработкой естественного языка, под эгидой ежегодного международного научного мероприятия. Так появился семинар «Компьютерная лингвистика», который каждый год расширяет свои горизонты, привлекая все больше участников со всего мира и приглашая специалистов не только из классических направлений (например, корпусная лингвистика), но и представителей науки о данных (data science), давая возможность высказаться и ученым, и представителям бизнеса. Успехи семинара — заслуга Виктора Павловича. Он навсегда останется в нашей памяти примером научной честности, креативности и безграничного энтузиазма. В связи с этим видится важной публикация итогов его последних проектов.

Итак, в сборнике представлено шесть статей.

В статье «Корпус текстов по корпусной лингвистике: состав и этапы формирования» обсуждается процесс создания корпуса статей по корпусной лингвистике, который был разработан на кафедре математической лингвистики Санкт-Петербургского государственного университета. Авторы описывают этапы формирования корпуса, включая унификацию форматов текстов, разметку именованных сущностей и генерацию аннотаций. Статья акцентирует внимание на значимости систематизации текстов и автоматизации их обработки.

Статья «Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем» посвящена созданию тематических моделей для корпуса текстов. Детально описываются методы, которые позволяют автоматически присваивать статьям метки тем, что значительно упрощает и ускоряет процесс анализа текстов. Работа отражает важность автоматизации анализа текстовых данных.

Статья «Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике» освещает результаты исследования методов метаразметки и визуализации данных в корпусе текстов. Визуализация связей между авторами, статьями и их аффилиациями предоставляет новые возможности для анализа и улучшает понимание структуры и динамики исследований в области корпусной лингвистики.

Статья «Частотные характеристики предлогов и их значений в базе данных предложных конструкций» является подведением итогов исследования, посвященного частотным характеристикам предлогов в русском языке. Авторы предлагают детальный анализ

использования предлогов на основе данных из предложных конструкций, что способствует более глубокому пониманию особенностей русского языка в контексте корпусной лингвистики.

Статья «Алгоритм сбора текстов для анализа тональности и тематического моделирования отзывов пациентов поликлиник» описывает разработку алгоритма для сбора текстов для анализа тональности отзывов. Особое внимание уделяется методике обработки текстов и их тематического моделирования, что может быть полезно в здравоохранении и смежных областях.

Сборник получился насыщенным и охватил широкий спектр актуальных тем в области компьютерной лингвистики и вычислительных онтологий. Он будет полезен как ученым, так и практикам, работающим над решением задач в различных сферах, где требуется автоматизация анализа текстовых данных и применение современных методов лингвистического исследования.

Редактор сборника
А. В. Чижик

Корпус текстов по корпусной лингвистике: состав и этапы формирования

О. А. Митрофанова, М. А. Адамова, Л. А. Букреева, А. К. Зернова,
А. А. Литвинова, В. С. Павликова, П. Ю. Сологуб

Санкт-Петербургский государственный университет

o.mitrofanova@spbu.ru, st110061@student.spbu.ru,
st110502@student.spbu.ru, st068103@student.spbu.ru,
st110228@student.spbu.ru, st109999@student.spbu.ru,
st095317@student.spbu.ru

Аннотация

Статья посвящена проблемам разработки корпуса статей по корпусной лингвистике, создаваемого на кафедре математической лингвистики СПбГУ. Корпус создан под руководством В. П. Захарова и включает в себя тексты докладов конференции «Корпусная лингвистика» с 2002 по 2021 гг., семинара «Компьютерная лингвистика и вычислительные онтологии» с 2011 по 2023 гг., а также некоторые другие материалы. В ходе работы над корпусным ресурсом была проведена унификация формата представления текстов, исследована структура статей. Осуществлены эксперименты по генерации ключевых слов и аннотаций в тех случаях, когда авторский текст не содержал данную информацию. Исследованы типы именованных сущностей, зафиксированных в корпусе, реализован алгоритм их разметки. Проведен анализ распределения докладов по тематическим блокам конференций в соответствии со схемой экспертной разметки.

Ключевые слова: корпусная лингвистика, материалы конференций, разметка, ключевые слова, аннотации, тематическая разметка, именованные сущности

Библиографическая ссылка: Митрофанова О. А., Адамова М. А., Букреева Л. А., Зернова А. К., Литвинова А. А., Павликова В. С., Сологуб П. С. Корпус текстов по корпусной лингвистике: состав и этапы формирования // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 13–29. DOI: 10.17586/2541-9781-2024-8-13-29.

1. Введение

Проект, представленный в данной статье, посвящен памяти основателя и руководителя Петербургской школы корпусной и компьютерной лингвистики Виктора Павловича Захарова, нашего учителя и коллеги, который с 2002 года был главным организатором конференций и семинаров, где обсуждались проблемы создания и применения корпусов текстов. За двадцатилетний период проведения научных встреч были собраны ценные материалы, которые связаны с историей корпусной и компьютерной лингвистики, с развитием основных направлений, с кругом проблем и предлагаемых решений, с исследованием этапов становления и изменений терминологии рассматриваемой предметной области, ее логико-понятийной схемы и принципов стандартизации. Цель проекта состояла в разработке комплексного корпусного и терминологического ресурса с возможностью многопараметрического поиска источников. Результаты предшествующих

исследований представлены в [1; 2; 3]. В данной статье рассматриваются следующие решенные нами задачи:

- формирование корпуса статей по корпусной лингвистике;
- типы информации, представленные в корпусе;
- разметка ключевых выражений в корпусе;
- генерация аннотаций статей;
- систематизация и разметка именованных сущностей.

Помимо этих задач были решены задачи систематизации рубрик в корпусе и мультимодальной тематической разметки, по формированию базы данных с метаинформацией и по разработке системы визуализации результатов поиска.

2. Состав и структура корпуса текстов ТКиКЛ

Процедура формирования корпуса ТКиКЛ на основе материалов конференций Corpora и IMS CompLing была комплексной, соответствовала протоколу, описанному в [4], и включала в себя несколько этапов.

Первый этап предполагал формирование электронной коллекции из текстов, опубликованных в материалах трудов конференций, которые включают статьи и тезисы (список изданий и их количественные параметры представлены в табл. 1). Тексты без аннотаций, авторских наборов ключевых слов и ссылок были преобразованы в файлы формата *.txt для дальнейшей обработки. Названия файлов были стандартизированы: в них обязательно входит название конференции (Corpora / IMS) и год публикации сборника. Статьи на английском языке не были включены в корпус и не извлекались из сборников.

На втором этапе разработки корпуса был разработан и применен код на языке программирования Python, который корректировал имена файлов для обеспечения единообразия, а также проводил лемматизацию всех файлов в папках разных годов с помощью библиотеки Pymorphy2, что дало дополнительное деление корпуса на тексты без лемматизации и с лемматизацией (Corpora_raw / Corpora_lemmatized, IMS_raw / IMS_lemmatized). Удаление нетекстовых элементов и лемматизация способствуют повышению качества анализа содержания текста и получить более точные результаты при использовании инструментов автоматической обработки текста.

Третий этап формирования корпуса состоял в сборке массива лемматизированных файлов (full_corpus) для дальнейшей их обработки, включающей следующие процедуры:

- автоматическое выделение ключевых слов и выражений;
- автоматическая генерация аннотаций;
- тематическое моделирование;
- автоматическая генерация меток тем.

Далее в нашей статье мы более подробно обсудим первые два этапа.

Четвертый этап включал сбор статистической информации о корпусе, автоматический подсчет количества токенов в текстах отдельных сборников с помощью счетчика, реализованного на языке Python.

Таким образом, процедура составления корпуса ТКиКЛ является трудоемким процессом, который включает в себя несколько этапов, начиная от сбора и лемматизации текстов до их систематизации. Благодаря использованию программных инструментов этот процесс был частично автоматизирован и упрощен.

Структура корпуса включает три каталога: Corpora, IMS, full_corpus. Первые два каталога подразделяются на еще два с необработанными и лемматизированными текстами: Corpora_raw, Corpora_lemmatized, IMS_raw, IMS_lemmatized. Каждый из этих каталогов включает папки годов с файлами статей соответствующих сборников. Названия папок маркированы тегами по следующему шаблону: для неразмеченных текстов — year, year_thesis (например, 2004, 2004_thesis); для лемматизированных — year_lem, year_thesis_lem (2004_lem, 2004_thesis_lem). Сами файлы унифицированы по шаблону: для

неразмеченных текстов — surname_conference name_year / thesis_year.txt (например, Gerd_CL_2006.txt, Gerd_CL_thesis_2004.txt, Masevich_IMS_2018.txt), для лемматизированных — surname_conference name_year / thesis year_lem.txt (Gerd_CL_2011_lem.txt, Alexeeva_IMS_2015_lem.txt).

Каталог с тегом IMS включают в себя 11 папок (с маркировкой от 2013 до 2023 г.), с тегом Corpora — 12 папок (2002, 2004, 2004_thesis, 2005, 2006, 2008, 2011, 2013, 2015, 2017, 2019, 2021). Сегмент корпуса, представляющий материалы конференции Corpora, в общей совокупности составили 442 файла, материалы семинара IMS по компьютерной лингвистике и вычислительным онтологиям — 201 файл. Общий размер корпуса — более 1 млн токенов.

Каталог full_corpus содержит 643 файла — все лемматизированные тексты двух конференций. Более подробное описание корпуса можно видеть в таблице 1 и на рисунках 1–5, где указаны сборники — источники материалов, а также их количественный состав.

Таблица 1. Количественный состав корпуса ТКиКЛ

№	Конференция – год	Число текстов	Количество токенов
1	Корпусная лингвистика и лингвистические базы – 2002	21	62634
2	Корпусная лингвистика – 2004 / тезисы	26 / 44	65412 / 15237
3	MegaLing – 2005	18	27277
4	Корпусная лингвистика – 2006	40	55524
5	Корпусная лингвистика – 2008	40	57388
6	Корпусная лингвистика – 2011	48	47749
7	Корпусная лингвистика – 2013	36	45484
8	Корпусная лингвистика – 2015	42	46963
9	Корпусная лингвистика – 2017	49	43517
10	Корпусная лингвистика – 2019	45	58828
11	Корпусная лингвистика – 2021	33	47835
		442	**541378**
12	IMS CompLing – 2013	48	111284
13	IMS CompLing – 2014	54	133133
14	IMS CompLing – 2015	12	30692
15	IMS CompLing – 2016	7	15323
16	IMS CompLing – 2017	19	43217
17	IMS CompLing – 2018	16	37396
18	IMS CompLing – 2019	14	41315
19	IMS CompLing – 2020	8	18552
20	IMS CompLing – 2021	7	16167
21	IMS CompLing – 2022	7	16452
22	IMS CompLing – 2023	9	22385
		201	**485916**

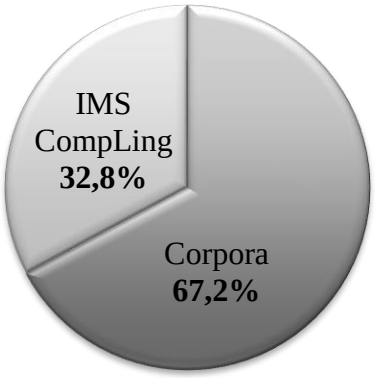

Рис. 1. Соотношение числа текстов в двух сегментах корпуса Corpora и IMS CompLing

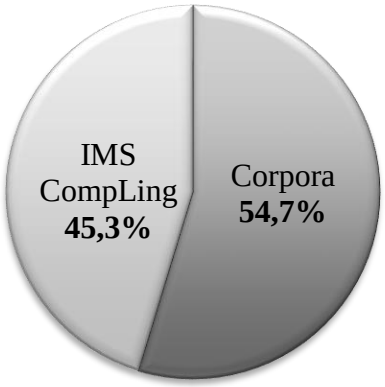

Рис. 2. Соотношение количества токенов в двух сегментах корпуса Corpora и IMS CompLing

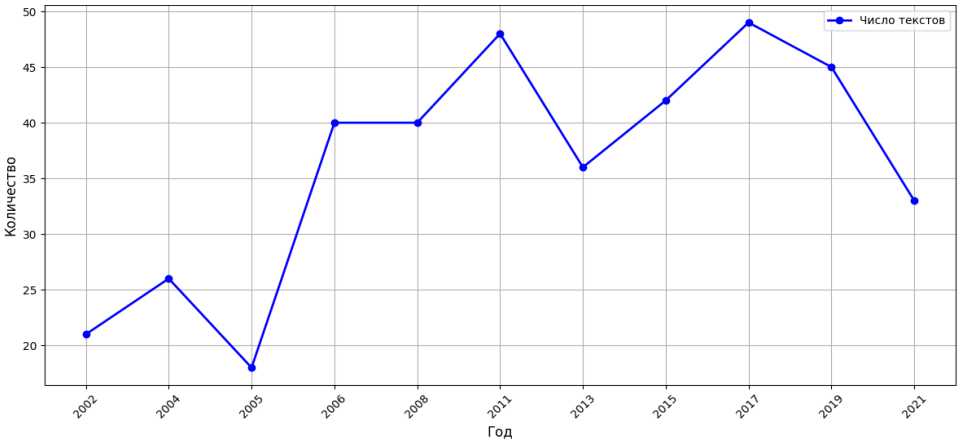

Рис. 3. Распределение текстов в корпусе по годам

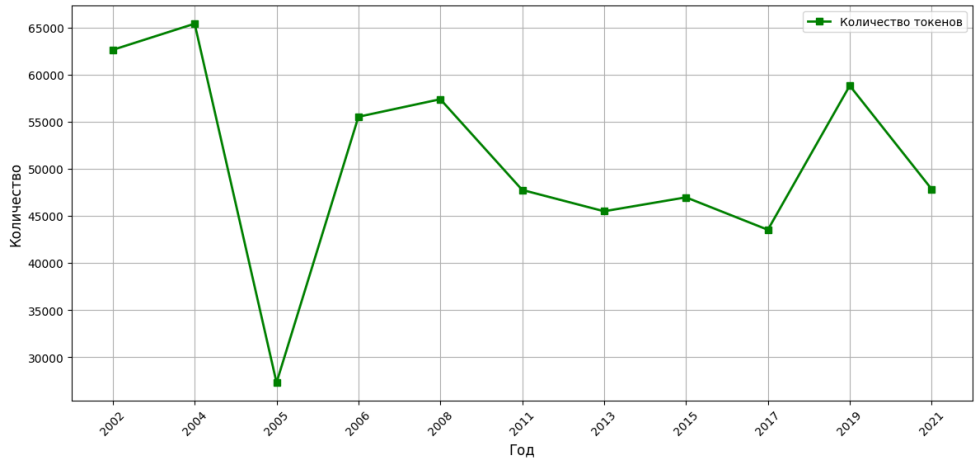

Рис. 4. Распределение токенов в корпусе по годам

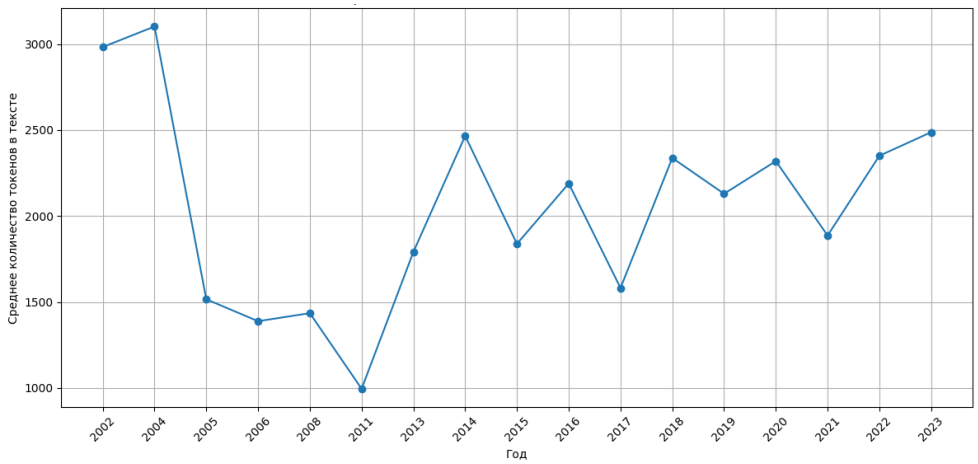

Рис. 5. Среднее количество токенов в тексте по годам

На основе данных из таблицы 1 и рисунков 1–5 можно сделать следующие выводы:

- из круговой диаграммы сравнения числа текстов (рис. 1) видно, что объем сегмента Corpora более чем в 2 раза выше, чем соответствующий сегмент IMS CompLing (статей IMS CompLing меньше, чем статей Corpora);
- по круговой диаграмме сравнения количества токенов (рис. 2) можно сделать вывод, что тексты сегмента IMS CompLing содержат больше токенов, чем тексты сегмента Corpora (статьи IMS CompLing длиннее статей Corpora);
- согласно распределению текстов, в корпусе по годам (рис. 3), наибольшее количество текстов было опубликовано в 2017 году (49 текстов), наименьшее количество текстов было опубликовано в 2005 году (18 текстов), общее количество текстов в корпусе имеет тенденцию к увеличению с течением времени;
- согласно распределению количества токенов в корпусе по годам (рис. 4), наибольшее количество токенов было внесено в корпус по текстам 2004 года (65412 токенов), наименьшее количество токенов было внесено в корпус по текстам 2015 года (46963 токенов); общее количество токенов в корпусе также увеличивается со временем (дополнительно можно отметить, что в 2013 году вклад материалов конференции IMS CompLing в корпус был наибольшим);

- наибольшее среднее количество токенов (средний объем текстов статей в токенах) было зарегистрировано в 2004 году (3101.88), а наименьшее в 2011 году (994.77) (рис. 5); в целом, наблюдается некоторая вариативность в значениях за исследуемый период, однако общая тенденция к изменению колеблется в пределах от 1388.10 до 3101.88.

В ходе составления корпуса были выявлены некоторые проблемы, связанные с его структурой и сбором материалов. Одной из основных проблем является необходимость восстановления отсутствующих компонентов текста: шаблоны оформления текстов статей менялись, отдельные статьи не имеют авторской аннотации и не содержат авторские наборы ключевых слов и словосочетаний. Для решения данной проблемы была проведена генерация ключевых выражений и аннотаций с применением моделей машинного обучения (см. разделы 3 и 4 данной статьи). Еще одной проблемой является отсутствие сопоставимого подкорпуса со статьями на английском языке, поскольку англоязычные тексты составляют значительную часть трудов конференций Corpora и IMS CompLing. Создание такого подкорпуса является задачей следующего этапа работы с корпусом, направленного на улучшение качества и репрезентативности корпуса для проведения дальнейших исследований.

3. Автоматическая генерация ключевых слов и словосочетаний в текстах корпуса ТКиКЛ

Автоматическое выделение ключевых слов и словосочетаний является необходимой процедурой в процессе подготовки научного текста, способствующей формированию информационно-поискового портрета текста. Наборы ключевых выражений помогают быстро оценить содержание текстов в ходе индексирования, рубрикации, суммаризации, упрощения и перефразирования [5; 6; 7; 8; 9; 10]. Методы выделения ключевых выражений разрабатывались прежде всего применительно к научным текстам с высокой концентрацией терминов и терминосочетаний, это объясняет ориентированность процедур автоматического выделения ключевых выражений на использование статистических признаков ключевых выражений, их структурной и лексико-грамматической организации (униграммы, биграммы, триграммы и т.д.), способы их ранжирования (регистрация их локализации в тексте, длина, встречаемость в составе других n-грамм), наличие одного корпуса текстов или пары корпусов — основного и фонового, возможность использования размеченных данных для организации процедур машинного обучения и т.д.

Автоматизация выделения ключевых выражений, равно как и ручная их разметка, является предметом дискуссий. Возникающие вопросы связаны с возможным несоответствием лексических единиц в реферативной и основной частях документа: зачастую назначаемые авторами ключевые выражения редко встречаются в тексте или вовсе в нем отсутствуют. В таких случаях неизбежно применение автоматических методов обработки данных. Базовыми количественными характеристиками, по которым можно оценить потенциальную значимость ключевых выражений для читателя, являются их плотность (отношение частоты употребления в тексте по отношению к его общему объему) и пространственно-позиционные признаки (расположение в документе). Принято считать, что наиболее информативны выражения, встречающиеся в заголовке, аннотации, в начальной части текста (первый абзац, первые несколько предложений), а также в конце текста (в заключении) [6].

В сборниках CL2002–2008, согласно использованному издательскому шаблону, отсутствовали авторские ключевые выражения, это обуславливает необходимость восстановления существующих лакун для унификации представления текстов в корпусе ТКиКЛ. Для автоматического извлечения ключевых выражений в нашем исследовании рассматривались разнородные алгоритмы, а именно, статистические: Log-Likelihood, TF-IDF, Хи-квадрат; гибридные (лингвостатистические): RAKE, YAKE, MultiRAKE, PullEnti,

RuTermExtract, графовые: TopicRank, с использованием машинного обучения: Spacy, KeyBERT [7; 8; 9; 10]. Рассматриваемый набор методов не является исчерпывающим. При отборе методов выделения ключевых выражений мы учитывали возможность их применения в работе с русскоязычными текстами, а также возможность извлечения n-грамм разной структуры (униграмм, биграмм, триграмм и т. д.). Основные методы выделения ключевых выражений учитывают не только их типичность для определенного документа или классов документов, но и их коллокационную природу, что важно для терминологически насыщенных текстов.

В [7] приведен анализ наборов ключевых выражений, выделенных в 50 текстах корпуса ТКиКЛ с учетом пространственно-позиционных и стилистически детерминированных характеристик ключевых выражений. В результате серии экспериментов были сопоставлены эталонные ключевые выражения, выделенные экспертами из первого сегмента текстов, и ключевые выражения, извлеченные из второго сегмента автоматическими методами. Наилучшие результаты показали алгоритмы PullEnti, RAKE и RuTermExtract. В [10] было проведено сравнение алгоритмов генерации ключевых выражений для аннотаций научных статей, в ходе которого было установлено, что самые высокие результаты по F-мере показывают алгоритмы YAKE и TopicRank. Полученные данные были применены в проекте НейроКРЯ по разметке ключевых выражений в корпусе региональной прессы, где был применен алгоритм RuTermExtract, хорошо зарекомендовавший себя в предыдущих экспериментах. Следуя нашим наблюдениям и опыту НейроКРЯ, мы приняли решение о генерации ключевых выражений для статей в сборниках *CL2002–2008* с помощью алгоритма RuTermExtract, отбирая первые три слова и три словосочетания с наибольшим весом. Пример разметки приведен в таблице 2.

Таблица 2. Примеры разметки ключевых выражений в текстах корпуса ТКиКЛ

№	Статья	Ключевые слова	Ключевые словосочетания
1	Андреев А. В. Архитектура информационно-поисковой системы для индоевропейского компьютерного тезауруса // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004	тезаурус, корпус, ИПС	индоевропейский компьютерный тезаурус, формат TEI, информационно-поисковая система
2	Апресян Ю. Д., Иомдин Л. Л., Санников А. В., Сизов В. Г. Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004	слово, дескриптор, корпус	семантическая информация, семантический словарь, семантическая роль
3	Беляева Л. Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004	текст, перевод, предложение	параллельный текст, машинный перевод, параллельный корпус
4	Захаров В. П. Толбаст С. П. Поисковая система сети Интернет и корпусные исследования // MegaLing 2005: Прикладная лингвистика в поисках новых путей. СПб., 2005	интернет, поиск, запрос	поисковая система, русский язык, корпусные исследования
5	Гарабик Р., Захаров В. П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006	текст, выравнивание, предложение	словацкий национальный корпус, пользовательский интерфейс, морфологический разметка

Продолжение таблицы 2

№	Статья	Ключевые слова	Ключевые словосочетания
6	Зубов А. В. Корпус текстов белорусского языка // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006	текст, корпус, кодирование	белорусский язык, корпус текстов, письменный текст
7	Герд А. С. Академическая лексикография как система корпусов // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006	словарь, слово, значение	словарный грамматика, академический словарь, теоретическая семантика
8	Копотев М. В. К построению частотной грамматики русского языка // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008	создание, корпус, Ханко	частотная грамматика, русский язык, частотные характеристики
9	Кустова Г. И. Электронный словарь степенной сочетаемости на базе Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008	словарь, значение, наречие	степенная сочетаемость, степенное слово, электронный словарь
10	Падучева Е. В. Прямая и косвенная диатеза ментального глагола: корпусное исследование // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008	глагол, мнение, знание	рематический акцент, прямая диатеза, пропозициональный актант

4. Автоматическая генерация аннотаций в текстах корпуса ТКиКЛ

Аннотация — это важный компонент структуры научной статьи, представляющий краткое и лаконичное изложение основного содержания исследования. Она представляет собой своего рода краткое резюме, которое помогает читателям быстро оценить, насколько статья соответствует их интересам и ожиданиям [11; 12; 13]. Структура аннотации, как правило, должна соответствовать требованиям IMRAD (Introduction, Methods, Results, and Discussion). Качественная аннотация научной статьи должна содержать следующие элементы: краткое изложение цели исследования, а также основных вопросов или задач, решаемых в работе; упоминание основных методов исследования, используемых для достижения поставленной цели; краткое описание основных результатов исследования или выводов, которые были получены в ходе работы. Помимо этого, аннотация может содержать уточнение, почему проведенное исследование важно, какие у него дальнейшие перспективы и какие практические или теоретические выводы можно сделать на его основе. Таким образом, аннотация должна быть ограниченного объема (в случае настоящего проекта до 250 слов) и информативной с точки зрения передачи содержания исходного текста, при этом структурированной и написанной доступным языком, чтобы читатели могли быстро понять суть исследования, не читая всю статью.

Автоматическая суммаризация текста широко применяется в различных областях, таких как информационные технологии, медицина, финансы, новости и другие, где требуется обработка большого объема информации для получения краткого обзора или анализа. Этот процесс может осуществляться с использованием различных методов и алгоритмов, предполагая суммаризация по предложениям, по документам, по корпусу текстов, по аспектам, одноязычную или многоязычную суммаризацию. Суммаризация представляет

собой вариант семантической компрессии, в ходе которой исходное содержание передается в тексте с сокращением плана выражения, при этом сходные механизмы используются при упрощении, когда результирующий текст должен быть формально проще и не обязательно короче [14], и перефразировании, когда исходный и итоговый тексты должны характеризоваться сходными содержанием и формой [15]. Два основных подхода к созданию аннотации текста: экстрактивная и абстрактивная суммаризация [16]. Основное отличие между этими двумя подходами заключается в том, как они обрабатывают исходный материал и формируют краткое содержание текста. Экстрактивный подход к суммаризации предполагает извлечение наиболее важных фрагментов оригинального текста (предложений или фраз) и комбинирование этих фрагментов для построения аннотации. В основном, при использовании такого подхода сохраняется структура и форма оригинального текста, так как экстрактивная суммаризация не предполагает генерации новых текстов или перефразирования уже имеющихся текстов. В отличие от экстрактивной суммаризации, абстрактивный метод суммаризации не ограничивается извлечением предложений из исходного текста, а предполагает порождение нового текста ограниченного объема с заданным в оригинале содержанием. Абстрактивная суммаризация позволяет не только переформулировать исходные предложения, но и генерировать новые, которых нет в оригинальном тексте. Такой подход сложнее с точки зрения технологической реализации, так как требует понимания текста и способности на его основании генерировать новое содержание.

В корпусе ТКиКЛ отсутствовали аннотации для статей CL 2002–2011, по этой причине было принято решение сгенерировать их автоматически при помощи алгоритмов суммаризации. В настоящем проекте к задаче автоматической генерации аннотаций текстов научных статей были применены два алгоритма суммаризации: в качестве алгоритма экстрактивной суммаризации был выбран алгоритм, представленный в библиотеке sumy [17], абстрактивная суммаризация осуществлялась при помощи модели T5 семейства Трансформер rut5_base_sum_gazeta [18]. Аннотации различаются степенью подробности и объемом: как видно из таблицы 3, аннотации sumy длиннее и конкретнее, тогда как аннотации rut5_base_sum_gazeta более краткие и обобщенные.

Таблица 3. Примеры аннотаций, сгенерированных для статей в текстах корпуса ТКиКЛ

№	Статья	Аннотация sumy	Аннотация rut5_base_sum_gazeta
1	Сергеева Е. М., Фивейская Е. А. Создание справочно-информационной базы данных для Лексического атласа русских народных говоров // Труды международной конференции «Корпусная лингвистика и лингвистические базы данных». СПб., 2002	Задача ЛАРНГ – показать в пространственной проекции (на карте) основные звенья словарного состава русских народных говоров, лексические и семантические различия в организации. Также в базе должна быть представлена следующая информация о картах, которые войдут в ЛАРНГ: данные об авторе-составителе; информация о материалах, полученных автором для составления карты; название карты (например, «ЛСЛ 123. Разработка такой базы данных входит в планы ИЛИ РАН на ближайшее время»	Создание справочно-информационной базы данных для Лексического атласа русских народных говоров (ЛАРНГ) является актуальной задачей русской диалектологии. Материалы будут иметь серьезное научное и культурно-историческое значение
2	Николаев И. С. Корпус текстов ижорских народных песен // Труды международной конференции «Корпусная лингвистика–2004». СПб., 2004	Работа с этим материалом показала, что для более полного сравнительно-исторического анализа ижорских диалектов необходимо изучить тексты, записанные у ижорцев ранее, а именно в конце XIX – начале XX вв. Несмотря на то, что в то время записей живой разговорной речи не велось, финскими учеными в Ингерманландии был собран обширный материал по устному народному творчеству, который	Корпус текстов ижорских народных песен в электронном виде был собран на кафедре математической лингвистики СПбГУ в рамках проекта «Полнотекстовая база данных по языкам и

Продолжение таблицы 3

№	Статья	Аннотация sumy	Аннотация rut5_base_sum_gazeta
2		был опубликован в многотомном собрании «Старые песни финского народа» [Suomen Kansan vanhat runot]. Решению всех этих и многих других проблем может способствовать создание корпуса текстов народных ижорских песен в электронном виде с соответствующей структурой и разметкой. Тем не менее, нам представляется, что, воспользовавшись опытом создания других корпусов текстов, можно решить перечисленные проблемы, а также те сложности, которые еще могут возникнуть при создании корпуса текстов ижорских народных песен	диалектам Северо-Запада России»
3	Котов А. А., Гопкало О. С. Русскоязычный эмоциональный корпус: коммуникативное взаимодействие в реальных эмоциональных ситуациях // Труды международной конференции «Корпусная лингвистика–2011». СПб., 2011	Эмоциональные корпуса важны для изучения общения с клиентами в состоянии стресса, для создания развлекательных технологий и для разработки эмоциональных компьютерных агентов: трехмерных компьютерных персонажей или роботов, способных взаимодействовать с человеком, распознавать его эмоции и правдоподобно имитировать собственные эмоции в процессе коммуникации. Для целей создания эмоциональных компьютерных агентов особый интерес представляют быстрые смены выражаемого эмоционального состояния и коммуникативных стратегий. Например, комбинация действий «поднимает брови» и «сжимает губы» (n = 47) может появляться в следующих контекстах: (a) Озадаченность: студент демонстрирует непонимание и озадаченность, смотрит в текст задания, поднимает брови, сжимает губы (20081219-zhum-a10, 20081227-fumo-a13), в некоторых случаях может при этом быстро моргать (20080717-c01, 20081227-fumo-a07)	В рамках проекта создания Русскоязычного эмоционального корпуса (REC) были фиксированы видеозаписи диалогов с клиентами в «службе одного окна» в одном из районов Москвы по вопросам оплаты коммунальных услуг
4	Ягунова Е. В. Исследование контекстной предсказуемости единиц текста с помощью корпусных ресурсов // Труды международной конференции «Корпусная лингвистика–2008». СПб, 2008. С. 396-403	В большей степени нас будут интересовать процедуры контекстной предсказуемости в рамках восприятия текста (речи), в меньшей степени мы обращаемся к данным порождения текста. Однако понятие «синтагматический сосед» требует уточнения; прежде всего, с точки зрения того, какая единица — словоформа или лемма — рассматривается в качестве коллоката. Таким образом, при решении разных задач контекстной предсказуемости оказывается важным сопоставлять данные по сочетаниям как словоформ, так и лексем	С помощью корпусных ресурсов можно рассматривать механизмы контекстной предсказуемости в рамках восприятия текста. Это может быть связано с вероятностями влияния разных позиций, которые способствуют (или не способствуют) адекватному восприятию соответствующих единиц текста

5. Автоматическая разметка именованных сущностей в текстах корпуса ТКиКЛ

Именованные сущности (Named Entities — NE) — это слова или словосочетания, которые выделяют предметы или явления в ряде аналогичных предметов или явлений. Именованные сущности в текстах представляют собой конкретные организации, объекты, даты, места, и другие имена, которые имеют определенное значение и могут быть идентифицированы как отдельные субъекты. Задача распознавания именованных

сущностей (Named Entity Recognition — NER) является важным этапом в извлечении информации (Information extraction, IE), которая состоит в автоматическом выделении структурированных данных из источников неструктурированной или слабоструктурированной информации и связана с информационным поиском и обработкой информации на естественных языках [19; 20; 21; 22].

Существуют различные методы выделения именованных сущностей в текстах, например: с применением создаваемых вручную наборов правил, с применением специализированных парсеров (например, библиотека Natasha [23], Yargy-parser [24]), основанных на статистических моделях с применением классического и глубинного машинного обучения (например, модели NER в проекте DeepPavlov [25]).

Особые именованные сущности в текстах могут быть связаны с уникальными или специфическими объектами, событиями или понятиями, которые имеют особое значение или статус. Они играют важную роль в анализе текстов, так как они содержат ценную информацию о контексте и содержании текста. Их распознавание и классификация может помочь выделить ключевые аспекты текста и информацию определенного вида (например, термины, названия организаций, персоналии и т.д.). Стоит также учитывать, что набор именованных сущностей и связи между ними на уровне вложенных сущностей (Nested Entities) [26] будет существенно различаться в зависимости от типа текста и его тематики.

Существующие программные комплексы, библиотеки, программы и программные интерфейсы приложений (API) для решения задачи извлечения именованных сущностей охватывают широкий тематический диапазон текстов и предлагают общий, стандартный набор именованных сущностей. Например, программа Stanford NER (CRFClassifier) выделяет следующие типы NE: Person; Location; Organization; Date; Time; Money; Percentage [27].

Однако для решения задачи информационного поиска в тематических текстах необходим более развернутый набор тегов для выделения именованных сущностей с более подробной классификацией. В ходе анализа ключевых слов из текстов, вошедших в корпус ТКиКЛ, были выделены следующие особые именованные сущности, характерные для рассматриваемой предметной области. В таблице 4 для каждого вида приведено название категории, возможный тег и примеры из корпуса. В ходе экспериментов с применением библиотек Natasha и yargy-парсера проведена разметка уникальных именованных сущностей в текстах корпуса ТКиКЛ.

Таблица 4. Уникальные именованные сущности в корпусе ТКиКЛ

№	Тип уникальной именованной сущности	Тег	Примеры
1	Названия конференций	CONF	Диалог
2	Язык	LAN	Китайский язык, русский язык
3	Названия моделей	MODEL	BERT, RuBERT
4	Проекты	PROJECT	IntelliText, CAT&kittens, Revita, Текстометр, Visualizing Russian, Русский конструктикон, RuSkell, CoCoCo
5	Стандарты	STANDARD	CTB CNS, PKU
6	Форматы разметки	FORMAT	CoNLL-U
7	Языки программирования	PR_LAN	Python
8	Библиотеки	LIB	UDPipe, Stanza
9	Алгоритмы	ALG	CWS (Chinese word segmentation), fastHan, LTP, PKUSeg, Ckiptagger
10	Корпусы	CORP	Русско-китайский параллельный корпус НКРЯ (ruzhcorp), НКРЯ, подкорпус RU-AC, CyberCAT, CAT, ruTenTen11
11	Тесты	TEST	Flesch Reading Ease, Flesch-Kincaid Grade

6. Тематическая рубрикация текстов в корпусе ТКиКЛ

В процессе формирования корпуса ТКиКЛ проводилась экспертная разметка текстов по рубрикам. Для этой цели была разработана схема рубрикации, содержащая темы работы секций конференций. Темы соответствуют названиям, предложенным челнами организационных комитетов конференций. Перед проведением экспертной разметки была осуществлена нормализация названий тем, результат представлен в таблице 5. Данная экспертная тематическая разметка будет использована для верификации результатов кластеризации текстов и тематических моделей, обученных на корпусе.

Таблица 5. Темы в схеме рубрикации текстов корпуса ТКиКЛ

№	Схема рубрикации
1	Общие вопросы корпусной лингвистики
2	Создание, разработка и применения корпусов
3	Статистические исследования на материале корпусов
4	Корпусы и лексикография
5	Морфология и синтаксис в корпусах
6	Семантика в корпусах
7	Обучающие корпусы
8	Исторические корпусы
9	Параллельные корпусы и машинный перевод
10	Речевые и мультимедийные корпусы
11	Корпусы художественных текстов

7. Заключение

В результате подготовки нового корпусного ресурса, включающего материалы конференции «Корпусная лингвистика» с 2002 по 2021 гг., семинара «Компьютерная лингвистика и вычислительные онтологии» с 2011 по 2023 гг., а также некоторые другие материалы, были проведены следующие работы: преобразование текстов в формат *.txt, фильтрация нетекстовых элементов, разметка метаданных (авторы, аффилиации, названия, наборы ключевых выражений, аннотации, названия конференций, годы издания, тематические рубрики, именованные сущности и т.д.). Была проведена унификация формата описания структуры текста, восстановлены лакуны — сгенерированы наборы ключевых выражений с применением алгоритма RuTrmExtract и аннотации с помощью алгоритмов экстрактивной и абстрактивной суммаризации.

Планы дальнейшего развития проекта включают в себя проведение экспертной разметки ключевых выражений, уточнение формата генерируемых аннотаций, проведение процедур тематического моделирования и разработка поискового сервиса для работы с данными корпуса ТКиКЛ.

Литература

[1] Митрофанова О. А., Захаров В. П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 321–328. URL: https://www.dialog-21.ru/digests/dialog2009/materials/pdf/49.pdf (дата обращения: 09.02.2024).

[2] Виноградова Н. В., Митрофанова О. А. Формальная онтология как инструмент систематизации данных в русскоязычном корпусе текстов по корпусной лингвистике // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008. С. 113–121.
URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (дата обращения: 09.02.2024).

[3] Виноградова Н. В., Митрофанова О. А., Паничева П. В. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды девятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский, 2007. URL: http://rcdl.ru/doc/2007/paper_31_v1.pdf (дата обращения: 15.02.2024).

[4] Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб., 2020. 234 с.

[5] Тихонова Е. В., Косычева М. А. Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology. 2022. Вып. 3 (4). С. 7–15. URL: https://elibrary.ru/item.asp?id=49446588 (дата обращения: 12.03.2024).

[6] Kamshilova O., Beliaeva L., Geikhman L. Author's Choice for Keyword List: Research Aspect // PRLEAL-2019. R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). Saint Petersburg, Russia, November 27, 2019. CEUR Workshop Proceedings. 2020. Vol. 2552. P. 47–59. URL: https://elibrary.ru/item.asp?id=42584043 (дата обращения: 12.03.2024).

[7] Митрофанова О. А., Гаврилик Д. А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Вып. 13 (4). С. 22–40. URL: https://elib.spbstu.ru/dl/2/j23-158.pdf/en/info (дата обращения: 25.03.2024).

[8] Гусева Д. Д., Митрофанова О. А. Ключевые выражения в русскоязычных научно-популярных текстах: сравнение восприятия устной и письменной речи с результатами автоматического анализа // Terra Linguistica. 2024. Вып. 15 (1). С. 20–35. URL: https://human.spbstu.ru/userfiles/files/articles/2024/1/20-35.pdf (дата обращения: 26.03.2024).

[9] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase extraction from the Russian corpus on Linguistics by means of KEA and RAKE algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS. P. 369–372. URL: https://elibrary.ru/item.asp?id=41112843 (дата обращения: 26.03.2024).

[10] Морозов Д. А. и др. Генерация ключевых слов для аннотаций русскоязычных научных статей / Морозов Д. А., Глазкова А. В., Тютюльников М. А., Иомдин Б. Л. // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2023. № 1. С. 54–66. URL: https://elibrary.ru/lxeizh (дата обращения: 28.03.2024).

[11] Aries A., Zegour D., Walid H. Automatic text summarization: What has been done and what has to be done // arXiv:1904.00688. 2019. P. 1–34. URL: https://arxiv.org/abs/1904.00688 (дата обращения: 29.03.2024).

[12] Nenkova A., McKeown K. Automatic summarization // Foundations and Trends in Information Retrieval. 2011. Vol. 5 (2–3). P. 103–233. URL: https://core.ac.uk/download/pdf/76383212.pdf (дата обращения: 30.03.2024).

[13] Allahyari M. et al. Text summarization techniques: a brief survey / Allahyari M., Pouriyeh S., ssefi M., Safaei S., Trippe E. D., Gutierrez J. B., Kochut K. // arXiv preprint arXiv:1707.02268. 2017. P. 397–405. URL: https://arxiv.org/abs/1707.02268 (дата обращения: 30.03.2024).

[14] Athugodage M., Mitrofanova O., Gudkov V. Transfer Learning for Russian Legal Text Simplification // Proceedings of the 3rd Workshop on Tools and Resources for People with

REAding DIfficulties (READI) @ LREC-COLING 2024. 2024. P. 59–69. URL: https://aclanthology.org/2024.readi-1.6/ (дата обращения: 30.03.2024).
[15] Gudkov V., Mitrofanova O., Filippskikh E. Automatically Ranked Russian Paraphrase Corpus for Text Generation // Proceedings of the Fourth Workshop on Neural Generation and Translation. Association for Computational Linguistics. 2020. P. 54–59. URL: https://aclanthology.org/2020.ngt-1.6/ (дата обращения: 30.03.2024).
[16] Pilault J. et al. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models / Pilault J., Li R., Subramanian S., Pal C. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2020. P. 9308–9319. URL: https://aclanthology.org/2020.emnlp-main.748/ (дата обращения: 30.03.2024).
[17] Automatic text summarizer // PyPI. URL: https://pypi.org/project/sumy/ (дата обращения: 30.03.2024).
[18] RuT5SumGazeta // Hugging Face. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (дата обращения: 30.03.2024).
[19] Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. Recognizing Named Entities in Specific Domain // Lobachevskii Journal of Mathematics. 2020. Vol. 41 (8). P. 1591–1602. URL: https://link.springer.com/article/10.1134/S199508022008020X (дата обращения: 30.03.2024).
[20] Костюк Д. М., Широков Н. К. Методы идентификации именованных сущностей в задачах обработки потока научных новостей // Менеджмент вузовских библиотек. Минск, 2021. С. 50–54. URL: https://elibrary.ru/item.asp?id=49171334 (дата обращения: 30.03.2024).
[21] Навроцкий А. А., Кривальцевич Е. В. Сравнительный анализ систем извлечения именованных сущностей из неструктурированных публицистических текстов // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня. Минск, 2020. С. 12–18. URL: https://elibrary.ru/item.asp?id=43934323 (дата обращения: 30.04.2024).
[22] Yadav V., Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models // Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 2018. P. 2145–2158. URL: https://arxiv.org/abs/1910.11470 (дата обращения: 02.04.2024).
[23] Natasha // GitHub Repository. URL: https://github.com/natasha/natasha (дата обращения: 02.02.2024).
[24] Yargy // GitHub Repository. URL: https://github.com/natasha/yargy (дата обращения: 02.02.2024).
[25] Named Entity Recognition (NER) // DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (дата обращения: 02.02.2024).
[26] NEREL // GitHub Repository. URL: https://github.com/nerel-ds/NEREL (дата обращения: 02.02.2024).
[27] Stanford NER // David Batista. URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (дата обращения: 02.02.2024).

Text Corpus on Corpus Linguistics: Composition and Stages of Formation

O. A. Mitrofanova, M. A. Adamova, L. A. Bukreeva, A. K. Zernova,
A. A. Litvinova, V. S. Pavlikova, P. S. Sologub

Saint–Petersburg State University

The article is dedicated to the challenges of creating a corpus of articles on corpus linguistics, which is being developed at the Department of Mathematical Linguistics of St. Petersburg State University (SPBU). The corpus is compiled under the supervision of V. P. Zakharov and includes texts from the «Corpus Linguistics» conference reports from 2002 to 2021, the «Computational Linguistics and Computational Ontologies» seminar from 2011 to 2023, as well as some other materials. During the development of the corpus resource, standardization of the text presentation format was carried out, and the structure of the articles was investigated. Experiments were conducted on the generation of keywords and annotations in cases where the original text did not contain this information. Types of named entities recorded in the corpus were examined, and an algorithm for their annotation was implemented. An analysis was conducted on the distribution of conference reports into thematic blocks according to the expert annotation scheme.

Keywords: corpus linguistics, conference materials, annotation, keywords, summaries, thematic annotation, named entities

Reference for citation: Mitrofanova O. A., Adamova M. A., Bukreeva L. A., Zernova A. K., Litvinova A. A., Pavlikova V. S., Sologub P. S. Text Corpus on Corpus Linguistics: Composition and Stages of Formation // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 13–29. DOI: 10.17586/2541-9781-2024-8-13-29.

Reference

[1] Mitrofanova O. A., Zaharov V. P. Avtomatizirovannyj analiz terminologii v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Kompyuternaya lingvistika i intellektualnye tekhnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog 2009» (Bekasovo, 27–31 maya 2009 g.). Vyp. 8 (15). M.: RGGU, 2009. S. 321–328. URL: https://www.dialog-21.ru/digests/dialog2009/materials/pdf/49.pdf (access date: 09.02.2024). (In Russian)

[2] Vinogradova N. V., Mitrofanova O. A. Formalnaya ontologiya kak instrument sistematizacii dannyh v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Trudy mezhdunarodnoj konferencii «Korpusnaya lingvistika – 2008». SPb., 2008. S. 113–121. URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (access date: 09.02.2024). (In Russian)

[3] Vinogradova N. V., Mitrofanova O. A., Panicheva P. V. Avtomaticheskaya klassifikaciya terminov v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Trudy devyatoj Vserossijskoj nauchnoj konferencii «Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollekcii» (RCDL–2007). Pereslavl-Zalesskij, 2007. URL: http://rcdl.ru/doc/2007/paper_31_v1.pdf (access date: 15.02.2024). (In Russian)

[4] Zaharov V. P., Bogdanova S. YU. Korpusnaya lingvistika. SPb., 2020. 234 s. (In Russian)

[5] Tihonova E. V., Kosycheva M. A. Effektivnye klyuchevye slova: strategii formulirovaniya // Health, Food & Biotechnology. 2022. Vyp. 3 (4). S. 7–15. URL: https://elibrary.ru/item.asp?id=49446588 (access date: 12.03.2024). (In Russian)

[6] Kamshilova O., Beliaeva L., Geikhman L. Author's Choice for Keyword List: Research Aspect // PRLEAL-2019. R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). Saint Petersburg, Russia, November 27, 2019. CEUR Workshop Proceedings. 2020. Vol. 2552. P. 47–59. URL: https://elibrary.ru/item.asp?id=42584043 (access date: 12.03.2024).

[7] Mitrofanova O. A., Gavrilik D. A. Eksperimenty po avtomaticheskomu vydeleniyu klyuchevyh vyrazhenij v stilisticheski raznorodnyh korpusah russkoyazychnyh tekstov // Terra Linguistica.

2022. Vyp. 13 (4). S. 22–40. URL: https://elib.spbstu.ru/dl/2/j23-158.pdf/en/info (access date: 25.03.2024). (In Russian)

[8] Guseva D. D., Mitrofanova O. A. Klyuchevye vyrazheniya v russkoyazychnyh nauchno-populyarnyh tekstah: sravnenie vospriyatiya ustnoj i pismennoj rechi s rezultatami avtomaticheskogo analiza // Terra Linguistica. 2024 Vyp. 15 (1). S. 20–35. URL: https://human.spbstu.ru/userfiles/files/articles/2024/1/20-35.pdf (access date: 26.03.2024). (In Russian)

[9] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase extraction from the Russian corpus on Linguistics by means of KEA and RAKE algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS. P. 369–372. URL: https://elibrary.ru/item.asp?id=41112843 (access date: 26.03.2024).

[10] Morozov D. A. i dr. Generaciya klyuchevyh slov dlya annotacij russkoyazychnyh nauchnyh statej / Morozov D. A., Glazkova A. V., Tyutyulnikov M. A., Iomdin B. L. // Vestnik NGU. Seriya: Lingvistika i mezhkulturnaya kommunikaciya. 2023. № 1. S. 54–66. URL: https://elibrary.ru/lxeizh (access date: 28.03.2024). (In Russian)

[11] Aries A., Zegour D., Walid H. Automatic text summarization: What has been done and what has to be done // arXiv:1904.00688. 2019. P. 1–34. URL: https://arxiv.org/abs/1904.00688 (access date: 29.03.2024).

[12] Nenkova A., McKeown K. Automatic summarization // Foundations and Trends in Information Retrieval. 2011. Vol. 5 (2–3). P. 103–233. URL: https://core.ac.uk/download/pdf/76383212.pdf (access date: 30.03.2024).

[13] Allahyari M. et al. Text summarization techniques: a brief survey / Allahyari M., Pouriyeh S., ssefi M., Safaei S., Trippe E. D., Gutierrez J.B., Kochut K. // arXiv preprint arXiv:1707.02268. 2017. P. 397–405. URL: https://arxiv.org/abs/1707.02268 (access date: 30.03.2024).

[14] Athugodage M., Mitrofanova O., Gudkov V. Transfer Learning for Russian Legal Text Simplification // Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024. 2024. P. 59–69. URL: https://aclanthology.org/2024.readi-1.6/ (access date: 30.03.2024).

[15] Gudkov V., Mitrofanova O., Filippskikh E. Automatically Ranked Russian Paraphrase Corpus for Text Generation // Proceedings of the Fourth Workshop on Neural Generation and Translation. Association for Computational Linguistics, 2020. P. 54–59. URL: https://aclanthology.org/2020.ngt-1.6/ (access date: 30.03.2024).

[16] Pilault J. et al. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models / Pilault J., Li R., Subramanian S., Pal C. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2020. P. 9308–9319. URL: https://aclanthology.org/2020.emnlp-main.748/ (access date: 30.03.2024).

[17] Automatic text summarizer // PyPI. URL: https://pypi.org/project/sumy/ (access date: 30.03.2024).

[18] RuT5SumGazeta // Hugging Face. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (access date: 30.03.2024).

[19] Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. Recognizing Named Entities in Specific Domain // Lobachevskii Journal of Mathematics. Vol. 41 (8). 2020. P. 1591–1602. URL: https://link.springer.com/article/10.1134/S199508022008020X (access date: 30.03.2024).

[20] Kostyuk D. M., Shirokov N. K. Metody identifikacii imenovannyh sushchnostej v zadachah obrabotki potoka nauchnyh novostej // Menedzhment vuzovskih bibliotek. Minsk, 2021. S. 50–54. URL: https://elibrary.ru/item.asp?id=49171334 (access date: 30.03.2024). (In Russian)

[21] Navrockij A. A., Krival'cevich E. V. Sravnitel'nyj analiz sistem izvlecheniya imenovannyh sushchnostej iz nestrukturirovannyh publicisticheskih tekstov // BIG DATA and Advanced Analytics = BIG DATA i analiz vysokogo urovnya. Minsk, 2020. S. 12–18. URL: https://elibrary.ru/item.asp?id=43934323 (access date: 30.04.2024). (In Russian)
[22] Yadav V., Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models // Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 2018. P. 2145–2158. URL: https://arxiv.org/abs/1910.11470 (access date: 02.04.2024).
[23] Natasha // GitHub Repository. URL: https://github.com/natasha/natasha (access date: 02.02.2024).
[24] Yargy // GitHub Repository. URL: https://github.com/natasha/yargy (access date: 02.02.2024).
[25] Named Entity Recognition (NER) // DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (access date: 02.02.2024).
[26] NEREL // GitHub Repository. URL: https://github.com/nerel-ds/NEREL (access date: 02.02.2024).
[27] Stanford NER // David Batista. URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (access date: 02.02.2024).

Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем

О. А. Митрофанова [1], Р. В. Голубев [1], П. А. Гусяцкая [1], К. В. Макеев [1], Е. А. Плюснина [1], Д. Д. Сухан [1,2], А. В. Трошина [1], А. А. Уткина [1]

[1] Санкт-Петербургский государственный университет, [2] Just AI

o.mitrofanova@spbu.ru, st110682@student.spbu.ru, st068584@student.spbu.ru, st110200@student.spbu.ru, st109958@student.spbu.ru, st110829@student.spbu.ru, st110338@student.spbu.ru, st110578@student.spbu.ru

Аннотация

В статье представлены результаты экспериментов по обучению семейства тематических моделей корпуса текстов по корпусной лингвистике, создаваемого на кафедре математической лингвистики СПбГУ под руководством В. П. Захарова. Тематическое моделирование корпуса ТКиКЛ осуществлено с помощью алгоритмов NMF, LSA, LDA, Biterm. Обобщение тем с помощью меток реализовано на основе обработки данных из выдачи информационно-поисковой системы, статических предсказывающих моделей Word2Vec, обученных на корпусе, а также большой языковой модели ChatGPT. Результаты тематического моделирования с назначением меток тем сопоставляются с данными о распределении докладов по тематическим блокам конференций в соответствии со схемой экспертной разметки.

Ключевые слова: корпусная лингвистика, материалы конференций, тематическое моделирование, метки тем, рубрикация

Библиографическая ссылка: Митрофанова О. А., Голубев Р. В., Гусяцкая П. А., Макеев К. В., Плюснина Е. А., Сухан Д. Д., Трошина А. В., Уткина А. А. Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 30–44. DOI: 10.17586/2541-9781-2024-8-30-44.

1. Введение

Данная статья посвящена решению задачи построения семейства тематических моделей корпуса текстов статей по корпусной лингвистике (далее корпус ТКиКЛ), составленного под руководством основателя Петербургской школы корпусной и компьютерной лингвистики Виктора Павловича Захарова. Разрабатываемый коллективом исследователей корпус представляет собой ценный источник информации о становлении и развитии методологии, ресурсов, понятийного аппарата и терминологии корпусной лингвистики. В статье [1] представлено описание процедуры формирования корпуса, разметки ключевых выражений в корпусе, генерации аннотаций и систематизации именованных сущностей. В статье [2] описаны эксперименты по формированию базы данных с метаинформацией и по разработке системы визуализации результатов поиска. Для улучшения качества поиска было проведено тематическое моделирование, результаты которого сопоставлены с ручной разметкой рубрик.

Под тематическим моделированием традиционно понимается особый способ построения структурно-семантической модели корпуса текстов, которая определяет взаимосвязи тем, документов и слов-тематизаторов [3]. Темы рассматриваются как скрытые факторы, представленные кластерами слов-тематизаторов. Каждый документ связан с одной или несколькими темами с некоторой вероятностью, при этом темы могут пересекаться. Наиболее распространенные методы тематического моделирования включают группу алгебраических моделей, например, латентный семантический анализ (Latent Semantic Analysis, LSA), неотрицательная матричная факторизация (Non-negative Matrix Factorization, NMF) и др., группу вероятностных моделей, например, вероятностный латентный семантический анализ (probabilistic Latent Semantic Analysis, LSA), латентный Распределение Дирихле (Latent Dirichlet Allocation, LDA) и т. д. В практических задачах широко используются мультимодальные версии тематических моделей, учитывающие дополнительные параметры корпусов (авторство текстов, время создания документов в корпусе, иерархия тем и т. д.), комбинируемые с моделями распределенных векторных вложений, например, BERTopic.

В статье представлены результаты построения тематических моделей корпуса ТКиКЛ с помощью алгоритмов NMF, LSA, LDA и Biterm. Параметры экспериментов были сходными, что обеспечивает объективность описания и сопоставления полученных результатов. Во всех экспериментах мы придерживались следующей схемы: проведение предварительной разметки n-грамм с помощью алгоритма выделения ключевых выражений RAKE [4], построение серии моделей с различным числом тем (5, 10, 15, 20), число слов-тематизаторов в темах (10, 15 20). В примерах, приводимых далее в статье, сохранено форматирование выдачи тематических моделей, предполагающее декапитализацию и отдельные случаи сохранения текстов в нелемматизированном варианте. Для оценки качества и интерпретируемости полученных моделей были определены значения агрегированной когерентности [5], перплексии [6] и энтропии [7]. Отдельные случаи расширения схемы экспериментов оговариваются в соответствующих разделах.

Тематические модели обеспечивают предпосылки для разведочного поиска в корпусах текстов. Повышение интерпретируемости моделей должно способствовать улучшению качества извлечения информации их текстов. Одним из факторов, влияющих на интерпретируемость тематических моделей, их адекватность решаемым задачам и исходным данным, является возможность обобщения тем с помощью меток [8; 9; 10]. Метка темы — это слово или словосочетание, отражающее общее содержание темы. Согласно традиции, темы условно обозначаются с помощью номера и первого слова-тематизатора, которое далеко не всегда является самым общим или типичным относительно темы. В автоматическом понимании текста разработаны формальные методы назначения меток тем, различающиеся источниками меток (внешними по отношению к корпусу и внутренними, использующими информацию из целевого корпуса), структурой меток (униграммы, биграммы, триграммы и т. д., выделяющиеся по лексико-грамматическим шаблонам), типами используемых алгоритмов. В [11] представлена апробация методов назначения меток тем на основе ИПС, с применением предсказаний дистрибутивно-семантических моделей и больших языковых моделей ChatGPT. Руководствуясь тем, что в экспериментах с корпусом научных новостных сообщений данный набор методов хорошо зарекомендовал себя, было принято решение воспроизвести его в проекте по тематическому моделированию корпуса ТКиКЛ.

2. Результаты тематического моделирования корпуса ТКиКЛ

2.1. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма NMF

Алгоритм NMF (Non-Negative Matrix Factorization, неотрицательная матричная факторизация) [12; 13] заключается в поиске для некой неотрицательной матрицы X двух

матриц (W, H), чье произведение будет являться приближением оригинальной матрицы X. В контексте тематического моделирования текстовых данных это означает, что для исходного корпуса подбираются матрицы «слова — темы» и «темы — документы», показывающие, соответственно, какие слова характеризуют каждую из тем, и как темы распределены по документам. К преимуществам алгоритма NMF относят высокую интерпретируемость результатов, вытекающую из неотрицательности элементов матриц, а также способность выявлять в данных более редкие и специфичные темы.

В настоящем проекте алгоритм NMF был применен к текстовым данным: каждая тема, таким образом, представляет из себя ранжированный по весам список слов и словосочетаний. Реализация алгоритма тематического моделирования NMF была осуществлена при помощи библиотеки scikit-learn [14]. Ход экспериментов предполагал предварительную разметку в корпусе униграмм и биграмм на основе алгоритма RAKE. Все слова, выделенные в составе биграмм, были затем удалены из неразмеченных текстов корпуса на этапе формирования списка уникальных униграмм. Биграммы были затем лемматизированы и представлены в корпусе в виде «практический_применение» — чтобы при обучении тематической модели для каждой биграммы формировался отдельный вектор. Список объединенных и лемматизированных биграмм был объединен со списком лемматизированных униграмм.

Далее была проведена серия экспериментов, целью которой было установить оптимальное количество тем, которые будет выделять модель NMF на корпусных данных. Для этого была проведена оценка когерентности для четырех моделей, выделивших 5, 10, 15 и 20 тем, соответственно. Наивысший показатель когерентности в группе моделей (≈ 0,44) был достигнут при 20 темах – это количество тем и было принято как рабочее значение параметра в финальной модели NMF. Данная модель была построена на корпусных данных, векторизованных при помощи метрики TF-IDF, в результате чего было получено 20 тем, представляющих собой отранжированные списки тематизаторов — униграмм и биграмм.

Результирующие темы демонстрируют не только высокий показатель когерентности, но и представляются достаточно интерпретируемыми при экспертной оценке. К примеру, в модели четко противопоставлены темы «Перевод» (*перевод, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо-, переводчик, многоязычный, словарь…*), «Коммуникативные стратегии» (*жест, ребенок, коммуникативный, движение, робот, мимика, поведение, эмоциональный, детский, рука, участник, человек, невербальный, социальный, собеседник, коммуникация, возраст, функция, мультимодальный, действие…*), «Медиапространство» (*театр, театральный, спектакль, сцена, зритель, интернет, новый, трансляция, александринский, режиссер, пространство, многопотоковый, видео, творческий, виртуальный, технология, театра, интерактивный, медиа, актер…*) и т.д.

Отдельно стоит упомянуть, что модель NMF успешно выделила в корпусе редкие темы, то есть темы, представленные в корпусе лексикой низкой частотности, к примеру тема «Финно-угорские языки» (*финский, ижорский, диалектный, диалект, песня, народный, прибалтийско-, карельский, говор, фонетический, вепсский, ингерманландия, топоним, топонимический, язык, текст, приток, топонимов, песен, звуковой…*), или «Тибетский язык» (*тибетский, разметка, грамматический, композит, тэг, токен, лексический, корпус, аффикс, буддийский, индийский, термин, разметить, традиция, сегментация, проект, трактат, традиции…*).

Заметим, что результаты применения алгоритма NMF на настоящий момент следует считать ориентиром в построении тематических моделей корпуса ТКиКЛ.

2.2. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма LSA

Латентный семантический анализ (LSA, Latent Semantic Analysis) — классический алгоритм построения дистрибутивных моделей корпусов текстов, основанный на матрично-векторных преобразованиях и отражающий близость значений и совместную встречаемость слов в корпусе [15; 16; 17]. Принцип работы LSA можно разбить на несколько этапов. На первом шаге текст предобрабатывается, затем токенам назначают веса (например, с помощью TF-IDF), и по этим весам строится матрица. В данном проекте для предобработки использовалась функция CountVectorizer в библиотеке scikit-learn [14]. Она токенизирует текст, после чего производит расчет вхождений, каждому токену присваивается уникальный целочисленный индекс, и эта информация приводится в формат матрицы. На финальной ступени матрица раскладывается методом сингулярного разложения (SVD, Singular Value Decomposition).

В экспериментах по обучению моделей LSA, согласно стандартной схеме, лучшие результаты были получены при выборе 15 тем. В этом случае темы четче разграничиваются, но при этом остаются довольно специализированными. Ниже приведены примеры общих тем: «Корпус как явление» (*текст, корпус, слово, являться, язык, система, русский, работа, данные, анализ, семантический, значение, информация, словарь, результат, иметь, использование, использовать, исследование, информационный…*); «Модель языка» (*слово, понятие, модель, термин, связь, язык, знание, отношение, электронный, семантический, корпус, развитие, определение, область, услуга, государственный, слов, поле, метафора, информационный…*); к более частным следует отнести лингвистические темы: «Семантика» (*слово, семантический, значение, ударение, иметь, глагол, понятие, класс, связь, отношение, стих, являться, предлог, объём, строка, определение, слов, модель, вид, часть…*); «Теория поэзии» (*ударение, объём, стих, строка, текст, слоговой, слог, слово, ударный, метр, икт, пропуск, место, электронный, показатель, интервал, схема, объёмный…*); прикладные лингвистические темы «Социальная сеть» (*социальный, сеть, пользователь, интернет, политический, сети, новый, сетевой, пространство, являться, человек, сми, текст, связь, коммуникация, исследование, аудитория, медиа, количество, сервис…*); «Электронное голосование» (*голосование, система, электронный, избиратель, голос, голосования, выборы, избирательный, список, интернет, проблема, слово, цифровой, ключ, бюллетень, возможность, дистанционный, помощь, кандидат, использовать…*).

Полученные темы органично вписываются в спектр исследовательских направлений, представленных в корпусе ТКиКЛ, однако для повышения интерпретируемости результатов модель LSA требует более точной настройки.

2.3. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма LDA

Скрытое распределение Дирихле (Latent Dirichlet allocation, LDA) — это широко используемый алгоритм вероятностного тематического моделирования, рассматривающий процесс определения тематической структуры текстов на основе семейства непрерывных многомерных вероятностных распределений [18]. Как известно, тематическая модель LDA частично решает проблему переобучения pLSA [19].

В экспериментах использовалась реализация алгоритма LDA в библиотеках scikit-learn [14] и genism [20]. Предобработка корпуса предполагала разметку в корпусе ключевых выражений (биграмм и триграмм) посредством алгоритма RAKE. Результаты обучения тематических моделей в библиотеках scikit-learn и genism отличаются долей неоднословных тематизаторов: например, LDA в scikit-learn в основном выделяет тематизаторы-униграммы, и единственной биграммой оказалось словосочетание «*социальная_сеть*», в то время как LDA в gensim генерирует темы с высокой долей биграмм и триграмм. С точки зрения интерпретируемости тем и равномерности распределения тем по документам следует отдать предпочтение варианту реализации LDA в библиотеке scikit-learn. Было проведено обучение серии моделей со сменой параметров (5, 10, 15 и 20 тем) и оценкой

когерентности. В результате экспериментов было установлено, что оптимальное число интерпретируемых наборов слов для LDA стремится к двадцати.

Среди полученных тем есть ядерные темы общего содержания, связанные с общей проблематикой корпуса ТКиКЛ, в частности, «Моделирование естественного языка» (*модель, алгоритм, формула, критерий, подобный, список, пример, метод, часть, параметр, следующий, ошибка, использовать…*), «Корпус текстов» (*корпус, словарь, русский, исследование, база, поиск, термин, материал, разметка, лингвистический, картотека, дескриптор, словоформа, анализ, лингвистика…*), «Представление текстов в корпусе» (*корпус, буква, словоформа, контекст, русский, написание, житие, вариант, словарь, рукопись, век, семантический, первый, термин, разметка…*). Примерно четверть сгенерированных тем соотносится с задачами семантического анализа, например: «Семантическая разметка» (*семантический, разметка, отношение, корпус, значение, разный, связь, признак, лингвистика, анализ, система, лексика, тип, общий…)*, «Формальные онтологии» (*онтология, корпус, понятие, отношение, возможность, класс, рамка, описание, элемент, слот, использование, экземпляр, система, иерархия, онторедактор…*), «Семантические отношения» (*отношение, семантический, словарь, синсет, значение, лексический, структура, существительное, связь, система, лексико-, глагол, база, часть, словоформа…*) и некоторые другие. Наряду с этим, одиночные темы представляют такие специфичные направления компьютерной и корпусной лингвистики, как «Морфосинтаксическая разметка» (*предложение, связь, оценка, корпус, синтаксический, парсер, узел, структура, отношение, этап, морфологический, речь, тип, случай, часть…*), «Звуковые корпусы текстов» (*речь, эда, материал, русский, устный, рассказ, речевой, корпус, живой, языковой, составлять, звуковой, вариант, запятая, точка…*).

2.4. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма Biterm

Модель битермов (Biterm Topic Model) [21] создана для распознавания тем в коротких текстах, таких как твиты и посты социальных сетей. Модели типа LSA или LDA недостаточно приспособлены для обработки текстов данного типа, поскольку неявно учитывают совместную встречаемость слов в документах, что проявляется в виде тематической разреженности данных в коротких текстах. BTM генерирует темы, напрямую моделируя шаблоны битермов (биграмм) для всего корпуса текстов. Восстановление битермов по шаблонам улучшает качество тем, а агрегированные шаблоны битермов для корпуса в целом решают проблему тематической разреженности.

BTM справляется с задачей тематического моделирования благодаря сэмплированию по Гиббсу. Основная идея сэмплирования заключается в генерации образцов из совместного распределения вероятностей путём итеративной выборки из условных распределений каждой переменной с учётом значений всех остальных переменных. Этот процесс позволяет получать выборки из сложных, высокоразмерных распределений, разбивая задачу на более простые условные шаги выборки. Сэмплирование по Гиббсу — это разновидность метода Монте-Карло с цепью Маркова, широко используемого в байесовской статистике и машинном обучении для аппроксимации апостериорных распределений.

Для экспериментов использовалась библиотека bitermplus [22]. На входе модели, помимо корпуса текстов нужно подать предполагаемое число тем. В ходе эксперимента были протестированы 5 значений: 5, 8, 10, 15 и 20 тем (здесь к стандартной схеме было добавлено еще одно значение — 8 тем). Примеры результирующих тем приведены ниже: «Информационные технологии» (*информационный, электронный, система, государственный, развитие, являться, научный, информация, работа, использование…*), «Корпус» (*текст, корпус язык, работа, словарь, система, анализ, данные, русский, разметка…*), «Семантика» (*слово, значение, семантический, являться, глагол, текст, форма, тип, случай, два…*) и т.д. Наибольшая когерентность наблюдается у модели с пятью темами, однако перплексия и энтропия у данной модели выше, чем у остальных. Средняя

когерентность почти не меняется с возрастанием количества тем, а энтропия даже падает. Результаты показывают, что оптимальное число тем определяется в промежутке между 16 и 20 темами.

3. Результаты генерации меток тем в корпусе ТКиКЛ

3.1. Генерация меток тем для текста с помощью ИПС

Одним из вариантов генерации меток тем является применение информационно-поисковых систем (ИПС). ИПС и тематическое моделирование тесно связаны, однако обычно именно метки тем используются для улучшения веб-поиска, обратная же схема встречается редко. Поскольку в основе всех современных ИПС лежит принцип отбора наиболее релевантных запросу документов, можно использовать это свойство для решения поставленной задачи. Так как тексты веб-документов могут быть разного объёма и содержать разнородную информацию, удобнее реализовать схему, при которой результаты выдачи используются не полностью, а частично. Можно рассмотреть два варианта — суммаризацию всего документа, например, с помощью моделей-трансформеров, либо использование только заголовков веб-страниц без учета основной части текста. Может показаться, что вторая опция ограничивает наши возможности, однако она более выгодна, так как при оценке релевантности ИПС учитывает заголовки с большим весом. Кроме того, это обеспечивает бОльшую вероятность получения связного текста, в отличие от автоматической суммаризации.

Методика генерации меток тем с помощью ИПС, примененная в нашем исследовании, является модификацией метода, представленного в [8; 11], и так же, как и исходный вариант, состоит из нескольких этапов.

На первом этапе в качестве входных данных используются списки тем, сгенерированных исследуемыми алгоритмами тематического моделирования. Для каждой метки тем отправляется запрос в ИПС Google [23] для получения поисковой выдачи. При этом все метки рассматриваются как единое предложение, как и в случае обычных запросов в ИПС, которые не всегда характеризуются синтаксической связностью. Использование Google обусловлено тем, что эта поисковая система не блокирует последовательные автоматические запросы к своему API, в отличие от Yandex. Полученная выдача далее фильтруется, рассматриваются 30 первых по релевантности документов.

На втором этапе для всех заголовков темы составляется матрица совместной встречаемости слов в контекстном окне [-1, 1], что позволяет выделить биграммы. Такую матрицу можно визуализировать как взвешенный граф, где рёбра — это связи в биграммах, а веса — встречаемость. Для слов, не встречавшихся друг с другом, устанавливается минимальный вес, равный 1.

Третий этап определяется как Power Iteration или применение степенного метода. Он используется также и в PageRank, алгоритме, имеющем ключевое значение в современных ИПС. Из матрицы, составленной на предыдущем этапе, собирается стартовое состояние: это словарь, где ключи — это все слова, а значения равны величине, обратной количеству слов. Затем в матрице совместной встречаемости все значения делятся на сумму элементов в этой строке. Наконец, запускается алгоритм сходимости, в ходе которой предыдущее стартовое состояние заменяется скалярным произведением предыдущего словаря на матрицу совместной встречаемости. Это происходит либо фиксированное число раз (в нашем алгоритме — 1000), либо пока разница между предыдущим и новым состоянием не становится меньше некоторого числа ε. Иначе говоря, рассчитывается собственный вектор для матрицы. Затем все слова сортируются по весу и из них отбирается n наиболее вероятных.

Два финальных этапа представляют собой формирование и фильтрацию полученных меток на основе правил. Для набора заголовков темы составляется список n-грамм.

Биграммы и триграммы составляются по правилам, а большие — из меньших по принципу паззла (конец одной биграммы равен началу предыдущей). Таким образом, максимальный размер n-грамм — 6 токенов. Правила составления n-грамм направлены на формирование наиболее частотных для русского языка словосочетаний — например, ADJ + NOUN или NOUN + NOUN (GEN, INSTR). Каждая n-грамма получает вес, равный сумме весов её составляющих, после чего отбираются первые 5 n-граммов по весу.

Постобработка включает фильтрацию n-грамм по следующим правилам: удаление повторов и совпадений; удаление меток с повторениями одного и того же слова; коррекция согласования внутри словосочетаний; отсеивание слишком коротких слов или случайных букв. На выходе метод производит от одной до трех биграмм для каждой темы, как правило, являющиеся осмысленными словосочетаниями. Примеры результирующих меток представлены в таблице 1.

Таблица 1. Примеры меток тем, сгенерированных ИПС

Темы (фрагмент выдачи)	**Метки ИПС**
онтология, понятие, свойство, отношение, знание, термин, связь, сущность, определение, объект, онтологии, класс, предметный, смысл, система, модель, являться, граф, множество, семантический…	*знания и онтология, онтология и тезаурусы, подход к процессам и системы*
мера, биграмм, коллокация, частота, словосочетание, коллокат, статистический, слово, сочетание, встречаемость, список, мер, словосочетаний, сочетаемость, значение, коллокаций, оценка, результат, корпус, эксперимент…	*выделение коллокаций, значение коллокаций, образование и наука*
учебный, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа…	*студенты и преподаватели в процессе, традиции к инновациям в обучении, инновации в обучении*
перевод, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо, переводчик, многоязычный, словарь…	*перевод для выравнивания, переводчик и редактор, поиск и ранжирование*
житие, текст, цитата, агиографический, житийный, рукопись, скат, разметка, написание, текста, древнерусский, рукописный, словоуказатель, издание, рукописи, фрагмент, дионисий, алексеева, глушицкого, представление…	*разметка в корпус агиографический текст, корпус агиографический текст, представление и анализ элементов структуры*
государственный, электронный, орган, гражданин, информационный, услуга, развитие, власть, социальный, правительство…	*информация о орган государственный власть, министерство социальный политика и труд, услуги для граждан и бизнес*
модель, алгоритм, тематический, слово, метод, текст, документ, тема, оценка, слов, результат, анализ, задача, распределение, количество, эксперимент, матрица, коллекция, вероятность, вероятностный…	*качество тематический модель для задача, плотность многомерных распределений в виде, методология и методы научных исследований*
…	*…*

Среди преимуществ использования ИПС для генерации меток тем следует отметить возможность генерации меток различной длины, более высокий уровень согласованности и объективности итоговых меток благодаря отбору релевантных комбинаций, а также достаточно высокую скорость исполнения. Метод способен продуцировать длинные осмысленные сочетания, такие как «типология ассоциативных словарей русского языка» или «основы цифровой грамотности и кибербезопасность». Кроме того, применение

поисковых методик понижает нестабильность некоторых базовых моделей, прежде всего, LDA.

Однако применение ИПС для тематического моделирования обладает и недостатками. В их числе сложная для настройки структура, построенная на разных принципах. Результаты не детерминированы, как и в случае нейросетей, и зависят от результатов работы поисковых систем, которые периодически обновляются. Результат зависит и от выбранной методики тематического моделирования: в данном исследовании лучший результат был получен для моделей LSA и NMF, где метки оказались более интерпретируемы. Метод сложно заставить производить фиксированное количество меток, а в некоторых случаях он может не создать ни одной. Поэтому рекомендуется использовать ИПС наряду с другими методами для получения лучшего результата.

3.2. Генерация меток тем для текста с помощью дистрибутивно-семантических моделей Word2Vec

Данный способ генерации меток тем относится к числу подходов, позволяющих назначать метки тем на основе внутренних по отношению к корпусу источников. Как предлагается в [9; 10; 11], мы применяли статические дистрибутивно-семантические модели типа Word2Vec [24] и рассматривали их предсказания как кандидаты в метки тем. Нейросетевая архитектура Word2Vec представляет контексты корпуса в виде векторов, которые при условии близости значения и употребления слов локализуются сходным образом, о чем свидетельствуют высокие значения косинусной меры. В Word2Vec предсказание близких лексических единиц осуществляется с помощью функции most_similar, допускающей генерацию ассоциатов как для отдельного слова, так и для группы слов, в нашем случае представляющей собой набор слов-тематизаторов, представляющих отдельную тему. Для предсказания кандидатов в метки тем на предобработанном корпусе ТКиКЛ были обучены две модели CBOW (Continuous Bag of Words) и Skip-gram, которые по-разному фиксируют отношения между словами в модели: если модель CBOW предсказывает потенциальные замены целевого слова с учетом контекста, то модель Skip-gram позволяет предсказывать элементы контекстного окружения для целевого слова. Для корректного обучения моделей корпус ТКиКЛ был токенизирован и повторно лемматизирован, для исключения попадания служебных слов и иных незнаменательных элементов в метках тем был подключен стоп-словарь. Частеречная разметка корпуса для того, чтобы исключить попадание наречий, прилагательных и иных частей речи кроме существительных в метки тем. Результаты экспериментов дают основания для дискуссии о статусе сгенерированных меток, которые действительно уточняют содержание тем, однако не обобщая их, а скорее расширяя. Примеры меток CBOW и Skip-gram представлены в таблице 2. Совпадения предсказаний двух типов моделей указывают на то, что повторяющиеся кандидаты в метки (выделены полужирным) являются релевантными для тем. Модели Word2Vec, обученные на корпусе с предварительной разметкой ключевых выражений с помощью алгоритма RAKE, генерируют повторяющиеся метки, что следовало бы избежать. Наиболее интерпретируемые результаты были получены в комбинации моделей Word2Vec и тем, порожденных моделями LSA и NMF.

Таблица 2. Примеры меток тем, сгенерированных моделями Word2Vec

Темы (фрагмент выдачи)	CBOW	Skip-gram
научный, система, информационный, библиотека, сервис, пользователь, поиск, ресурс, электронный, данные, информация, полнотекстовый, база, поддержка, программный, проект, запрос, поисковый, публикация, доступ…	*интерфейс, карта, контент, пользовательский*	*вебсайт, протокол, ипс, навигация*

Продолжение таблицы 2

Темы (фрагмент выдачи)	CBOW	Skip-gram
онтология, понятие, свойство, отношение, знание, термин, связь, сущность, определение, объект, онтологии, класс, предметный, смысл, система, модель, являться, граф, множество, семантический…	*иерархия, именовать, элементарный, родовидовой*	*таксономия, экземпляр, иерархия, родовидовой*
словарь, русский, словарный, слово, языка, словаря, язык, лексикографический, корпус, толковый, картотека, грамматический, словарей, академический, цитата, бас, слова, новый, создание, рнк…	*словник, двуязычный, одноязычный, зализняк*	*указатель, словник, грамматик, бумажный*
морфологический, синтаксический, разметка, слово, грамматический, форма, предложение, словоформа, разбор, существительное, неоднозначность, автоматический, ошибка, парсер, омонимия, часть, анализатор, снятие, падеж, вариант…	*частеречной, тег, помета, лемматизация*	*частичный, частеречной, морфема, лемматизация*
учебный, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа…	*перспектива, филология, методический, методология*	*учитель, будущий, привлечение, методический*
семантический, глагол, значение, валентность, слово, лексический, контекст, синсет, лексико-, актант, употребление, иметь, описание, отношение, семантика, класс, единица, синтаксический, языковой…	*ядро, синонимия, стилистический, номинация*	*сосед, синонимия, корень, синтаксема*
модель, алгоритм, тематический, слово, метод, текст, документ, тема, оценка, слов, результат, анализ, задача, распределение, количество, эксперимент, матрица, коллекция, вероятность, вероятностный…	*статистика, гипотеза, ранжирование, классификатор*	*lda, отзыв, классификатор, кластеризация*
…	…	…

3.3. Генерация меток тем для текста с помощью большой языковой модели ChatGPT

Проводимое исследование открывает новые возможности в тестировании больших языковых моделей, в частности, мощной языковой модели ChatGPT, созданной OpenAI. В эксперименте использовалась модель GPT-3.5 [26]. Модель способна генерировать текст, имитируя стиль носителя языка и понимая контекст. Далее рассмотрим применение ChatGPT к генерации меток с учетом весов слов (в этом состоит модификация протокола, реализованного в сходных задачах [11; 25]).

В ходе эксперимента на вход модели подавались темы в виде наборов слов-тематизаторов с их весами. При обращении к модели использовались промпты, например, «Используя слова из списка, составь несколько общих выражений и выдели главное слово». Предсказанные кандидаты в метки тем и ключевые слова были сохранены в таблицах Excel. ChatGPT успешно выделял общие выражения, адекватно отражающие тематику текстов корпуса ТКиКЛ, однако в ходе анализа данных были выявлены некоторые особенности. Во-первых, ChatGPT при использовании одного и того же чата ChatGPT может запоминать структуру диалога и тем самым в предсказаниях опирается на излишне широкий контекст, что приводит к семантическим сдвигам в генерации меток тем. Во-вторых, веса слов-

тематизаторов оказывают влияние на процесс генерации, что может привести к искажению результатов. В-третьих, в промпте желательно явно указывать ожидаемое количество кандидатов в метки для получения реалистичных результатов.

ChatGPT демонстрирует высокий потенциал для генерации меток, но требует аккуратного использования. Память контекста и влияние весов могут сказаться на точности результата. Верификация полученных меток была проведена с привлечением схемы рубрикации, содержащей темы работы секций конференций [1]. Рубрики из схемы экспертной рубрикации приведены в последней колонке таблицы 3.

Таблица 3. Примеры меток тем, сгенерированных ChatGPT

Темы (фрагмент выдачи)	**Метки ChatGPT**	**Главное слово в теме, выбор ChatGPT**	**Рубрика из схемы экспертной рубрикации**
научный, система, информационный, библиотека, сервис, пользователь, поиск, ресурс, электронный, данные, ***информация****, полнотекстовый, база, поддержка, программный, проект, запрос, поисковый, публикация, доступ*	*«Информационные ресурсы в научных исследованиях», «Электронные библиотеки и доступ к информации», «Сервисы для поиска научных данных»*	*информация*	Создание, разработка и применения корпусов
мера*, биграмм, коллокация, частота, словосочетание, коллокат, статистический, слово, сочетание, встречаемость, список, мер, словосочетаний, сочетаемость, значение, коллокаций, оценка, результат, корпус, эксперимент*	*«Статистический анализ коллокаций в корпусе», «Оценка частоты и сочетаемости слов», «Меры и статистика в анализе словосочетаний»*	*мера*	Статистические исследования на материале корпусов
словарь*, русский, словарный, слово, языка, словаря, язык, лексикографический, корпус, толковый, картотека, грамматический, словарей, академический, цитата, бас, слова, новый, создание, рнк*	*«Лексикографический анализ русского языка», «Словарные ресурсы и их создание», «Грамматический аспект в словарях»*	*словарь*	Корпусы и лексикография
морфологический, синтаксический, ***разметка****, слово, грамматический, форма, предложение, словоформа, разбор, существительное, неоднозначность, автоматический, ошибка, парсер, омонимия, часть, анализатор, снятие, падеж, вариант*	*«Морфологическая и синтаксическая разметка», «Грамматический анализ предложений», «Автоматический разбор текста и ошибки»*	*разметка*	Морфология и синтаксис в корпусах
учебный*, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа*	*«Обучение и образование с использованием корпусов», «Английский язык в образовательной среде», «Корпусный анализ ошибок студентов»*	*учебный*	Обучающие корпусы

Продолжение таблицы 3

Темы (фрагмент выдачи)	Метки ChatGPT	Главное слово в теме, выбор ChatGPT	Рубрика из схемы экспертной рубрикации
***перевод**, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо, переводчик, многоязычный, словарь*	*«Перевод и анализ текстов в параллельном корпусе», «Машинный перевод и создание терминологических словарей», «Текстовый анализ и создание переводческих ресурсов»*	*перевод*	Параллельные корпусы и машинный перевод
***житие**, текст, цитата, агиографический, житийный, рукопись, скат, разметка, написание, текста, древнерусский, рукописный, словоуказатель, издание, рукописи, фрагмент, дионисий, алексеева, глушицкого, представление*	*«Житийные тексты и их агиографические особенности», «Написание и разметка древнерусских рукописей», «Текстологический анализ житийных материалов»*	*житие*	Исторические корпусы
…	…	…	…

4. Заключение

Разработка специализированных корпусов текстов, к которым относится корпус ТКиКЛ, созданный коллективом авторов под руководством В. П. Захарова, требует как тщательной подготовки текстов, так и создания инструментария для автоматизации извлечения, и структурирования информации в корпусе. По этой причине столь важно успешное проведение экспериментов по тематическому моделированию корпуса ТКиКЛ, которое показало особенности структурно-семантической и тематической организации корпуса. Важно отметить, что в построении тематических моделей корпуса авторы следовали принципу мультимодальности и учитывали возможность совмещения базового протокола тематического моделирования с автоматическим выделением ключевых выражений и автоматическим назначением меток тем. Наиболее интерпретируемыми оказались результаты, полученные с помощью алгоритма NMF с метками тем, сгенерированными с помощью ChatGPT. Объективность полученных результатов подтверждается соответствием между автоматически назначенными метками и рубриками их схемы экспертной разметки, составленной на основе программ работы конференций, материалы которых представлены в корпусе ТКиКЛ.

Литература

[1] Митрофанова О. А., Адамова М. А., Букреева Л. А., Зернова А. К., Литвинова А. А., Павликова В. С., Сологуб П. С. Корпус текстов по корпусной лингвистике: состав и этапы формирования // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 12–28. DOI: 10.17586/2541-9781-2024-8-12-28.

[2] Сухан Д. Д., Плюснина Е. А. Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии.

Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 44–59. DOI: 10.17586/2541-9781-2024-8-44-59.

[3] Воронцов К. В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. М.: URSS, 2023. 208 с.

[4] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings. FRC CSC RAS. 2018. P. 369–372.

[5] Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. P. 262–272.

[6] Heinrich G. Parameter estimation for text analysis: Technical report. 2005. P. 1-32.

[7] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. 2018. Т. 512. P. 1192–1204.

[8] Ерофеева А., Митрофанова О. Автоматическое назначение меток тем в тематических моделях русскоязычных корпусов текстов // Структурная и прикладная лингвистика. Т. 12. СПб., 2019. С. 122–147.

[9] Kriukova A., Erofeeva A., Mitrofanova O., Sukharev K. Explicit Semantic Analysis as a Means for Topic Labeling // Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings. Springer, Cham. 2018. P. 167–177.

[10] Mitrofanova O., Kriukova A., Shulginov V., Shulginov V. E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science. Springer. 2021. Vol. 1357. P. 102–114.

[11] Mitrofanova O. A., Athugodage M. M., Ten L. V. Topic Label Generation in the Popular Science Corpus // Digital Geography: Proceedings of the International Conference on Internet and Modern Society (IMS 2023). Springer, 2023. (В печати)

[12] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. 2020. Vol. 12469, pt. 2. P. 134–152.

[13] Kuang D., Choo J., Park H. Nonnegative matrix factorization for interactive topic modeling and document clustering // Partitional clustering algorithms. 2015. P. 215–243.

[14] Scikit-learn // Scikit-learn. URL: https://scikit-learn.org/ (дата обращения: 09.02.2024).

[15] Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25 (2–3). P. 259–284.

[16] Чижик А. В. Использование методов тематического моделирования для оценки степени влияния СМИ на общественное настроение // Компьютерная лингвистика и вычислительные онтологии. Вып. 5. (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24–26 июня 2021 г. Сборник научных статей). СПб.: Университет ИТМО, 2021. С. 70–78.

[17] Кирина М. А. Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 2. С. 93–109.

[18] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. Vol. 3. P. 993–1022.

[19] Hofmann T. Probabilistic latent semantic indexing // ACM SIGIR Forum. 2017. Vol. 51 (2). P. 211–218.
[20] Gensim // Gensim. URL: https://radimrehurek.com/gensim/ (дата обращения: 09.02.2024).
[21] Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts // WWW 2013. Proceedings of the 22nd International Conference on World Wide Web. 2013. P. 1445–1456.
[22] biterm 0.1.5 // PyPI. URL: https://pypi.org/project/biterm/ (дата обращения: 09.02.2024).
[23] Google // Google. URL: https://www.google.ru/ (дата обращения: 09.02.2024).
[24] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 09.02.2024).
[25] Митрофанова О. А. Поиск и ранжирование текстов в специальном корпусе на основе тематического моделирования // Труды Международной конференции «Корпусная лингвистика — 2023», SPb Corpora 2023, 21–23 июня 2023 г. СПб.: Изд-во СПбГУ, 2024. (В печати)

Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment

O. A. Mitrofanova [1], R. V. Golubev [1], P. A. Gusyatskaya [1], K. V. Makeev [1], E. A. Pliusnina [1], D. D. Sukhan [1,2], A. V. Troshina [1], A. A. Utkina [1]

[1] Saint–Petersburg State University, [2] Just AI

The article presents novel experimental results concerning experiments aimed at training a family of topic models of the corpus on Corpus Linguistics, developed at the Department of Mathematical Linguistics, St. Petersburg State University under the supervision of V. P. Zakharov. Topic modelling of the corpus was carried out using NMF, LSA, LDA, Biterm algorithms. Ggeneralization of topics using labels is implemented on the basis of processing data from the output of an information search engine, static predictive Word2Vec models trained on the corpus, as well as a large ChatGPT language model. The results of topic modelling with the assignment of topic labels are compared with data on the distribution of reports by conference thematic blocks of in accordance with the expert markup scheme.

Keywords: corpus linguistics, conference materials, topic modelling, topic labels, rubrication

Reference for citation: Mitrofanova O. A., Golubev R. V., Gusyatskaya P. A., Makeev K. V., Pliusnina E. A., Sukhan D. D., Troshina A. V., Utkina A. A. Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 30–44. DOI: 10.17586/2541-9781-2024-8-30-44.

Reference

[1] Mitrofanova O. A., Adamova M. A., Bukreeva L. A., Zernova A. K., Litvinova A. A., Pavlikova V. S., Sologub P. S. Korpus tekstov po korpusnoj lingvistike: sostav i etapy formirovaniya // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXVII Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2024, Sankt-Peterburg, 24–26 iyunya 2024 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2024. S. 12-28. DOI: 10.17586/2541-9781-2024-8-12-28. (In Russian)

[2] Sukhan D. D., Pliusnina E. A. Metarazmetka i vizualizaciya dannyh v korpuse tekstov po korpusnoj lingvistike // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXVII Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2024, Sankt-Peterburg, 24–26 iyunya 2024 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2024. S. 44–59. DOI: 10.17586/2541-9781-2024-8-44-59. (In Russian)
[3] Vorontsov K. V. Veroyatnostnoe tematicheskoe modelirovanie: Teoriya regulyarizacii ARTM i biblioteka s otkrytym kodom BigARTM. M.: URSS, 2023. 208 s. (In Russian)
[4] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings. FRC CSC RAS. 2018. P. 369–372.
[5] Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. P. 262–272.
[6] Heinrich G. Parameter estimation for text analysis: Technical report. 2005. P. 1–32.
[7] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. 2018. T. 512. P. 1192–1204.
[8] Erofeeva A., Mitrofanova O. Avtomaticheskoe naznachenie metok tem v tematicheskih modelyah russkoyazychnyh korpusov tekstov // Structural and applied linguistics. Volume 12. St. Petersburg, 2019. P. 122–147. (In Russian)
[9] Kriukova A., Erofeeva A., Mitrofanova O., Sukharev K. Explicit Semantic Analysis as a Means for Topic Labeling // Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings. Springer, Cham. 2018. P. 167–177.
[10] Mitrofanova O., Kriukova A., Shulginov V., Shulginov V. E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science. Springer. 2021. Vol. 1357. P. 102–114.
[11] Mitrofanova O. A., Athugodage M. M., Ten L. V. Topic Label Generation in the Popular Science Corpus // Digital Geography: Proceedings of the International Conference on Internet and Modern Society (IMS 2023). Springer, 2023. (In print)
[12] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. 2020. Vol. 12469, pt. 2. P. 134–152.
[13] Kuang D., Choo J., Park H. Nonnegative matrix factorization for interactive topic modeling and document clustering // Partitional clustering algorithms. 2015. P. 215–243.
[14] Scikit-learn // Scikit-learn. URL: https://scikit-learn.org/ (access date: 09.02.2024).
[15] Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25 (2–3). P. 259–284.
[16] Chizhik A. V. Ispol'zovanie metodov tematicheskogo modelirovaniya dlya ocenki stepeni vliyaniya SMI na obshchestvennoe nastroenie // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vyp. 5. (Trudy XXIV Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2021, Sankt-Peterburg, 24–26 iyunya 2021 g. Sbornik nauchnyh statej). SPb.: Universitet ITMO, 2021. S. 70–78. (In Russian)
[17] Kirina M. A. Sravnenie tematicheskih modelej na osnove LDA, STM i NMF dlya kachestvennogo analiza russkoj hudozhestvennoj prozy maloj formy // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul'turnaya kommunikaciya. 2022. T. 20, № 2. S. 93–109. (In Russian)

[18] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. Vol. 3. P. 993–1022.
[19] Hofmann T. Probabilistic latent semantic indexing // ACM SIGIR Forum. 2017. Vol. 51 (2). P. 211–218.
[20] Gensim // Gensim. URL: https://radimrehurek.com/gensim/ (access date: 09.02.2024).
[21] Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts // WWW 2013. Proceedings of the 22nd International Conference on World Wide Web. 2013. P. 1445–1456.
[22] biterm 0.1.5 // PyPI. URL: https://pypi.org/project/biterm/ (access date: 09.02.2024).
[23] Google // Google. URL: https://www.google.ru/ (access date: 09.02.2024).
[24] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (access date: 09.02.2024).
[25] Mitrofanova O. A. Poisk i ranzhirovanie tekstov v special'nom korpuse na osnove tematicheskogo modelirovaniya // Trudy Mezhdunarodnoj konferencii «Korpusnaya lingvistika — 2023», SPb Corpora 2023, 21–23 iyunya 2023 g. SPb.: Izd-vo SPbGU, 2024. (In Russian; in print)

Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике

Д. Д. Сухан [1,2], Е. А. Плюснина [1]

[1] Санкт-Петербургский государственный университет, [2] Just AI

sukhandaniel@gmail.com, lizapl00@mail.ru

Аннотация

В статье представлены результаты проекта по представлению и визуализации метаданных для корпуса статей по корпусной лингвистике, разработанного на кафедре математической лингвистики СПбГУ. Корпус создан под руководством В. П. Захарова и включает в себя тексты докладов конференции «Корпусная лингвистика» с 2002 по 2021 гг., семинара «Компьютерная лингвистика и вычислительные онтологии» конференции IMS с 2011 по 2023 гг., а также некоторые другие материалы. В ходе работы над корпусным ресурсом был унифицирован формат разметки данных о статьях и их авторах и реализован алгоритм автоматизированного дополнения метаинформации. Осуществлены эксперименты по визуализации связей между элементами метаданных с использованием инструментов для построения графов Gephi, WebOWL, Protégé, библиотек PyGraphviz и NetworkX для языка программирования Python. Проведен анализ результатов визуализации, реализован поиск и навигация по построенным графам в формате веб-страницы.

Ключевые слова: корпусная лингвистика, материалы конференций, графовый анализ, метаразметка, визуализация, информационный поиск, онтологии, именованные сущности

Библиографическая ссылка: Сухан Д. Д., Плюснина Е. А. Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 45–60. DOI: 10.17586/2541-9781-2024-8-45-60.

1. Введение: типы метаинформации в корпусе и необходимость их систематизации

Настоящая статья посвящена вопросам реализации экстралингвистической разметки и визуализации метаданных корпуса статей по корпусной лингвистике. Корпус был собран студентами и сотрудниками кафедры математической лингвистики СПбГУ в рамках проекта, инициированного организатором конференции «Корпусная лингвистика» В. П. Захарова. Корпус включает статьи, опубликованные в сборниках конференций «Корпусная лингвистика», а также конференции IMS (семинар «Компьютерная лингвистика и вычислительные онтологии» и секции по компьютерной лингвистике прошлых лет) общим числом 643 статьи и 1027294 токена[1].

Основная цель создания корпуса состояла в систематизации изданных научных материалов с применением методов тематического моделирования, генерации ключевых выражений

[1] Материал доступен на GitHub

и аннотаций (см. статьи авторских коллективов в данном издании). Разработанный ресурс также доступен для дальнейших исследований через репозиторий на GitHub.

Любой корпус, предназначенный для многократного использования, нуждается в метаразметке. Она не только снабжает исследователя дополнительной информацией, но и открывает широкие возможности для поиска по корпусу, фильтрации его материалов и визуализации данных [8]. Таким образом, метаинформация способствует более детальному исследованию на уровне не только текстов, но и подкорпусов. Она также позволяет осуществлять дополнительную внешнюю критику источников, учитывающую хронологические, тематические и другие факторы. Кроме того, корпус может быть оснащен собственным корпусным менеджером, организованным в формате веб-интерфейса с возможностью поиска и визуализации выбранной информации.

Если рассматривать корпус как цельное собрание текстов, то названия статей и их метаданные (например, авторов) можно рассматривать как именованные сущности (обычные или вложенные), обладающие соответствующими полями. Весь корпус с этой точки зрения представляет собой многоуровневую базу данных с несколькими типами сущностей, связанных друг с другом через гиперссылки. Эти связи, в свою очередь, становятся доступными исследователю с помощью визуализации — например, с использованием методов автоматического построения графов.

Визуализация в современных корпусных менеджерах — например, в Национальном корпусе русского языка (НКРЯ) или SketchEngine — играет настолько важную роль, что учитывается при проектировании этих информационных систем [1]. Однако, речь чаще идет об отображении результатов поиска по корпусу, в то время как визуализации метаданных уделяется меньше внимания — обычно, разработчики используют функционал простых графиков, например, круговых диаграмм. На наш взгляд, визуализация экстралингвистических данных играет не меньшую роль. Многоуровневая разметка современных корпусов является фактически сетью связанных именованных сущностей разного типа, представимой в виде формальной онтологии. Данная сеть позволяет пользователю сосредоточиться на тех сущностях, которые его интересуют, и в удобном графическом формате анализировать связи между ними.

Полный корпус однородных текстов, таких как статьи конференции, можно также представить и как электронную библиотечную систему [4]. В таких системах доступно большое количество экстралингвистической информации, требующей организации, а современные онлайн-технологии позволяют организовать визуализацию. Исследования, проведенные в Лос-Аламосской национальной лаборатории в начале 2010-х гг., показали, что представление библиотечной метаинформации в формате графов, таких как RDF, отвечает запросам пользователей. Кроме того, поскольку последние заинтересованы в получении информации различных типов, оптимальным вариантом является использование для графов объединенных данных из нескольких онтологий (например, отдельных наборов полей для статей и авторов) [10]. На наш взгляд, для любого корпуса статей может быть применена аналогичная концепция, объединяющая разные иерархии понятий в единую сеть.

В настоящей статье мы представим полный цикл процесса визуализации метаданных корпуса в формате графов с элементами библиотечно-поисковой системы. Указанный цикл включает метаразметку, подбор подходящих инструментов для визуализации и итоговое представление в виде веб-страницы. Кроме того, мы покажем, какие возможности для анализа открываются с помощью графового представления метаинформации корпуса.

2. Подготовка метаинформации

Метаразметка, или добавление метаданных любых типов к текстам, является одной из обязательных стадий создания каждого корпуса. Обычно она осуществляется непосредственно после предварительной обработки и токенизации текстового материала. Метаразметка предполагает присвоение разнообразных типов данных, однако традиционно

делится на структурную, лингвистическую и экстралингвистическую. В данной статье главное внимание уделено последнему типу разметки: добавлению внешней информации о текстах. В отличие от других типов метаразметки, которые в настоящее время осуществляются с высокой степенью автоматизации, внешняя разметка требует обработки разнородной информации преимущественно вручную [2].

Поскольку основной единицей корпуса являются статьи конференций, избранная нами разметка преимущественно состоит из относящихся к ней полей. Для статей нашего корпуса такими ключевыми полями являются «автор», «год издания», «конференция» и некоторые другие. При этом следует отметить уникальность поля «автор»: несмотря на то, что авторы в рамках корпуса вторичны по отношению к статьям и выступают в роли метатегов, они сами могут быть представлены в качестве именованных сущностей, имеющих несколько полей. Среди таких полей можно перечислить «ФИО», «аффилиацию», «личные данные» (например, адрес электронной почты или личный сайт) и другие.

Подполя, относящиеся к авторам статей, также удобны и для целей визуализации, поскольку дают более полную информацию об авторах и дополнительные возможности по кластеризации элементов корпуса. Таким образом, фактически в качестве основы выбрана метасущность — пересечение «автор-статья». Можно также представить метаинформацию корпуса в виде пересечения двух таблиц реляционной базы данных, где каждая строка соответствует отдельной паре «автор-статья», что обеспечивает и удобство для реализации информационного поиска.

Наконец, при определении списка полей учитывалась их потенциальная польза для пользователя корпуса. Параметры включают как те, которые могут быть визуализированы (например, аффилиация или ФИО), так и те, которые представляют интерес только в качестве справочной информации внутри библиографической карточки. К последним относятся, например, параметры, имеющиеся не у всех статей (допустим, ссылки на оригинальный текст в формате pdf, отсутствующие у ранних статей начала 2000-х гг.), а также параметры-списки, имеющие переменную длину (например, дополнительные аффилиации или ссылки на сайты авторов).

Как отмечает В. П. Захаров, экстралингвистическая разметка, как правило, осуществляется вручную [2]. Тем не менее, в ряде случаев поиск данных был нами автоматизирован. Например, список ссылок на личные веб-страницы авторов был получен автоматически с использованием библиотеки Requests для языка программирования Python путем фильтрации результатов автоматической выдачи по предустановленному списку подходящих сайтов. Для формирования поискового запроса использовалась уже доступная информация по персоне: так, добавление инициалов увеличивало точность поиска на 50 %, а аффилиации — еще на 20 %. Полный список параметров с кратким обоснованием причин их выбора представлен в таблице.

В ходе подготовки были выделены подходящие поля для метаинформации, определена ее пригодность для различных задач и получены данные, необходимые для дальнейшей работы. Кроме того, был разработан стандарт сокращения длинных названий аффилиаций, а также исправлен ряд ошибок и неточностей в исходных данных, включая лакуны в выходных данных статей, опечатки в инициалах. Также был расширен формат представления ФИО с целью исключения смешения однофамильцев, иногда приводившее в печатных сборниках к исчезновению персоны из списка авторов. Большое число тезок среди авторов означает, что такие вложенные именованные сущности, как ФИО авторов статей, не могут размечаться автоматически, и требуют ручной проверки.

Для того, чтобы отобразить при визуализации степень участия авторов в конференции, каждому из них были добавлены дополнительные веса. За индивидуальную статью автор получал 1 балл, за статью в соавторстве вес W назначался в соответствии с формулами:

- $W = 2 / (n + 1)$ — для автора, указанного первым;
- $W = 1 / (n + 1)$ — для последующих авторов (где n — количество соавторов).

Баллы первого автора выше потому, что, как правило, первый в списке авторов является руководителем группы, отвечающим за итоговое качество статьи и ее представление в ходе выступления, что накладывает на него дополнительную ответственность.

Таблица. Список метаданных корпуса

№	Поле	Определяет	Пригодность	Способ аннотации	Причина выбора
1	ФИО автора	Автор, статья	Карточка, визуализация	Вручную	Гарантирует уникальность именованных сущностей
2	Конференция	Статья	Карточка	Автоматически	Добавлены для удобства формирования подкорпусов
3	Год издания	Статья	Карточка	Автоматически	
4	Ссылка на размещение статьи	Статья	Карточка	Вручную	Удобство навигации по корпусу
5	Аннотация статьи	Статья	Карточка	Автоматически	
6	Ключевые слова	Статья	Карточка	Вручную	
7	Тематические метки	Статья	Карточка	Автоматически	Цель создания корпуса
8	Аффилиация автора	Автор, статья	Карточка, визуализация	Вручную	Гарантирует уникальность именованных сущностей
9	Контакты автора	Автор	Карточка	Вручную	Интеграция с поисковыми системами
10	Ссылки на электронные страницы автора	Автор	Карточка	Автоматически	

Заключительным этапом стала автоматическая очистка данных от лишних пробелов, некорректных символов и нормализация длины названия статей и аффилиации для удобства отображения на графе, и приведение к единому csv-формату.

3. Визуализация экстралингвистических данных

Для удобства визуализации метаинформации корпуса можно представить данные как формальную онтологию [1]. В настоящий момент существует разнообразие методов построения онтологий. Пользователю доступны онлайн-редакторы, например WebVOWL, офлайн-программы, например Gephi, Protégé, а также библиотеки для различных языков программирования, например, PyGraphViz, NetworkX для Python. Разные программные средства обладают своими преимуществами и недостатками, а также по-разному ведут себя в зависимости от объемов обрабатываемых данных. Для определения лучшего решения для визуализации метаданных нашего корпуса мы провели ряд экспериментов.

Визуализация онтологии, как правило, производится путем построения графа. В случае с метаданными корпуса вершинами графа могут служить разнообразные элементы данных (например, названия статей или авторов), а ребра обеспечивают связь между ними (к примеру, связывая соавторов статей). В связи с объемом данных в корпусе единственным доступным вариантом является автоматическое построение графа. Внешний вид графа в этом случае определяется параметром укладки. Укладка подразумевает расположение вершин и ребер таким образом, что разным вершинам соответствуют различные точки, а кривые, соответствующие ребрам (исключая их концевые точки), не проходят через точки, соответствующие вершинам, и не пересекаются [5].

Поскольку предобработка данных проводилась средствами языка программирования Python, встроенные в него библиотеки для построения графов были рассмотрены в первую очередь. Так, библиотека NetworkX для языка Python предназначена для базовой работы с графами и другими сетевыми структурами. Библиотека способна создавать различные типы графов: простые, ориентированные и взвешенные. Преимуществом является интеграция с другими библиотеками для визуализации и анализа данных в Python, такими как Matplotlib, Pandas и др. NetworkX способен проводить манипуляции с различными видами данных, составляющими вершины графов, например, текстами, изображениями, электронными таблицами, временными рядами и т. д. [9].

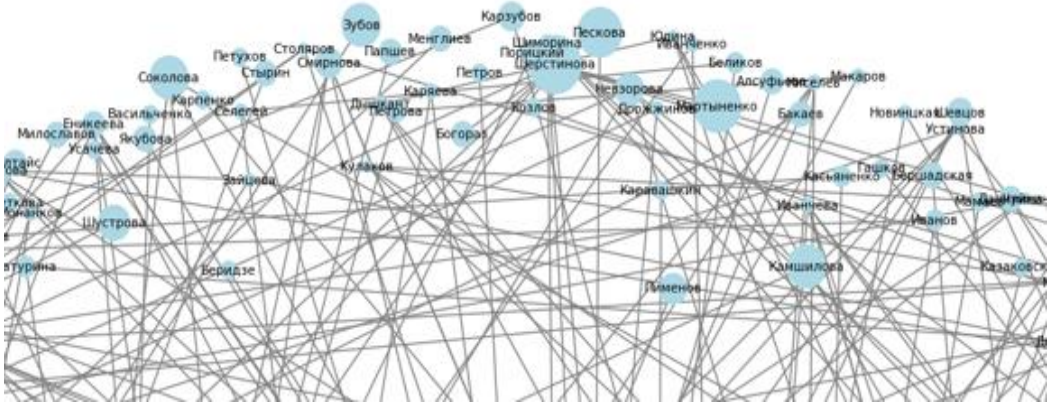

Рис. 1. Пример взаимного расположения элементов и связей вида «автор-автор» в укладке spring_layout для библиотеки NetworkX)

NetworkX обладает большим разнообразием типов построения и укладки графов. Подробная документация и простота использования позволяют легко создавать небольшие графы. Однако минусом библиотеки является низкая гибкость встроенных методов, невозможность управления отдельными элементами графа, а также малый набор инструментов, позволяющих избежать взаимного наложения вершин и ребер. Единственным способами избежать наложения являются простое расталкивание элементов путем пропорционального расширения размера графового пространства. Это приводит к тому, что вершины отображаются в виде точек на окружности, связанных большим числом случайным образом построенных линий, создавая визуально привлекательный, но непригодный для анализа рисунок (рис. 1). Все это делает NetworkX неподходящим вариантом при большом количестве данных.

Решением для больших объемов данных является использование библиотеки PyGraphViz для Python. Библиотека основана на мощных алгоритмах инструмента построения графов Graphviz, позволяя управлять параметрами напрямую через Python. Интеграция PyGraphViz с NetworkX позволяет создавать базовые NetworkX-графы, настраивая каждый элемент отдельно, а затем перевести их в формат PyGraphViz.

PyGraphViz предлагает многочисленные варианты построения графов. Для метаданных корпуса, где один автор может соответствовать нескольким статьям, а одна статья — нескольким авторам, наилучший результат показал тип графа MultiDiGraph, соответствующий графам с кратными ребрами. Для графа метаданных была использована иерархическая укладка «dot», успешно создающая ориентированные ребра в условиях максимально ограниченного пространства. Кроме того, PyGraphViz дает доступ к дополнительным настройкам укладки, включая встроенные алгоритмы предотвращения наложения и скалирования размера, такие как overlap_scaling, увеличивающий масштаб до тех пор, пока элементы не окажутся друг от друга на достаточном расстоянии. Это позволяет эффективно использовать все доступное графу пространство, размещая элементы

равномерно. Вместе с этим, PyGraphViz предоставляет возможность использования нестандартного формата соединений (например, ломаных), что гарантирует обтекание вершин графа его ребрами и, благодаря этому, размещать в узлах всю необходимую текстовую информацию (в нашем случае, названия статей, инициалы авторов и т.д.), обеспечив ее читабельность (см. пример на рис. 2).

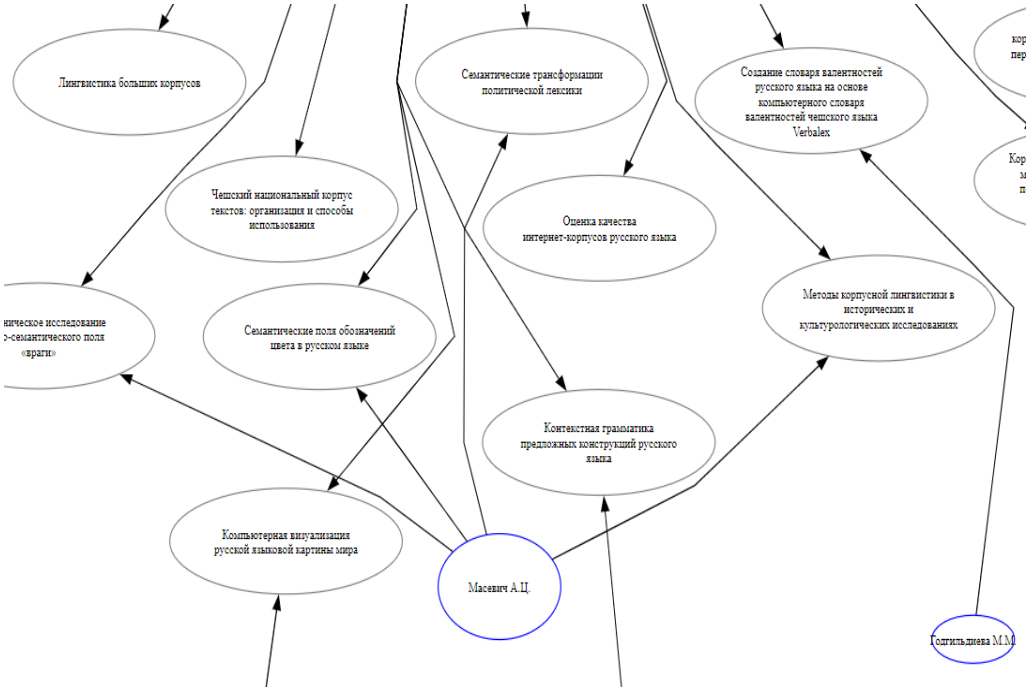

Рис. 2. Пример укладки графа библиотекой PyGraphViz

Нами также было рассмотрено ПО на основе языка для определения и создания веб-онтологий — Web Ontology Language (OWL) [3]. Код на этом языке состоит из двух частей: заголовок, в который включаются версия, примечания и импортируемые онтологии, а также тело, в котором описываются классы, свойства, аксиомы. Веб-инструмент на основе языка OWL (WebVOWL) позволяет самостоятельно создать онтологии или загрузить уже готовый JSON-файл. Экспортировать онтологию можно в форматах JSON, SVG, Tex или TTL. WebVOWL позволяет пользователю создавать онтологии для обмена информацией в электронной коммерции (GoodRelations), онтологии связей между людьми (PersonasOnto) и другие. Для визуализации именованных сущностей наилучшим вариантом является PersonasOnto.

Несмотря на обширный функционал, набор параметров настройки визуализации невелик. В редакторе отсутствует возможность выбора размера вершин и ребер, недоступны автоматическая кластеризация и укладка графа. В силу слабой интерпретируемости формата представления данных WebVOWL значительно уступает другим инструментам в контексте задач настоящего исследования.

Protégé — поддерживаемая на языке программирования Java среда для создания онтологий различных предметных областей [10]. Пользователю по предварительной регистрации доступна как онлайн-, так и офлайн-версия данного редактора. Protégé поддерживает различные форматы: RDF/XML, Turtle, OWL/XML, OBO и другие. Импорт доступен только в формате OWL «автор-автор», однако, есть опция создания иерархических

классов посредством txt-документа. Данные в виде связей были загружены и визуализированы (рис. 3).

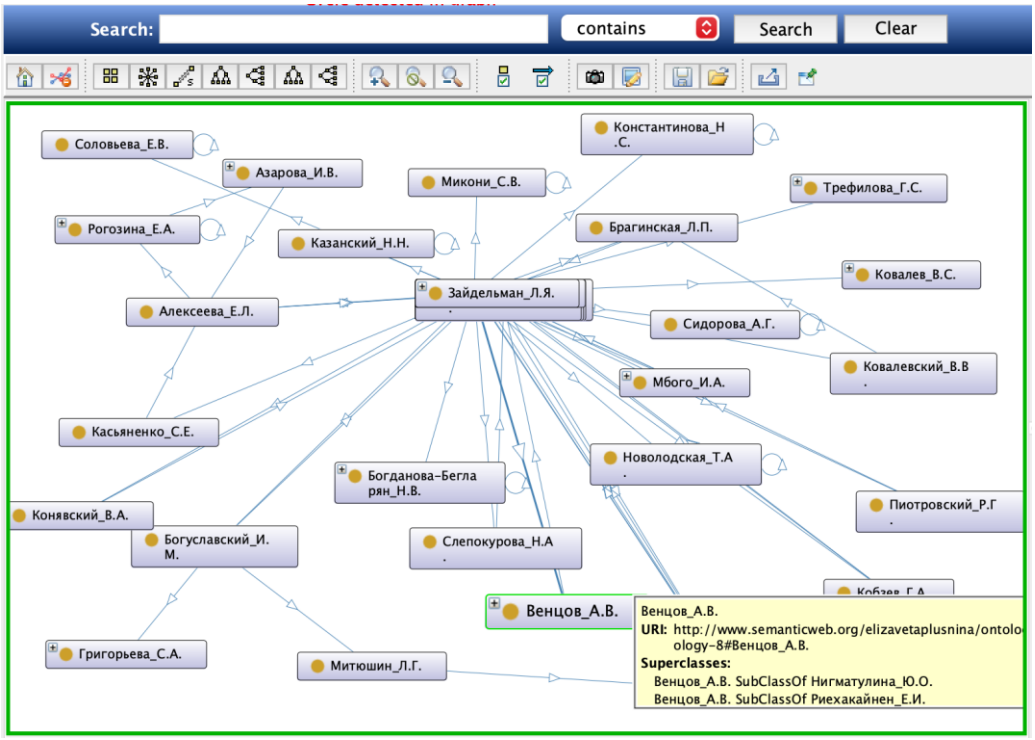

Рис. 3. Пример визуализации связи «автор-автор» в программе Protégé

Программа поддерживает построение ориентированных графов и предлагает возможность интерактивного взаимодействия. При клике на необходимую вершину пользователю предоставляется информация об элементе графа: URI, подклассах данного элемента, а также другая информация, которая вводится пользователем собственноручно. Однако, в Protégé отсутствует автоматическая укладка графа, что значительно затрудняет кластеризацию элементов. Наконец, ресурс не позволяет изменять цвет, размер или вид вершин и ребер графа. Отсутствие данных, важных для визуализации настроек, сподвигли нас сместить фокус внимания с языка OWL на программное обеспечение на базе инструмента построения графов GraphViz.

Недостатки, связанные с настройкой параметров, компенсируются в редакторе Gephi. С помощью Gephi, пользователь способен манипулировать структурами, формами и цветами, чтобы выявить скрытые закономерности [7]. В отличие от описанных выше редакторов, Gephi не поддерживает форматы RDF, JSON или OWL. Несмотря на это, пользователю предоставляется большая альтернатива: GraphViz DOT, CSV и другие. Однако возможность экспорта графа реализована только в трех форматах: PDF, SVG или PNG.

Для загрузки данных в редактор необходимо предварительно разделить их на две таблицы: таблица вершин и таблица ребер; после загрузки данных, программа автоматически создаст граф. Таблицу можно просматривать и редактировать в разделе «Лаборатория данных». В разделе обработки пользователю предоставляется граф и окна с редактированием параметров. В окне «Appearance» можно настроить цвет, размер и шрифт вершин и ребер. Помимо встроенных параметров отображения, доступна возможность добавления собственных, зависящих от значений в таблице данных. Например, назначив всем авторам метку «1», а всем статьям метку «0», получим два

кластера данных, каждому из которых можно сопоставить отдельные настройки внешнего вида.

В Gephi представлено разнообразие укладок графа, начиная от случайных, заканчивая сложными математическими алгоритмами. Один из таких алгоритмов — «ForceAtlas» и его модифицированная версия «ForceAtlas2». Последняя направлена на работу с большими объемами данных и содержит модели сил притяжения, гравитации и отталкивания по закону Гука [9], а также учитывает вес ребер при укладке. Иными словами, вершины итерационно притягиваются или отталкиваются друг от друга в пространстве визуализации в зависимости от их взаимного расположения и наличия связей. Пользователь имеет возможность самостоятельно определять момент остановки укладки, поскольку она происходит в реальном времени. На выходе данного алгоритма можно увидеть максимально наглядную раскладку графа, так как при проектировании метода был сделан акцент на качестве визуализации [9]. После «ForceAtlas2» мы применили укладку «Noverlap», которая исключает наложения вершин друг на друга. Наконец, с помощью алгоритмов «Расширение» и «Сокращение» пользователь может менять расстояние между вершинами, не теряя при этом кластеризацию.

Помимо описанных выше преимуществ Gephi, можно отметить также возможность создавать изогнутые ребра. При визуализации большого объема данных такие ребра делают граф визуально более привлекательным. Для анализа полученного графа пользователь может воспользоваться встроенных статистическими методами. Благодаря этому можно не просто посчитать количество ребер и вершин, но также и силу связи между ними. Как было описано выше, импортировать полученный граф можно в разных форматах, а также можно настроить размер страницы, отступы, ориентацию и фон. В виду такого большого объема различных настроек скорость работы сервиса снижается пропорционально увеличению числа данных.

Таким образом, эксперименты показывают преимущества библиотеки PyGraphViz и редактора Gephi для визуализации метаданных корпуса. Редактор Gephi имеет удобный пользовательский интерфейс и содержит все необходимые для нашего исследования инструменты. В то же время, библиотека PyGraphViz удобна при работе с языком Python, а набор доступных настраиваемых параметров ограничен количеством интегрируемых библиотек.

Все графы представлены в двух форматах: SVG и PDF. Формат SVG сохраняет внутреннюю структуру и расположение графа на плоскости и в дальнейшем может быть использован для создания интерактивной HTML-страницы. PDF-формат, в отличие от представленного также PNG, удобен при детальном рассмотрении графа в силу сохранения качества при приближении документа.

Среди возникших трудностей в процессе преобразования файла в доступный читателю формат следует отметить следующие. Во-первых, объем данных — текстовые данные, которые занимают много пространства в графе и накладываются друг на друга. Автоматическая укладка не всегда решает проблему, а ручная обработка затратна по времени. Во-вторых, возникла необходимость дополнительного назначения весов для участников конференции с целью более информативной визуализации о количестве их участия в конференциях. В-третьих, остается нерешенной проблема имен. В нее включается проблема полных тезок, а также людей, сменивших свою фамилию за данный период. Отдельную трудность представляют иностранные имена.

4. Результаты экспериментов по визуализации и примеры анализа

В результате визуализации с помощью редактора Gephi было получено три ориентированных графа на основе трех различных связей: «автор-статья», «автор-автор» и «аффилиация-автор». В каждом из них, в соответствии с выбранными метаданными, были выполнены автоматическая кластеризация, укладка графа и настройка параметров.

В настройку параметров были включены: изменение размера вершин, изменение цвета вершин и ребер, изменение размера шрифта, ограничение текста по количеству символов.

Для кластеризации связи «автор-статья» вершинами являются авторы и их статьи. Вершины с названиями статей имеют одинаковый размер, в отличие от авторов, где размер вершины напрямую зависит от веса. Метками ребер для данного графа послужили аффилиации (рис. 4).

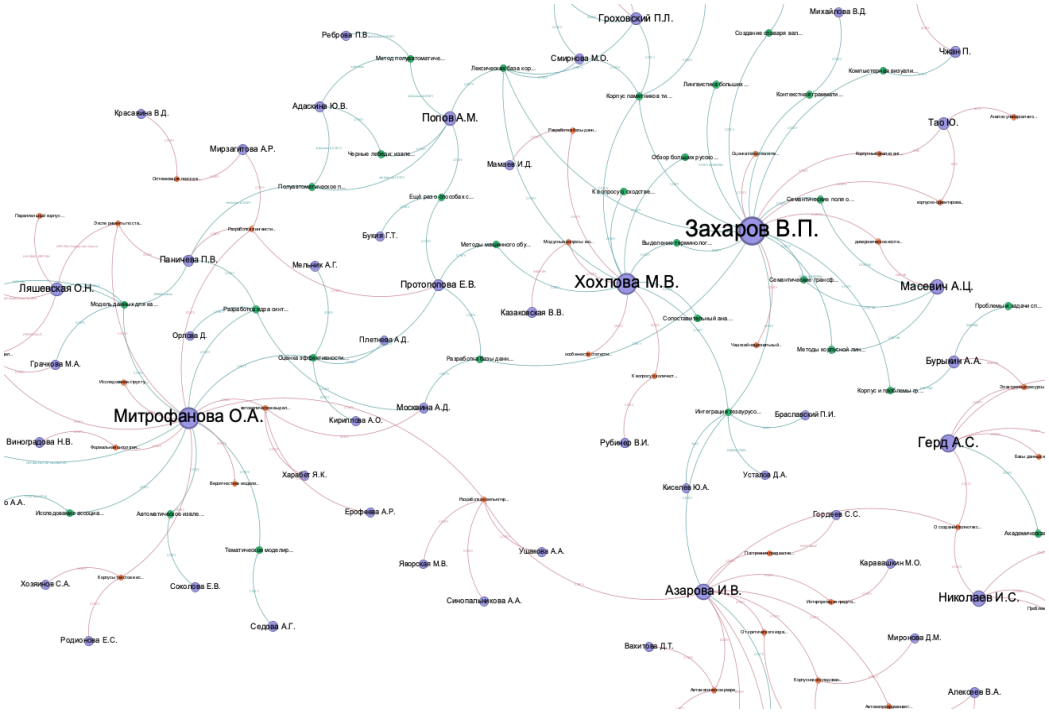

Рис. 4. Пример визуализации связи «автор-статья»

Граф «автор-статья» центрирован на таких авторах, как В. П. Захаров, М. В. Хохлова, О. А. Митрофанова. Согласно статистическим данным корпуса, из 531 автора эта тройка обладает самой высокой публикационной активностью. Так, В. П. Захаров опубликовал 20 статей, О. А. Митрофанова — 16 и М. В. Хохлова — 14. Расположение В. П. Захарова в центре самого крупного кластера логически соотносится с его статусом как организатора конференции Corpora и члена организационного комитета семинара «Компьютерная лингвистика и вычислительные онтологии». Авторы О. А. Митрофанова и М. В. Хохлова имеют одинаковый размер вершины, однако разное количество статей. Это объясняется тем, что М. В. Хохлова имеет больше индивидуальных публикаций, в то время как О. А. Митрофанова опубликовала большое количество статей с соавторами, на что также указывает и больший размер кластера вокруг неё.

Построенный граф также позволяет изучить различные кластеры авторов и совместные исследования представителей различных научных групп (рис. 5).

На рисунке можно отметить два изолированных кластера ученых, наибольшее количество авторов одной статьи по корпусу достигло восьми человек. В среднем, количество авторов статьи составляет два человека. На графе были выделены кластеры устойчивых соавторов, например К. К. Боярского и Е. А. Каневского, а также другие группы, которые совместно публиковали свои статьи в конференции IMS и Corpora (рис. 6).

Для кластеризации связи «автор-статья» вершинами являются авторы. В данном типе графа была приведена настройка размера вершин, в соответствии с назначенными весами. В качестве меток ребер выступают названия статей (рис. 7).

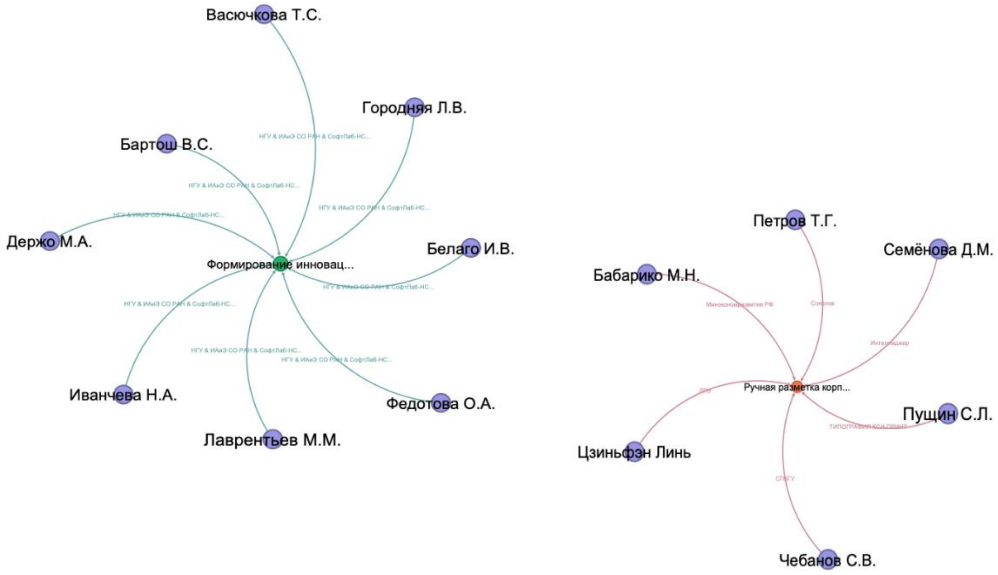

Рис. 5. Пример кластеризации авторов

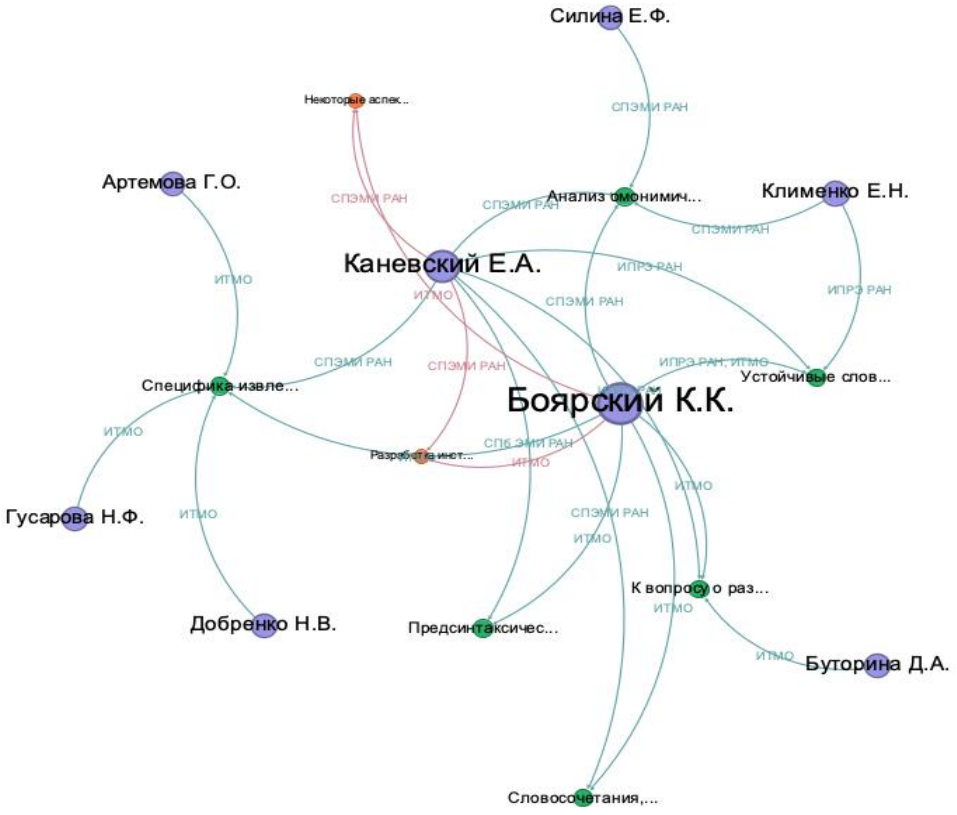

Рис. 6. Пример кластера устойчивых соавторов

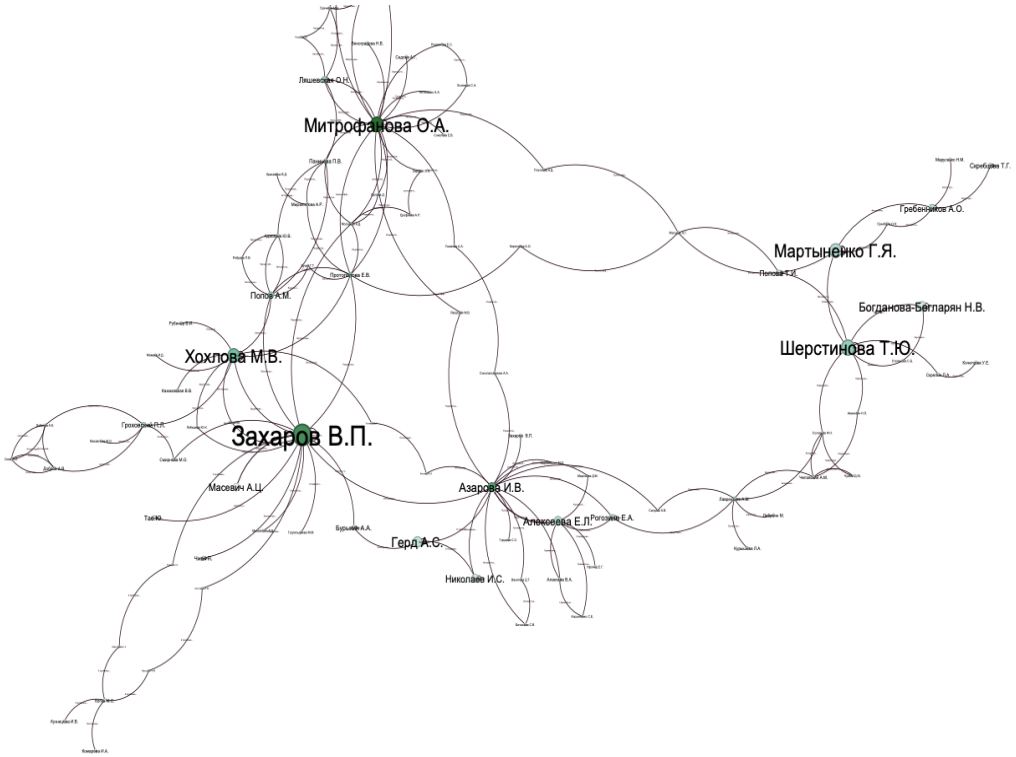

Рис. 7. Пример визуализации связи «автор-автор»

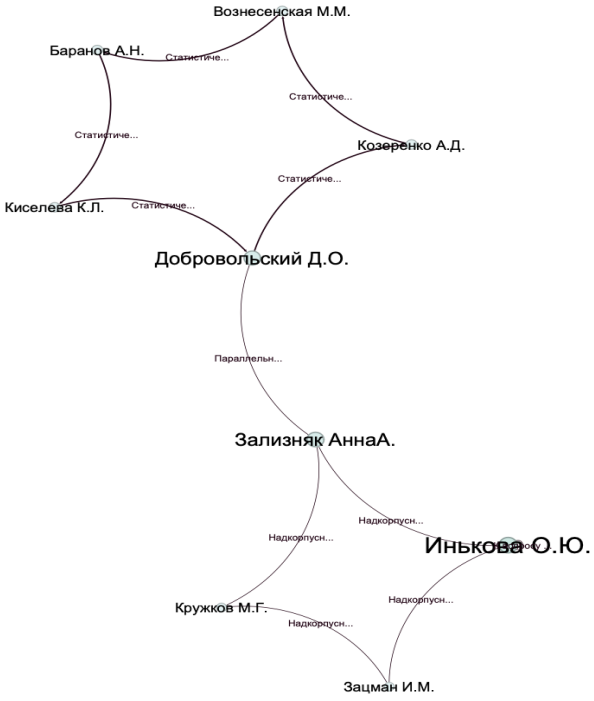

Рис. 8. Пример графа с петлей

Граф связей «автор-автор» повторяет эту структуру: фокус внимания вновь на авторах статей. В отличие от графа «автор-статья», в данной визуализации допустимо наличие графов с петлями, если статья была написана одним ученым без соавторов, но такая визуализация трудночитаема, так как название ребра и названия вершины накладываются друг для друга (рис. 8).

Данная визуализация помогает проследить, как могут быть связаны между собой представители различных научных групп.

Заключительным экспериментом стала визуализация связи «аффилиация-автор». Чтобы вершины можно было отличить от вершин авторов, они были увеличены в размере для наглядности (рис. 9).

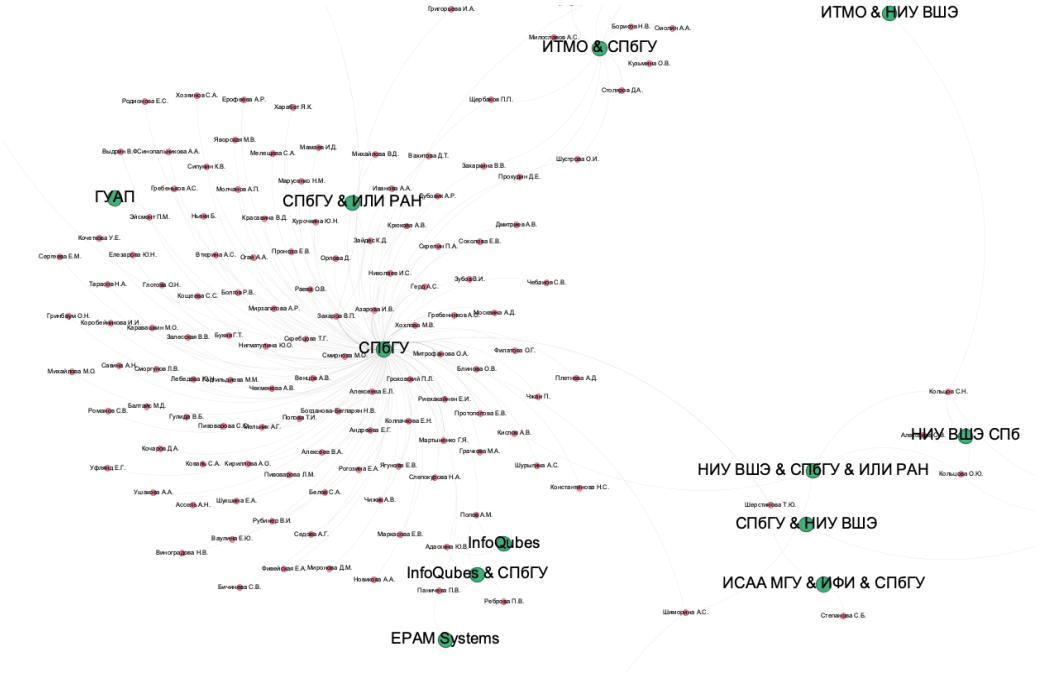

Рис. 9. Пример визуализации связи «аффилиация-автор»

Граф аффилиации центрирован на ведущих вузах Санкт-Петербурга: СПбГУ, ИТМО и ВШЭ. С помощью данного графа можно определить не только ученых, относящихся к той или иной организации, но также и взаимодействие организаций между собой. Некоторые исследователи публиковали статьи с разными аффилиациями, в зависимости от требований конкретной конференции или в связи со сменой места работы. Согласно статистике, из СПбГУ были опубликованы 274 автора, что превышает половину участников. Авторов из ИТМО оказалось 36, на третьем месте Институт русского языка РАН с 31 публикацией и на одну меньше набрала ВШЭ. Однако, почти столько же статей исследователи из ВШЭ опубликовали совместно с авторами из других организаций. Более 80 авторов имели две и более аффилиаций, вплоть до четырех.

На наш взгляд, графы, фиксирующие соотношение элементов метаразметки статей корпуса, могут функционировать не только как способы визуализация, но и как разновидность формальной онтологии. Кроме того, было запланировано размещение корпуса на веб-странице проекта. По этим причинам все графы, созданные при помощи Gephi и PyGraphViz, были выгружены в формате SVG — подвиде XML для графовых файлов, сохраняющем информацию о типе и взаимном расположении элементов. Каждая

вершина графа была снабжена собственной невидимой HTML-ссылкой, уникальной для элемента.

После этого граф был включен в состав веб-страницы, написанной на HTML и CSS. Пользователю доступны: выбор графа для отображения, навигация по графу, поиск любого элемента по названию. Кроме того, при клике мышью на элемент (статью или автора) средствами JavaScript подгружается библиотечная карточка, содержащая дополнительную метаинформацию из числа той, которую невозможно разместить на основном графе. Для статей карточка содержит выходные данные, аннотацию, информацию об авторах, тематические метки статей и т. д. Карточки авторов демонстрируют список статей и соавторов исследователя, его аффилиацию и предоставляют доступ к ссылкам на внешние интернет-страницы с публикациями и информацией об авторе. При клике на элементы карточки, соответствующие узлам графа, пользователь переносится к соответствующему узлу (рис. 10).

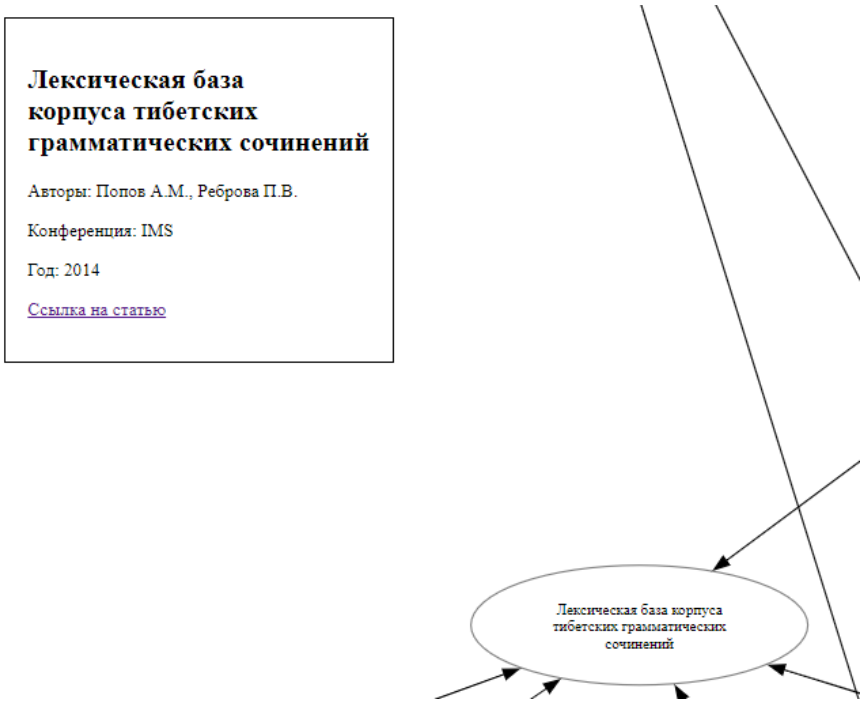

Рис. 10. Пример библиографической карточки статьи на веб-странице метаданных корпуса

Веб-страницы подобного формата представляют собой вид библиотечной онтологии, в которой для навигации используется визуальная составляющая в виде графа, а также ссылочные элементы. Визуализация, на наш взгляд, играет существенную роль в поиске данных, позволяя, например, находить похожие на целевую статьи без дополнительных сложных SQL-запросов, давая возможность избегать громоздких форм. Кроме того, данные об авторах доступны в сжатом виде, что избавляет читателя от необходимости поиска персоналии в интернете. Автоматическое обновление и графа, и HTML-страницы обеспечивают бесперебойное функционирование системы.

5. Заключение

Экстралингвистические метаданные корпуса играют важную роль в любом корпусном менеджере. Они обеспечивают пользователя дополнительной информацией, позволяют

обрабатывать большие данные, а также структурируют корпус и обеспечивают возможность информационного поиска. Работа с метаданными любого корпуса, вне зависимости от его размера, представляет собой трудоемкую задачу, состоящую из нескольких самостоятельных этапов.

В рамках настоящего исследования нами был применен комплексный подход к сбору и применению метаданных для нового корпуса статей по корпусной лингвистике. Основной акцент в исследовании был сделан на пользе метаданных для визуализации экстралингвистической информации о корпусе. Был проведен сравнительный анализ ряда программных средств визуализации формальных онтологий в виде графов. Программы для визуализации формальных онтологий и отношений между сущностями показали свою эффективность для визуализации внешней разметки корпуса, однако конкретный результат значительно зависит от размера базы данных. Несмотря на то, что большинство алгоритмов построения графов становятся менее эффективными с ростом объема данных, некоторые из них способны строить системы отношений между многочисленными сущностями. Среди них выделяются схемы, построенные на алгоритме силового отталкивания, поскольку они одновременно обеспечивают равномерное заполнение пространства и препятствуют наложению элементов друг на друга. Наилучший результат с точки зрения внешнего вида показала программа Gephi. Экстраполируя применение этих методов на еще большие данные, доступные, например, в крупных корпусах русского языка, можно говорить об их практической эффективности.

Визуализация отношений между сущностями экстралингвистической разметки корпуса обеспечивает пользователя еще одним способом навигации по нему. Кроме того, в доступном формате демонстрируются базовые кластеры элементов метаданных корпуса, что дает дополнительную статистическую информацию для любого исследования. Единственным недостатком применения графового метода является большой размер итогового изображения. В полном виде и на максимальном отдалении она становится нечитабельной. В настоящей работе эта проблема решается за счет двух факторов. Во-первых, графовые файлы могут масштабироваться, что позволяет сосредоточиться на отдельном участке данных. Во-вторых, чтобы облегчить переход к нужному участку, граф содержит интерактивные возможности поиска, позволяющие легко найти нужную статью или автора и моментально перейти к ним. Кроме того, важным преимуществом графовых библиотек (прежде всего, PyGraphViz) является возможность автоматизации построения визуализации. Данные для построения схемы подгружаются из базы автоматически и могут быть отфильтрованы образом, удобным пользователю — например, по определенным полям, по части данных или пользовательскому подкорпусу.

Практическим результатом исследования представлены в виде веб-страницы, составленной с помощью средств HTML, CSS и скриптов на Python и JavaScript, обеспечивающая онлайн-доступ к результатам визуализации, интегрированная с репозиторием хранения корпуса. При этом метаинформация, признанная неподходящей для использования внутри графов, доступна в интерактивном режиме в виде библиографических карточек для каждой статьи и справочных материалов для каждого автора непосредственно на веб-странице, где размещен граф. Это позволяет пользователям получать доступ к максимальном объему экстралингвистической информации онлайн и в удобном формате. В перспективе планируется размещение материалов в открытом доступе на сайте СПбГУ.

Исследование показало применимость графового метода визуализации корпусной метаинформации в корпусах небольшого размера. В практической плоскости результаты графовой визуализации показали потенциал для встраивания в более крупные структуры — такие, как корпусные менеджеры.

Однако, следует отметить, что научная работа в этом направлении не завершена и может быть продолжена. Среди потенциальных направлений можно рассмотреть увеличения объема данных как за счет добавления новых метаданных, так и за счет работы с корпусами

большего размера, улучшение укладки графов, а также применение графового анализа для работы не только с внешней, но и с лингвистической разметкой материалов корпуса. Учитывая растущую роль интерактивных интернет-ресурсов в жизни ученых, мы вправе ожидать новых исследований в этом направлении.

Литература

[1] Гладилин С. А. и др. Прототип корпусной платформы нового поколения для НКРЯ // Сборник 28-й международной конференции по компьютерной лингвистике и интеллектуальным технологиям (15–18 июня 2022 г., Москва). М., 2022. С. 1043–1054.

[2] Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб., 2020. 234 с.

[3] Лебедев С. В., Нгуен Н. Т., Баймуратов И. Р., Жукова Н. А. Анализ средств визуализации OWL-онтологий // Труды 5-ой Международной научной конференции «Технологическая перспектива в рамках евразийского пространства: новые рынки и точки экономического роста». Санкт-Петербург, 7–8 ноября 2019 г. СПб.: Центр научно-производственных технологий «Астерион», 2019. С. 273–274.

[4] Славута Т. А. Национальный корпус русского языка как информационно-библиографический ресурс // Материалы Всероссийской научно-практической конференции «Динамика библиотечно-информационного обеспечения образования, науки и культуры». Омск: Омский государственный технический университет, 2020. С. 175–181.

[5] Укладка графа // Большая российская энциклопедия. М.: Большая российская энциклопедия, 2017. URL: https://bigenc.ru/c/ukladka-grafa-09f2e3 (дата обращения: 14.04.2024).

[6] Хоай Л., Тузовский А. Ф. Использование онтологии в электронных библиотеках // Известия Томского политехнического университета. 2012. Т. 320, № 5. С. 36–41.

[7] Bastian M., Heymann S., Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks // Proceedings of the International AAAI Conference on Web and Social Media, 2009. 2009. Vol. 3 (1). P. 361–362.

[8] Dash N. Language Corpora Annotation and Processing. Singapore: Springer, 2021. P. 71–90.

[9] Jacomy M., Venturini T., Heymann S., Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software // PLOS One. 2014. Vol. 9 (6). URL: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0098679&type=printable (дата обращения: 14.04.2024).

[10] Powell J., Collins L., Martinez M. Semantically Enhancing Collections of Library and Non-Library Content // D-Lib Magazine. 2010. Vol. 16 (7/8). URL: https://www.dlib.org/dlib/july10/powell/07powell.html (дата обращения: 14.04.2024).

[11] Sivakumar R., Arivoli P. Ontology Visualization Protégé Tools ― a Review // International Journal of Advanced Information Technology (IJAIT). 2011. Vol. 1 (4). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3429010 (дата обращения: 14.04.2024).

Meta Tagging and Visualization for the Corpora Linguistics Texts Corpora

D. D. Sukhan [1,2], E. A. Pliusnina [1]

[1] Saint–Petersburg State University, [2] Just AI

This paper presents the project representation and visualization results of metadata for a corpora linguistics corpus of articles. The corpus, which was built at the Saint Petersburg State University Computer and Applied Linguistics department under V. P. Zakharov supervision, included papers

texts from the conference «Corpus Linguistics» which were published from 2002 to 2021, the IMS conference workshop «Computational Linguistics and Computational Ontologies» paper texts from 2011 to 2023, as well as some other materials. The author and article data markup format were unified, and an automated algorithm for adding information of metadata has been implemented. Experiments were carried out to visualize connections between metadata elements using graph tools such as Gephi, WebOWL, Protégé, as well as Python programming language libraries PyGraphviz and NetworkX. The visualisation results were analysed, moreover the created graphs search and navigation in the web page format was implemented.

Keywords: corpora linguistics, conferences materials, graph analysis, metatagging, visualization, informational search, ontologies, named entities

Reference for citation: Sukhan D. D., Pliusnina E. A. Meta Tagging and Visualization for the Corpora Linguistics Texts Corpora // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 45–60. DOI: 10.17586/2541-9781-2024-8-45-60.

Reference

[1] Gladilin S. A. i dr. Prototip korpusnoi platformy novogo pokoleniia dlia NKRIA // Sbornik 28-i mezhdunarodnoi konferentsii po komp'iuternoi lingvistike i intellektual'nym tekhnologiiam (15-18 iiunia 2022 g., Moskva). M., 2022. S. 1043–1054. (In Russian)
[2] Zakharov V. P., Bogdanova S. Yu. Korpusnaia lingvistika. SPb., 2013. 234 s. (in Russian)
[3] Lebedev S. V., Nguen N. T., Bajmuratov I. R., ZHukova N. A. Analiz sredstv vizualizacii OWL-ontologij // Trudy 5-oj Mezhdunarodnoj nauchnoj konferencii «Tekhnologicheskaya perspektiva v ramkah evrazijskogo prostranstva: novye rynki i tochki ekonomicheskogo rosta». Sankt-Peterburg, 7–8 noyabrya 2019 g. SPb.: Centr nauchno-proizvodstvennyh tekhnologij «Asterion», 2019. S. 273–274. (In Russian)
[4] Slavuta T. A. Natsional'nyi korpus russkogo iazyka kak informatsionno-bibliograficheskii resurs // Materialy Vserossiiskoi nauchno-prakticheskoi konferentsii «Dinamika bibliotechno-informatsionnogo obespecheniia obrazovaniia, nauki i kul'tury». Omsk: Omskii gosudarstvennyi tekhnicheskii universitet, 2020. S. 175–181. (In Russian)
[5] Ukladka grafa // Bol'shaia rossiiskaia entsiklopediia. Moskva: Bol'shaia rossiiskaia entsiklopediia, 2017. URL: https://bigenc.ru/c/ukladka-grafa-09f2e3 (access date: 14.04.2024). (In Russian)
[6] Khoai L., Tuzovskii A. F. Ispol'zovanie ontologii v elektronnykh bibliotekakh // Izvestiia Tomskogo politekhnicheskogo universiteta. 2012. T. 320, № 5. S. 36–41. (In Russian)
[7] Bastian M., Heymann S., Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks // Proceedings of the International AAAI Conference on Web and Social Media. 2009. Vol. 3 (1). P. 361–362.
[8] Dash N. Language Corpora Annotation and Processing. Singapore: Springer, 2021. P. 71–90.
[9] Jacomy M., Venturini T., Heymann S., Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software // PLOS One. 2014. Vol. 9 (6). URL: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0098679&type=printable (access date: 14.04.2024).
[10] Powell J., Collins L., Martinez M. Semantically Enhancing Collections of Library and Non-Library Content // D-Lib Magazine. 2010. Vol. 16 (7/8). URL: https://www.dlib.org/dlib/july10/powell/07powell.html (access date: 14.04.2024).
[11] Sivakumar R., Arivoli P. Ontology Visualization Protégé Tools — a Review // International Journal of Advanced Information Technology (IJAIT). 2011. Vol. 1 (4). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3429010 (access date: 14.04.2024).

Частотные характеристики предлогов и их значений в базе данных предложных конструкций

В. В. Выборная, А. М. Гончарова, А. А. Родина

Санкт-Петербургский государственный университет

vvybornaa@gmail.com, sssparzha@gmail.com, rodinany@gmail.com

Аннотация

В статье описываются частотные характеристики соотношения предлогов и их значений, а также исследуется синтаксическая неоднозначность предложных конструкций в русском языке. Материалом исследования послужила база данных предложных конструкций, созданная в ходе проекта «Квантитативная грамматика русских предложных конструкций», который разрабатывался на кафедре математической лингвистики Санкт-Петербургского государственного университета, а также корпус из 200 синтаксически неоднозначных предложений, заимствованных из диссертационного исследования Д. А. Черновой «Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование». Анализ данных проведен с помощью инструментов модуля Pandas и других библиотек Python. Мы исходим из положения о том, что предлоги, в особенности первообразные, в разных контекстах реализуют разные значения. Мы рассматриваем соотношение предлогов и приписываемых им семантических меток в целях выявления закономерностей распределения предлогов по синтаксемам, а также определяем наиболее частотные синтаксемы и предлоги среди синтаксически неоднозначных предложений на основе многослойного перцептрона. Результаты данного исследования могут быть полезны при решении задач по снятию омонимии и вносят вклад в понимание структуры синтаксически неоднозначных предложений в русском языке, указывая на доминирующую роль синтаксемы «тематив» в их структуре.

Ключевые слова: русские предлоги, предложные конструкции, значение предлогов, синтаксемы, синтаксическая неоднозначность

Библиографическая ссылка: Выборная В. В., Гончарова А. М., Родина А. А., Частотные характеристики предлогов и их значений в базе данных предложных конструкций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 61–69. DOI: 10.17586/2541-9781-2024-8-61-69.

1. Введение

Данная статья основана на результатах, полученных в ходе выполнения проекта РФФИ «Квантитативная грамматика русских предложных конструкций» [1, с. 17]. Данный проект разрабатывался на кафедре математической лингвистики Санкт-Петербургского государственного университета. Основной целью проекта стала разработка комплексного квантитативного лексико-грамматического описания русских предлогов и предложных конструкций. В результате работы над проектом была сформирована база данных предложных конструкций, насчитывающая 11122 контекста. Представленные в базе данных контексты послужили материалом для настоящего исследования. Контексты размечены

сразу по нескольким критериям. Для каждой предложной конструкции указывается предлог, управляющее слово, его лемма и часть речи, указывается также зависимое слово, его лемма, часть речи, падеж и число. Более того, каждой предложной конструкции приписывается семантическая метка, определяющее значение, реализующееся в данном контексте [1, с. 17].

Значения, приписываемые предложным конструкциям, представлены как семантические классы, выделенные на основе синтаксем «Синтаксического словаря» Г. А. Золотовой [2]. Такое решение мотивировано тем, что значение предложной конструкции невозможно разложить на значение предлога и значение падежной формы. Предложная конструкция рассматривается как единое целое, т. е. как синтаксема, наделенная определенным значением. При этом одна синтаксема может быть представлена сразу несколькими парами «предлог — падежная форма» [1, с. 20].

Сложность работы с предлогами заключается в том, что предлоги представляют собой весьма неоднородную группу с неясными и неструктурированными значениями. В результате предлоги зачастую остаются без внимания. Например, при автоматическом анализе текста предлоги, как правило, помечаются как «стоп-слова». Однако нельзя забывать о том, что предлоги «передают четкие семантико-синтаксические отношения между знаменательными словами» [3, с. 9]. При семантически ориентированном анализе семантико-синтаксические отношения, выражаемые предлогами, безусловно, оказываются важным аспектом. Более того, ряд исследований показывает, что существует определенная закономерность в распределении служебных единиц, в том числе и предлогов, в разных типах и стилях текстов [4; 5]. Еще одна область, в которой значение предлогов оказывается ключевым, – компьютерное зрение. Интеграция задач обработки естественного языка и компьютерного зрения позволяет найти новые перспективные подходы к описанию и поиску объектов в визуальном пространстве. Предлоги при этом рассматриваются как указатели на пространственные отношения.

В рамках настоящего исследования мы сосредоточились на выявлении закономерностей распределения предлогов по синтаксемам. Это позволит нам получить более точное представление обо всём объёме значений, которые реализуются отдельным предлогом в составе различных предложных конструкций. При этом результаты нашего исследования могут быть полезны при решении целого ряда задач: задач по снятию омонимии, задач атрибуции текстов, задач по определению стилей и типов текстов, задач визуального описания и т. д.

2. Описание частотных характеристик предлогов и их значений

На основании контекстов, представленных в базе данных, можно выделить 5 наиболее частотных синтаксем: локатив (2053 контекстов), темпоратив (1463 контекста), тематив (1334 контекста), объект (943 контекста) и директив (890 контекстов). При этом каждая синтаксема представлена некоторым набором предлогов, которые в определенных контекстах реализуют значение, соответствующее семантической метке. Так, синтаксема локатив чаще всего реализуется при помощи следующих предлогов: в, на, по, за, у. Для синтаксемы темпоратив наиболее частотными являются предлоги в, на, до, за, после. Синтаксема директив чаще всего реализуется при помощи предлогов в, из, на, с, к.

Нетрудно заметить, что, по большому счету, наборы предлогов для каждой синтаксемы несильно отличаются друг от друга. Например, предлог «в» входит в пятерку самых частотных предлогов каждой группы. Следовательно, большую ценность представляет то, в каких именно контекстах тот или иной предлог реализует определенное значение и как на основе этого мы можем описать семантику того или иного предлога.

2.1. Первообразные предлоги и их значения

2.1.1. Предлог «в»

Характерной особенностью первообразных предлогов является их полисемичность. Способность первообразных предлогов реализовывать разные значения в разных контекстах определяет их частотность. Самым частотным предлогом оказывается предлог «в», который охватывает 13 синтаксем, и в базе данных предложных конструкций представлен сразу 3 146 контекстами (рис. 1).

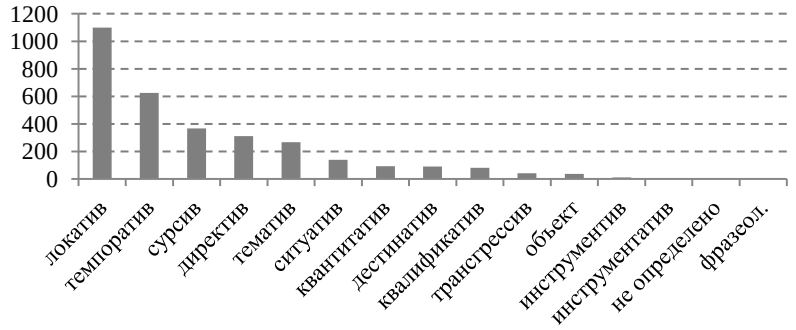

Рис. 1. Столбчатая диаграмма распределения синтаксем предлога «в»

Значение локатива («*происходит в городе*») реализуется почти в два раза чаще, чем значение темпоратива («*в ближайшее время*»). Примечательно то, что следующей по частотности синтаксемой для данного предлога оказывается сурсив, синтаксема со значением «источник информации» («*говорится в письме*»).

2.1.2. Предлог «на»

Второй по частотности предлог, предлог «на», представлен вдвое меньшим числом контекстов по сравнению с предлогом «в». Несмотря на значительную разницу в частотности, предлог «на» охватывает 11 синтаксем (рис. 2).

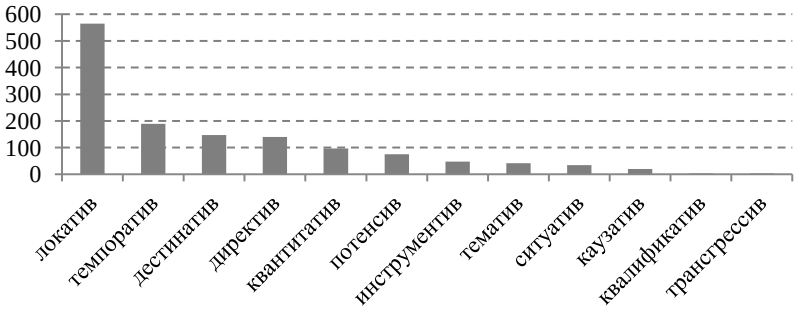

Рис. 2. Столбчатая диаграмма распределения синтаксем предлога «на»

Чаще всего контексты с предлогом «на» реализуют синтаксему локатив, т. е. значение местонахождения («*работал на киностудии*»). Компонент, выражающий временные характеристики, представлен среди прочего следующими контекстами: «*на прошлой неделе*», «*выступит на церемонии*». Чуть меньше 300 контекстов приходится в совокупности на синтаксемы дестинатив и директив, выражающие значения назначения предмета («*цена на газ*») и направления действия или движения («*подняться на второй этаж*») соответственно. В 96 конструкциях реализуется компонент, содержащий

количественные характеристики, синтаксема квантитатив («*сократился на семь процентов*»).

2.1.3. Предлог «о»

Предлог «о» представлен лишь 856 контекстами и охватывает всего две синтаксемы (рис. 3).

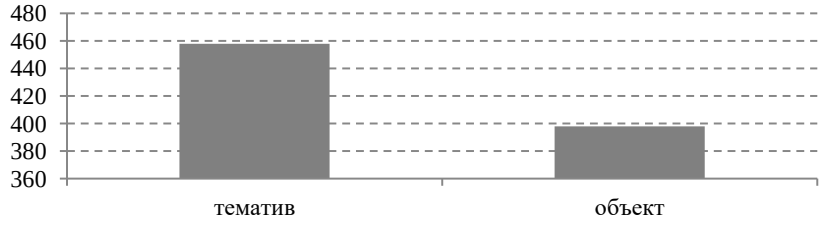

Рис. 3. Столбчатая диаграмма распределения синтаксем предлога «о»

Среди самых частотных первообразных предлогов предлог «о» оказывается семантически наиболее четко очерченным. При этом две синтаксемы: тематив и объект, делят практически поровну весь объём контекстов с данным предлогом. В «Синтаксическом словаре» Г. А. Золотовой тематив и объект определены достаточно четко: тематив — тема оцениваемой ситуации, объект — компонент с предметно-вещественным значением, подвергающийся воздействию [2, с. 431]. Однако при составлении базы данных предложных конструкций, по-видимому, не всегда оказывалось возможным однозначно определить, какая именно синтаксема реализуется в том или ином контексте. Это подтверждается конструкциями, получившими сразу две семантических метки: «*фильмы о Гарри Поттере*», «*уведомить о результатах*» и т. д. Совершенно ясно, что подобные контексты реализуют одновременно две синтаксемы. Отсюда возникают трудности с разграничением конструкций с предлогом «о» на практике.

2.1.4. Предлог «по»

779 контекстов в базе данных предложных конструкций содержат предлог «по», причем данный предлог охватывает 10 синтаксем (рис. 4).

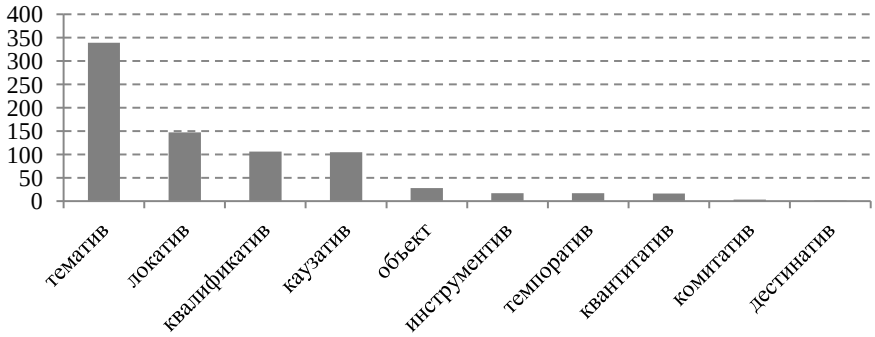

Рис. 4. Столбчатая диаграмма распределения синтаксем предлога «по»

Наиболее характерным для предлога «по» оказывается синтаксема тематив («*комитет по обороне*»). Почти вдвое реже реализуется синтаксема локатив («*гулять по городу*»). Больший интерес вызывают следующие две синтаксемы: квалификатив и каузатив. Поскольку для других производных предлогов данные синтаксемы не так характерны, в совокупности с локативом и темативом они очерчивают семантику предлога «по».

Квалификатив определяется как компонент, обозначающий качество, свойство предмета («*фасад по чертежам*») [2, с. 431]. Каузатив выражает значение причины действия или проявления признака, свойства («*прибывшие по вызову*»). При этом как каузатив, так и квалификатив для предлога «по» обнаруживают одинаковую частотность: на каждую из синтаксем приходится примерно по 100 контекстов.

2.1.5. Предлог «с»

Предлог «с» представлен в базе данных 768 контекстами и, как и первый по частотности предлог «в», охватывает 13 синтаксем (рис. 5). Последний факт ещё раз указывает на разрозненность значений первообразных предлогов. Предлог «с» примечателен среди прочего тем, что наиболее частотная для него синтаксема, комитатив, оказывается лишь на десятом месте среди всех синтаксем первообразных предлогов. В «Синтаксическом словаре» комитатив определяется как «компонент, обозначающий сопровождающее действие, признак, сопутствующий предмет, соучаствующее лицо» («*дом с апартаментами*», «*два с половиной*») [2, с. 431]. Следующая по частотности синтаксема — объект («*гулять с друзьями*», «*пакет с деньгами*»).

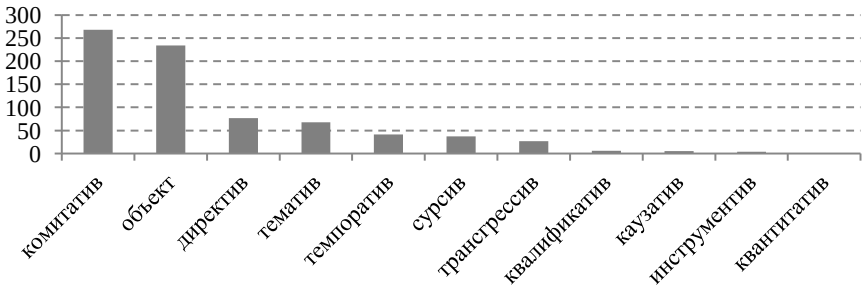

Рис. 5. Столбчатая диаграмма распределения синтаксем предлога «с»

2.2. Производные предлоги и их значения

Путем автоматического анализа свыше 1000 примеров использования было получено распределение синтаксем производных предлогов по частоте. Наиболее частотными оказались следующие: темпоратив, квалификатив, тематив, каузатив, локатив (рис. 6).

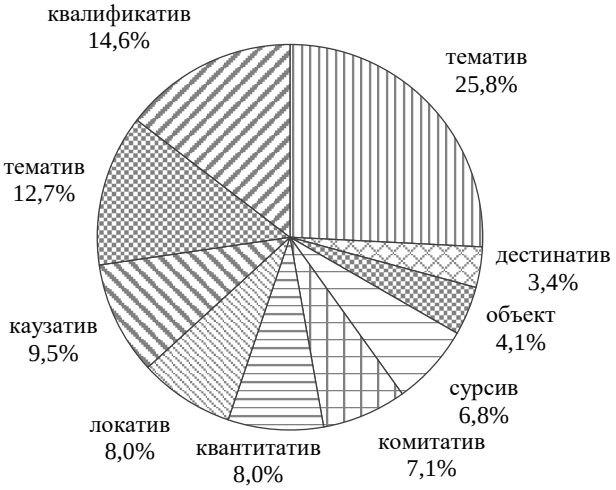

Рис. 6. Круговая диаграмма распределения синтаксем производных предлогов

2.2.1. Предлог «после»

Наиболее частотным среди производных предлогов оказывается предлог «после». Однако в сравнении с первообразными предлогами мы замечаем, что частотность производных в десятки раз меньше (ср. более 3000 контекстов для предлога «в» и чуть больше 100 контекстов для предлога «после»). При этом число производных предлогов почти в 8 раз превосходит число первообразных. Нужно отметить тот факт, что производные предлоги обнаруживают более чёткую семантическую структуру, что объясняется мотивированностью их знаменательными частями речи [3, с. 10]. Так, предлог «после» охватывает лишь три синтаксемы (ср. диапазон значений предлога «в» насчитывает 13 синтаксем). Наиболее характерной для этого предлога оказывается синтаксема темпоратив («*после долгого перерыва*»). Вдвое реже встречаются контексты с данным предлогом, размеченные как каузатив («*задержанные после массовой драки*»).

2.2.2. Предлог «около»

Вторым по частотности среди производных предлогов оказывается предлог «около» (90 контекстов). В первую очередь мы замечаем синтаксемы квантитатив («*около двух метров*») и темпоратив («*около месяца назад*»), на каждую из которых приходится примерно по 40 контекстов. И всего лишь 11 конструкций с предлогом «около» размечены как локатив («*припаркованный около дома*»).

2.2.3. Предлог «между»

Предлог «между» представлен 48 контекстами. При этом большая часть контекстов реализуют значение синтаксемы объект («*разница между сборами и выплатами*»). Нельзя не заметить, что в случае с предлогом «между» все контексты размечаются однозначно, мы не наблюдаем двух семантических меток (тематив и объект) у конструкций с данным предлогом. Всего лишь 11 контекстов приходится на синтаксему локатив («*граница между Турцией и Сирией*»).

2.2.4. Предлог «в связи с»

39 контекстов в базе данных предложных конструкций содержат предлог «в связи с», при этом все контексты почти поровну делятся между синтаксемами каузатив («*отложить в связи с неявкой*») и тематив («*утверждают в связи со вступлением*»). Заметим также, что и в случае с предлогом «в связи с» имеет место неоднозначность. Равное распределение контекстов по двум синтаксемам явно указывает на то, что в действительности синтаксемы тематив и каузатив часто пересекаются.

2.2.5. Предлог «по мнению»

Следующий по частотности предлог — «по мнению» — представлен 37 контекстами, все из которых получили метку сурсив («*по мнению строителей*»). Значение источника информации оказывается единственным для данного предлога.

3. Анализ синтаксической неоднозначности: выявление наиболее частотных синтаксем и предлогов в корпусе неоднозначных предложений

В рамках исследования была изучена синтаксическая неоднозначность на материале корпуса из 200 предложений, заимствованных из диссертационного исследования Д. А. Черновой «Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование» [6]. Целью эксперимента являлось определение наиболее проблематичных для анализа синтаксем и предлогов.

Для выявления наиболее частотных предлогов в синтаксически неоднозначных предложениях был проведен компьютерный анализ на языке Python, включая использование стандартных библиотек для обработки строк и модуля collections для количественного анализа элементов текста. Разбиение текста на предложения осуществлялось с помощью библиотеки nltk, а подсчет частотности каждого предлога проводился с использованием структуры данных Counter.

Результаты анализа указывают на значительную частотность употребления предлога «в», составившего 35,29 % от общего числа употреблений предлогов, «на» — 21,32 %, «о» — 13,24 %, «с» — 12,5 %, «за» — 5,88 %, «у» — 5,15 %, «под», «после», и «над» составили 2,21 %.

Для определения наиболее часто встречающейся синтаксемы был задействован классификатор на основе многослойного перцептрона (MLP). В качестве обучающих данных использовались предложения, размеченные по основным синтаксемам русского языка: локатив, темпоратив, тематив, объект и директив, которые были выявлены в рамках текущего исследования. Была применена модель SentenceTransformer, основанная на предварительно обученной модели «DeepPavlov/rubert-base-cased-sentence», для преобразования предложных конструкций в векторные представления.

Классификатор был обучен на основе базы данных предложных конструкций, созданной в рамках исследовательского проекта «Квантитативная грамматика русских предложных конструкций» [7]. Из общего числа 11 122 контекстов 80 % были выделены для обучающей выборки, а 20 % использовались для тестирования. Для обучения классификатора был использован оптимизационный алгоритм Adam с параметром регуляризации равным 0,001 и максимальным количеством итераций обучения, установленным на уровне 10.

Средняя точность классификатора, взвешенная по объему выборки каждого класса, составила 76 % (табл.). Применение классификатора к исходному набору данных выявило, что синтаксема «тематив» преобладает в 71 % случаев, что указывает на ее доминирующую роль в структуре синтаксически неоднозначных предложений.

Таблица. Оценка работы MLP классификатора

Синтаксемы	**Меры**			
	precision	**recall**	**f1-score**	**support**
директив	0,91	0,65	0,76	115
локатив	0,82	0,92	0,87	303
объект	0,54	0,52	0,53	139
тематив	0,60	0,63	0,61	198
темпоратив	0,92	0,88	0,90	176
accuracy			0,76	931
macro average	0,76	0,72	0,73	931
weighted average	0,76	0,76	0,75	931

Таким образом, результаты исследования указывают на значительную частотность употребления предлога «в» в синтаксически неоднозначных предложениях, а также на преобладание синтаксемы «тематив» в структуре этих предложений.

4. Заключение

Семантическая структура предлогов оказывается недостаточно четко очерченной, но это совсем не означает, что значения предлогов не поддаются структурному описанию. Даже на примере, хоть и небольшой, но семантически размытой, группы первообразных предлогов мы видим, что самые частотные предлоги обнаруживают отчетливые различия в наборе реализуемых значений. Более того, частота встречаемости синтаксемы прямо коррелирует с количеством ее значений. Самые частотные предлоги и синтаксемы оказываются самыми неоднозначными, а со снижением частотности предлога сужается и

диапазон его значений. Интересно то, что данная закономерность характерна именно для производных предлогов. На примере же первообразных предлогов мы увидели, что и не самые частотные предлоги могут охватывать свыше 10 синтаксем.

Полученные результаты могут быть полезны для дальнейших исследований в области синтаксического и семантического анализа текста и разработки методов автоматизированного извлечения синтаксических структур.

Литература

[1] Захаров В. П., Азарова И. В., Головина А. В., Гудков В. В., Москвина А. Д. Квантитативная онтология и база данных русских предлогов // Вестник РФФИ. Гуманитарные и общественные науки. 2022. № 109. С. 17–26.

[2] Золотова Г. А. Синтаксический словарь: репертуар элементарных единиц русского синтаксиса. 4-е изд. М.: Наука, 1988. 440 с.

[3] Азарова И. В., Захаров В. П., Москвина А. Д. Семантическая структура русских предложно-падежных конструкций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXI Международной объединенной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 30 мая – 2 июня 2018 г. Сборник научных статей). — СПб.: Университет ИТМО, 2018. С. 9–16.

[4] Митрофанова О. А., Москвина А. Д. О роли статистики предлогов в определении стилистической принадлежности русскоязычных текстов // International Journal of Open Information Technologies. 2020. Т. 8, № 11. С. 91–96.

[5] Сичинава Д. В. Об одном лингвистическом параметре типологии текстов: коэффициент «под/над» // Научно-техническая информация. Серия 2. 2003. № 10. С. 27–35.

[6] Чернова Д. А. Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование: автореф. на соиск. ученой степ. канд. филолог. наук: 10.02.19 — теория языка. СПб., 2016. 23 с.

[7] Квантитативная грамматика русских предложных конструкций / Захаров В. П. [и др.] // Github. URL: https://vintagentleman.github.io/qt_prep_gram/ (дата обращения: 31.03.2024).

Frequency Characteristics of Russian Prepositions and Their Meanings

V. V. Vybornaya, A. M. Goncharova, A. A. Rodina

Saint–Petersburg State University

The article describes the frequency characteristics of the preposition's ratio and their meanings, and also explores the syntactic ambiguity of prepositional constructions in the Russian language. The research material was a database of prepositional constructions created during the project «Quantitative Grammar of Russian Prepositional Constructions» developed at the Department of Mathematical Linguistics of Saint Petersburg State University, as well as a corpus of 200 syntactically ambiguous sentences borrowed from D. A. Chernova's doctoral research «The Process of Processing Syntactically Ambiguous Sentences: A Psycholinguistic Study». Data analysis was conducted using tools from the pandas module and other Python libraries. We start from the position that prepositions, especially non-derived ones, implement different meanings in different contexts. We consider the ratio of prepositions and the semantic labels attributed to them with the aim of identifying patterns in the distribution of prepositions across syntactic constructions, and also determine the most frequent syntactic constructions and prepositions among syntactically ambiguous sentences based on a multilayer perceptron. The results of this study can be useful in addressing tasks related to disambiguation and contribute to the

understanding of the structure of syntactically ambiguous sentences in the Russian language, indicating the prevailing role of the «topic» synatxeme in their structure.

Keywords: Russian prepositions, prepositional constructions, prepositional meanings, synatxemes, syntactic ambiguity

Reference for citation: Vybornaya V. V., Goncharova A. M., Rodina A. A. Frequency Characteristics of Russian Prepositions and Their Meanings // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 61–69. DOI: 10.17586/2541-9781-2024-8-61-69.

References

[1] Zaharov V. P., Azarova I. V., Golovina A. V., Gudkov V. V., Moskvina A. D. Kvantitativnaya ontologiya i baza dannyh russkih predlogov // Vestnik RFFI. Gumanitarnye i obshchestvennye nauki. 2022. № 109. P. 17–26. (In Russian)

[2] Zolotova G. A. Sintaksicheskij slovar': repertuar elementarnyh edinic russkogo sintaksisa. 4-e izd. M.: Nauka, 1988. 440 s. (In Russian)

[3] Azarova I. V., Zaharov V. P., Moskvina A. D. Semanticheskaya struktura russkih predlozhno-padezhnyh konstrukcij // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXI Mezhdunarodnoj ob"edinennoj konferencii «Internet i sovremennoe obshchestvo», IMS-2019, Sankt-Peterburg, 30 maya – 2 iyunya 2018 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2018. S. 9–16. (In Russian)

[4] Mitrofanova O. A., Moskvina A. D. O roli statistiki predlogov v opredelenii stilisticheskoj prinadlezhnosti russkoyazychnyh tekstov // International Journal of Open Information Technologies. 2020. T. 8, № 11. S. 91–96. (In Russian)

[5] Sichinava D. V. Ob odnom lingvisticheskom parametre tipologii tekstov: koefficient «pod/nad» // Nauchno-tekhnicheskaya informaciya. Seriya 2. 2003. № 10. S. 27–35. (In Russian)

[6] Chernova D. A. Process obrabotki sintaksicheski neodnoznachnyh predlozhenij: psicholingvisticheskoe issledovanie: avtoref. na soick. uchenoj step. kand. filolog. nauk: 10.02.19 — teorija jazyka. SPb., 2016. 23 s. (In Russian)

[7] Kvantitativnaya grammatika russkih predlozhnyh konstrukcij / Zaharov V. P. [i dr.] // Github. URL: https://vintagentleman.github.io/qt_prep_gram/ (access date: 31.03.2024). (In Russian)

Алгоритм сбора текстов для анализа тональности и тематического моделирования отзывов пациентов поликлиник

А. Д. Белкин, М. С. Коган, М. В. Болсуновская

Санкт-Петербургский политехнический университет Петра Великого

redloin@mail.ru, m_kogan@inbox.ru, bolsun_mv@spbstu.ru

Аннотация

Пользовательский фидбек является ценным источником информации для оценки качества услуг, оказываемых медицинскими учреждениями, в частности поликлиниками, и выявления проблемных аспектов. Однако, процесс сбора пользовательских отзывов представляет достаточно сложную техническую задачу ввиду отсутствия единых требований сбора отзывов пациентов на сайтах поликлиник, которым бы следовали все учреждения здравоохранения. Собираемые традиционным способом данные такого типа, как правило, не являются открытыми. В данной работе рассматривается процесс сбора и анализа пользовательских отзывов о медицинских учреждениях на примере поликлиник города Санкт-Петербурга. Для сбора данных был разработан алгоритм на языке программирования Python, использующий веб-скрапинг. Собранный набор данных содержит более 64 тысяч отзывов о 350 поликлиниках. Собранные данные будут предобработаны и использованы для обучения моделей анализа тональности и тематического моделирования. Приводятся результаты предварительного анализа пользовательских отзывов, которые характеризуются большим лингвистическим и стилистическим разнообразием. Полученные результаты могут быть использованы медицинскими учреждениями для улучшения качества обслуживания и повышения удовлетворенности пациентов. Кроме того, модель может быть применена для сравнительного анализа различных учреждений и выявления областей, требующих внимания.

Ключевые слова: отзывы пациентов поликлиник, открытые данные, Яндекс Карты, веб-скрапинг, естественная обработка языка, анализ тональности, тематическое моделирование

Библиографическая ссылка: Белкин А. Д., Коган М. С., Болсуновская М. В., Алгоритм сбора текстов для анализа тональности и тематического моделирования отзывов пациентов поликлиник // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). – СПб: Университет ИТМО, 2024. С. 70–78. DOI: 10.17586/2541-9781-2024-8-70-78.

1. Введение

По мере стремительного развития Интернета новый вид приобретают и различные онлайн-платформы, социальные сети, форумы и т. д. Некоторые подобные площадки агрегируют отзывы интернет-пользователей, описывающие какой-либо товар или заведение.

Отзыв представляет собой текст произвольной длины, написанный пользователем и выражающий его личное отношение к тому или иному объекту. Часто платформы дают возможность оценивать степень удовлетворенности в баллах от 1 до 5.

Пользовательский фидбек может быть очень ценным источником информации для организаций, направленных на предоставление различных услуг людям [1]. В отзывах пользователи могут указывать на неочевидные проблемы, недостатки или несоответствие действительности информации, указанной на сайте организации. Это может помочь быстро выявить и своевременно устранить недочеты. Большое количество положительных отзывов влияет на приток новых клиентов, так как люди в первую очередь обращают внимание на рейтинг и оценки. Также, положительные отзывы могут быть использованы в рекламных целях, демонстрируя их на сайте или в социальных сетях. С другой стороны, отзывы могут помочь в оценке конкурентных заведений или товаров, например, понять сильные и слабые стороны, выявить новые тренды и возможности для роста.

Эффективными инструментами изучения пользовательских отзывов являются анализ тональности и тематическое моделирование [2]. Тематическое моделирование позволяет выделить ключевые темы и проблемы, которые обсуждаются в отзывах [3]. Вдобавок, можно сравнивать различные группы пользователей и выявлять их потребности и предпочтения.

В свою очередь, анализ тональности комментариев и отзывов помогает понять то, какие эмоции выражают пользователи. Обученный на достаточно большом объеме данных алгоритм анализа тональности поможет исследовать данные, не имеющие таких же оценок или меток, как на онлайн-платформах. Существует множество традиционных и современных способов сбора отзывов, которые используются в настоящее время. К традиционным можно отнести бумажные формы для отзывов, как, например, книги жалоб и предложений или заполнение опросных листов [4]. Среди современных способов можно выделить электронные опросы, мониторинг социальных медиа и телефонные интервью. Такие данные не содержат конкретной метрики, говорящей об эмоциональном отношении опрошенного. В таких случаях и может пригодиться модель анализа тональности.

Особенно важно понимать потребности и опыт пациентов в сфере здравоохранения, так как их мнение является ключевым показателем качества. Интернет дал людям возможность выражать свои мнения на различных платформах, что создало большой объем данных, исследование которых может выявить тенденции, не замеченные ранее, в медицинских услугах. Исследователи из Имперского колледжа Лондона собрали комментарии пациентов с сайта NHS Choices за 2008, 2009 и 2011 годы (13 802 текста) и применили методы анализа тональности [5]. Данные за 2010 год (6 412 комментариев) использовались для проверки точности предсказания. Целью было обучение модели машинного обучения для автоматического определения рекомендаций пациентов, чистоты медицинского учреждения и уважительного отношения. Алгоритм классифицировал комментарии на основе примеров и проверял точность предсказания, сравнивая результаты с оценками пациентов по шкале Ликерта. Исследование показало высокую точность предсказаний, подтверждая соответствие прогнозов результатам опросов.

В другом исследовании группа ученых собрала корпус данных с сайта Otzovik, состоящий из 1,4 миллиона пользовательских отзывов на русском языке о здоровье и лекарствах, а также 500 подробно аннотированных отзывов [6]. Корпус называется RuDReC и находится в публичном доступе. Ученые затем использовали эти тексты для обучения моделей Multi-BERT, RuBERT и RuDR-BERT, сравнив их способность извлекать из отзывов пользователей информацию о названии лекарств, побочных эффектах, заболеваниях и здоровье.

В еще одной работе исследователи так же собрали корпус из 2 800 комментариев пользователей на форуме Otzovik. В статье анализируется то, как выбор различных компонентов модели влияет на точность распознавания сущностей. Исследователи применяют тот же подход к аннотированию текстов, что и авторы [6], и обучают на них

алгоритм нейронной сети. Результатом работы стало то, что был установлен новый уровень точности извлечения биомедицинских сущностей для русского языка на полноразмерном размеченном корпусе [7]. Собранные для исследования данные могут быть получены по запросу.

Также стоит отметить статью, в которой описывается способ выявления связей между лекарствами на основе семантического анализа текстовых данных [8]. Авторы применили веб-краулинг для сбора текстов с различных медицинских порталов. Всего было собрано примерно 2,5 млн текстов, из которых были извлечены названия препаратов, описание их действия и другие сущности. Затем полученные данные были векторизованы и сравнены между собой. Эксперименты показали, что такой подход может успешно идентифицировать препараты с похожими терапевтическими эффектами.

Однако в машинном обучении существует проблема нехватки размеченных данных, необходимых для обучения моделей. В открытом доступе не всегда можно найти достаточно информации или уже готовых «датасетов» с метками. Особенно это касается такого домена, как медицинские учреждения. В данном случае нехватка частично обусловлена тем, что структура сайтов поликлиник и больниц не располагает отдельной секцией для отзывов и оценки работы учреждения. Кроме того, очень много данных находится в закрытом доступе, так как они содержатся на бумажных носителях и не оцифрованы.

В последнее время обсуждают важность открытости данных и кода для сотрудничества государств в борьбе с глобальными проблемами. Доклад ЮНЕСКО подчеркивает, что открытые данные обеспечивают доступ к необходимой информации для исследований, политики и мер против глобальных кризисов, таких как COVID-19 [9]. Искусственный интеллект помогает анализировать большие объемы данных, выявляя скрытые закономерности для принятия решений. Для публикации данных нужно выполнить три этапа: разработка политики управления и обмена данными, решение о публикации данных и привлечение пользователей. На первом этапе определяются обязательства, принципы, законодательные ссылки, причины недоступности данных, сбор и обучение пользователей. Затем принимается решение о том, какие наборы данных будут опубликованы и на условиях лицензирования. После этого данные публикуются на веб-сайте. Этап привлечения пользователей включает в себя информирование, консультирование, поддержку международного взаимодействия и предотвращение злоупотреблений. Также, необходимо регулярно обновлять данные для сохранения их актуальности. Можно заметить, что предлагаемая процедура является цикличной поскольку после каждой публикации новых данных, она должна заново проходить все описанные этапы.

Тем не менее, далеко не все страны придерживаются указанных выше рекомендаций и принципов. По этой причине нехватка качественных наборов данных, необходимых для обучения алгоритмов машинного обучения, является актуальной проблемой во многих областях науки.

В связи с этим было принято решение сформировать подобный корпус с нуля при помощи компьютерных средств [10].

2. Разработка алгоритма

Для сбора информации и создания корпуса был разработан алгоритм и реализован на языке программирования Python. В качестве источника данных, с которого программа собирает отзывы, был выбран вебсайт онлайн-карт «Яндекс Карты». Формирование набора данных опирается на оценки, выставленные пользователями учреждениям. На сайте они имеют вид «звездочек» в диапазоне от 1 до 5.

Принцип работы алгоритма сводится к тому, что сначала задаются параметры, по которым будет идти поиск организаций и учреждений на сайте. В результате поиска выпадает список карточек организаций, который затем прокручивается вниз, пока не

достигнет заданного числа искомых организаций. С каждой карточки собирается ссылка, ведущая на соответствующие отзывы. Каждая ссылка обрабатывается, а отзывы, находящиеся на странице, прокручиваются до конца. Затем из них, оценок и имен пользователей формируется набор данных при помощи библиотеки «pandas». Наконец, сформированный «датасет» передается внешней таблице, вбирающей в себя все результаты обработки.

В нашем исследовании для реализации алгоритма использовались следующие библиотеки Python: «Selenium», «Beautiful Soup», «pandas» и «pymorphy3». Первые две библиотеки из перечисленных, необходимы для взаимодействия с веб-браузером и извлечения данных с сайта, соответственно. «Selenium» автоматизирует работу веб-браузеров, а «Beautiful Soup» облегчает «веб-скраппинг», разбирая HTML- и XML-файлы [11].

Программная реализация алгоритма состоит из 6 функций, которые в процессе работы многократно вызываются. Прежде всего задаются параметры веб-драйвера, который будет моделировать действия пользователя в браузере, давая тем самым доступ к необходимой информации.

На первом этапе указывается, какой браузер из 5 доступных будет использоваться для работы, поскольку структура сайтов для разных браузеров отличается. Для данного исследования был выбран «Google Chrome». Для ускорения работы кода следует применить специальный параметр, запрещающий использовать режим с графическим интерфейсом браузера. Данный режим весьма полезен, но только при разработке и экспериментах, так как он показывает каждое действие, которое было запрограммированно. В случае обычного использования программы нет необходимости следить за процессом моделирования работы браузера, так как это может оказать дополнительную нагрузку на компьютер.

На втором этапе задаются функции, которые затем объединяются в главной, исполняющей и автоматизирующей весь процесс. Первая функция очень важна, так как она необходима для имитации пролистывания страницы при помощи ползунка, поскольку одной из главных сложностей сбора данных стала процедурная загрузка отзывов и учреждений на сайте. Без этого решения программа может собирать только первые 5 предзагруженных текстов и ссылок.

Третья функция отвечает за циклический поиск в коде сайта ссылок, ведущих с карточек найденных учреждений на отзывы к ним.

Следующая функция совершает парсинг отзывов, поочередно переходя по заданному списку ссылок. Поскольку сайт не всегда успевает загрузиться быстро, необходимо искусственно замедлить работу кода, давая возможность прогрузиться всем нужным элементам. В зависимости от количества отзывов программа подсчитывает приблизительное число прокруток ползунка, чтобы загрузить все тексты. Для этого используется библиотека «time», которая тормозит процесс выполнения кода на нужное количество секунд. Также на этом этапе вместе с отзывами собираются и соответствующие им оценки от 1 до 5.

Стоит отметить, что не все пользователи ставят оценки. Хотя по нашим наблюдениям это довольно редкое явление, его необходимо учитывать: в противном случае, сталкиваясь с отсутствием оценки, код будет аварийно завершаться. Такие отзывы включаются в набор данных, но со специальной меткой. На этапе предобработки данных подобные тексты будут удалены или размечены вручную, если их количество будет достаточно большим. Помимо оценок данная функция собирает имена пользователей, которые затем используются для определения пола комментатора.

На финальном этапе главная функция принимает на вход параметры поиска данных: тип организации, город для сбора отзывов, район и количество организаций, которое нужно обработать. Количество организаций указывается исследователем или иным пользователем, желающим сформировать набор данных в каком-либо домене. Эти параметры попадают в поисковую строку на сайте, после чего начинают работать другие функции. При этом

функция контролирует, чтобы параметры вводились правильно. Если какой-то параметр не был введен, то в работе будет использовано базовое значение этого параметра.

Если было найдено недостаточное число организаций, запускается цикл, дополняющий список ссылок до тех пор, пока не наберется нужное количество. В случае, если было обработано слишком много сайтов, то список будет обрезан до обозначенного числа. Во избежание бесконечного поиска количество итераций ограничено максимальным числом допустимых циклов, к которым прибавляется еще 5 дополнительных.

Когда процесс сбора завершен, функция, используя цикл, объединяет все полученные данные в единый датасет и возвращает его для дальнейшей работы.

В рамках исследования программа собрала 64 451 отзыв о 350 поликлиниках Санкт-Петербурга. Время работы программы составило примерно 3 часа. В таблице приведены 5 примеров из собранного набора отзывов (в отзывах сохранена авторская орфография; ФИО упомянутых в отзывах сотрудников медучреждений были закодированы путём сокращения имен, фамилий и отчеств до инициалов). Фамилии упомянутых врачей скрыты с целью сохранения безопасности личной информации.

Таблица. Примеры отзывов из полученного набора данных

Номер строки	Текст отзыва	Оценка
60250	Хорошая поликлиника, нареканий нет!	5
43027	Лучшая поликлиника, все врачи отличные специалисты. В ПК есть лекотека со своим логопедом, психологом, массажистом во благо развития здоровья детей. Отдельно отмечу невролого до года О. О. В. и педиатра Т. В. Ф. внимательное отношение к детям и бережное отношение ко мне как к маме. Отвечали на все мои вопросы	5
40054	Сложно попасть к специалистам	3
3586	Хорошие специалисты	5
7818	Нехватает специалистов. Электронная очередь и по записи. Все равно у кабинетов толпа без билетников. Лезут без очереди. Спрашивается зачем тогда предварительная запись, когда все и так проходят. Бардак. Хотя врачи попадаются толковые. Но некоторые спешат, так как торопятся поскорее убежать на платный прием	3

Как показано на рисунке, основные темы, выявленные в отзывах, включают обсуждение медицинского персонала, медицинских учреждений, процедур, записей и самих процессов на прием ко врачу, а также инфраструктуры.

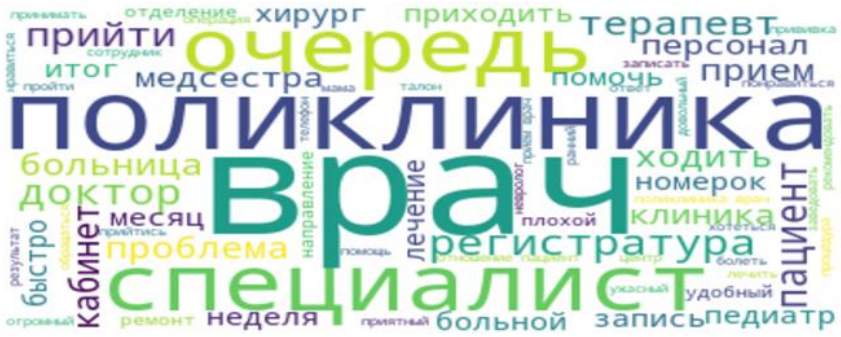

Рис. Облако слов, сформированное на основе отзывов

Поскольку нет строгих правил и ограничений, диктующих то, как надо писать отзывы, в них содержится множество индивидуальных лингвистических особенностей пользователей, которые будут кратко описаны в следующем разделе.

3. Предварительный лингвистический и стилистический анализ собранного датасета

Пациенты поликлиник, оставляющие отзывы, часто используют в своих сообщениях разнообразные фразеологизмы, сленг и эмодзи. Они также часто совершают грамматические и пунктуационные ошибки и выделяют отдельные слова заглавными буквами. По причине того, что отзывы посвящены медучреждениям, пользователи часто описывают свой опыт посещения врачей, упоминая конкретные даты, имена врачей и другую информацию. К примеру, некоторые пациенты выражают благодарность отдельным врачам, которые оказали им помощь. Другие, наоборот, критикуют кого-то, обращаются к администрации больницы или поликлиники, призывая принять меры.

Зачастую пользователи выражают свое отношение весьма эмоционально, используя в тексте скобочки для демонстрации эмоции. Их количество может варьироваться от одной до нескольких десятков. Они могут выражать как радость («))»), так и печаль («((»), а их количество указывает на силу испытанных эмоций. Более того, в некоторых случаях множественные скобочки сопровождают язвительные, саркастические и иронизирующие комментарии. Кроме того, пациенты могут ставить многоточие. В каких-то случаях многоточие используется при перечислении большого списка объектов, например специальностей врачей, либо для выражения расстройства, когда пациент не получил того, что ожидал. Также, для акцентирования внимания на каких-либо конкретных аспектах, пользователи выделяют слова заглавными буквами. Как правило, это встречается в негативных отзывах, например, при описании большого времени ожидания в очереди («ЧАС») или при неудовлетворительном качестве обслуживания в платной поликлинике («ПЛАТНАЯ»).

Вдобавок ко всему перечисленному, пациенты регулярно используют слэнг, неформальные слова и выражения. Они могут быть специфичны для определенных социальных групп, возрастных категорий или регионов и зачастую могут меняться и развиваться со временем (выявление этих закономерностей может стать предметом отдельного исследования в будущем). Например, часто можно встретить преобразованные для простоты употребления медицинские термины. Вместо слова «флюорография» вероятнее увидеть слово «флюшка». Очень много пользователей используют большое количество разговорных выражений в своих отзывах. Так, они подчеркивают ту или иную эмоцию, которую стремятся донести до других. К примеру, некоторые пользователи так описывают хорошую работу медперсонала: «Медсестра колет как богиня» или «просто крутой профессионал своего дела». Негативные отзывы так же очень богаты на примеры: «аж волосы дыбом», «фикалии в уши льют...» или «с горем пополам профосмотр был пройден».

Отдельно можно упомянуть эмодзи, которые встречаются в пользовательских отзывах. Эмодзи — это набор символов, представляющих широкий спектр эмоций, объектов, действий и идей. Они широко используются в современной коммуникации в социальных сетях, мессенджерах, форумах и других онлайн-платформах. Эмодзи позволяют выразить эмоции более точно и наглядно, чем простые текстовые сообщения, а в некоторых случаях, они могут заменять слова или целые фразы, экономя место и делая сообщение более лаконичным.

4. Дальнейшее исследование

Полученный набор данных станет основой для обучения ряда моделей анализа тональности, из которых будет выбрана наиболее эффективная. Однако перед этим данные следует предобработать. Для этого из данных будут извлечены в отдельный столбец эмодзи, которые были использованы пользователями для усиления эмоциональной составляющей отзыва. Столбец с эмодзи может быть использован для улучшения качества модели

тональности текста. Кроме того, данные нужно проверить на дубликаты, так как некоторые пользователи могут писать несколько полностью совпадающих отзывов. Они могут быть посвящены разным учреждениям, но одинаковые комментарии от одного пользователя снижают качество данных. В полученном наборе данных обнаружилось чуть более 2 тысяч дубликатов. Затем данные разделяются на токены, слова приводятся к основе и лемматизируются. Это необходимо для уменьшения размера словаря и улучшения обобщающей способности модели. После этого данные преобразуются в числовой формат, необходимый для обучения модели. Для этого используется один из существующих методов векторизации текстовых данных.

За обучением модели анализа тональности следует обучение алгоритма тематического моделирования. Выделение ключевых тем в отзывах с разделением на положительные и отрицательные, а также с разделением пользователей на группы позволит проанализировать наиболее важные потребности, недостатки и преимущества, которые находятся в текстах.

5. Заключение

Итогом данного исследования стал алгоритм, который собирает пользовательские отзывы на сайте «Яндекс Карты», используя настраиваемые параметры. Также было проанализировано влияние отзывов об различных организациях и составлен набор данных, содержащий отзывы пользователей о поликлиниках в городе Санкт-Петербург. Эти данные будут использованы для обучения модели анализа тональности. Сам алгоритм легко настраивается на сбор аналогичных данных о медучреждениях или других организациях, предоставляющих услуги населению, расположенных в любом регионе страны, при условии, что информация о них содержится на сайте «Яндекс Карты».

Обученная модель может помочь оценить удовлетворенность пациентов услугами медицинских учреждений, которые в свою очередь принять во внимание выявленные проблемы и предпринять меры по улучшению качества услуг. Более того, модель может использоваться для сравнительного анализа различных медицинских учреждений с целью выявления лучших практик и определения областей, в которых могут быть внесены улучшения.

Литература

[1] Dhahak K., Huseynov F. The Impact of Online Consumer Reviews (OCR) on Online Consumers' Purchase Intention // Journal of Business Research-Turk. 2020. Vol. 12 (2). P. 990–1005. URL: https://isarder.org/2020/vol.12_issue.2_article01.pdf (дата обращения 21.03.2024).

[2] Самигулин Т., Джурабаев А. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. № 1. URL: https://rrinformation.ru/journal/annotation/2376/ (дата обращения: 01.03.2024).

[3] Косарева Е., Давыдик Н. Применение тематического моделирования для интеллектуального анализа отзывов на русском языке // Дистанционные образовательные технологии: сб. трудов V Междунар.науч.-практ. конф., Симферополь, 22–25 сент. 2020 г. Симферополь: ИТ «АРИАЛ», 2020. С. 228–231. URL: https://elib.grsu.by/doc/65168 (дата обращения: 20.03.2024).

[4] Бурый М., Кравцова Т. Книга отзывов и предложений как элемент совершенствования качества услуг // E-Scio. 2023. №5 (80). URL: https://e-scio.ru/?p=21000 (дата обращения: 20.03.2024).

[5] Greaves F., Ramirez-Cano D., Millett C., Darzi A., Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online // Journal of Medical Internet Research. 2013. Т. 15 (11): e239. DOI: 10.2196/jmir.2721. URL: https://www.jmir.org/2013/11/e239/ (дата обращения: 24.03.2024).

[6] Tutubalina E., Alimova I., Miftakhutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews // Bioinformatics. 2020. T. 37 (2). DOI: 10.1093/bioinformatics/btaa675.
[7] Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Selivanov A., Rylkov G., Ilyin V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models // Applied Sciences. 2022. T. 12 (1). DOI: 10.3390/app12010491.
[8] Tutubalina E., Miftakhutdinov Z., Nugmanov R., Madzhidov T., Nikolenko S., Alimova I., Tropsha A. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects // Russian Chemical Bulletin. 2017. Vol. 66 (11). P. 2180–2189.
[9] Ziesche S. Open data for AI. What now? Paris: UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385841. 64 p. (дата обращения: 24.03.2024).
[10] Чижик А., Мельникова С., Захаров В. Социальное картирование на основании анализа тональности комментариев в социальных сетях // International Journal of Open Information Technologies. 2022. Т. 10, № 11. С. 75–79. URL: http://injoit.org/index.php/j1/article/view/1434 (дата обращения: 04.03.2024).
[11] Бастрикина В. Проектирование веб-скрапера для получения данных с сайтов книжных издательств // Актуальные проблемы авиации и космонавтики. 2018. № 14. С. 125–127.

Text Collection Algorithm for Sentiment Analysis and Topic Modelling of Patient Reviews for Polyclinics

A. D. Belkin, M. S. Kogan, M. V. Bolsunovskaya

Peter the Great St. Petersburg Polytechnic University

User feedback is a valuable source of information for assessing the quality of services provided by health care facilities, particularly polyclinics, and identifying problematic aspects. However, the process of collecting user feedback is technically challenging, as there are no uniform requirements for collecting patient reviews on polyclinic websites that all health care institutions should follow. This type of data collected in the traditional way is usually not open. This paper examines the process of collecting and analysing user feedback on health care institutions on the example of polyclinics in the city of St. Petersburg. For data collection, an algorithm was developed in the Python programming language using web scraping. The collected data set contains more than 64 thousand reviews of 350 polyclinics. The collected data will be preprocessed and used to train tone analysis and topic modelling models. The results of a preliminary analysis of user reviews, which are characterised by a large linguistic and stylistic diversity, are presented. The results can be used by healthcare providers to improve the quality of care and patient satisfaction. In addition, the model can be applied to benchmark different institutions and identify areas that require attention.

Keywords: polyclinic patients' feedback, open data, Yandex Maps, web scraping, natural language processing, sentiment analysis, topic modelling

Reference for citation: Belkin A. D., Kogan M. S., Bolsunovskaya M. V. Text Collection Algorithm for Sentiment Analysis and Topic Modelling of Patient Reviews for Polyclinics // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 70–78. DOI: 10.17586/2541-9781-2024-8-70-78.

Reference

[1] Dhahak K., Huseynov F. The Impact of Online Consumer Reviews (OCR) on Online Consumers' Purchase Intention // Journal of Business Research-Turk. 2020. Vol. 12 (2). P. 990–1005. URL: https://isarder.org/2020/vol.12_issue.2_article01.pdf (access date: 21.03.2024).

[2] Samigulin T., Djurabaev A. Analiz tonalnosti teksta metodami mashinnogo obucheniya // Nauchnyy rezultat. Informatsionnye tekhnologii. 2021. № 1. URL: https://rrinformation.ru/journal/annotation/2376/ (access date: 01.03.2024). (In Russian)

[3] Kosareva E., Davydik N. Primenenie tematicheskogo modelirovaniya dlya intellektual'nogo analiza otzyvov na russkom yazyke // Distantcionnye obrazovatel'nye tekhnologii: sb. trudov V Mezhdunar. nauch.-prakt. konf., Simferopol', 22–25 sent. 2020 g. — Simferopol': IT "ARIAL", 2020. — S. 228–231. URL: https://elib.grsu.by/doc/65168 (access date: 20.03.2024). (In Russian)

[4] Buryy M., Kravtsova T. Kniga otzyvov i predlozheniy kak element sovershenstvovaniya kachestva uslug // E-Scio. 2023. № 5 (80). URL: https://e-scio.ru/?p=21000 (access date: 20.03.2024). (In Russian)

[5] Greaves F., Ramirez-Cano D., Millett C., Darzi A., Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online // Journal of Medical Internet Research. 2013. T. 15 (11): e239. DOI: 10.2196/jmir.2721. URL: https://www.jmir.org/2013/11/e239/ (access date: 24.03.2024).

[6] Tutubalina E., Alimova I., Miftakhutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews // Bioinformatics. 2020. T. 37 (2). DOI: 10.1093/bioinformatics/btaa675.

[7] Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Selivanov A., Rylkov G., Ilyin V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models // Applied Sciences. 2022. T. 12 (1). DOI: 10.3390/app12010491

[8] Tutubalina E., Miftakhutdinov Z., Nugmanov R., Madzhidov T., Nikolenko S., Alimova I., Tropsha A. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects // Russian Chemical Bulletin. 2017. Vol. 66 (11). P. 2180–2189.

[9] Ziesche S. Open data for AI. What now? Paris: UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385841. 64 p. (access date: 24.03.2024).

[10] Chizhik A., Mel'nikova S., Zaharov V. Social'noe kartirovanie na osnovanii analiza tonal'nosti kommentariev v social'nyh setyah // International Journal of Open Information Technologies. 2022. T. 10, № 11. S. 75–79.URL: http://injoit.org/index.php/j1/article/view/1434 (access date: 04.03.2024). (In Russian)

[11] Bastrykina V. Proektirovanie veb-skrapera dlya polucheniya dannykh s saytov knizhnykh izdatel'stv // Aktual'nye problemy aviatsii i kosmonavtiki. 2018. № 14. S. 125–127. (In Russian)

Сведения об авторах

Адамова Мария Антоновна, Санкт-Петербургский государственный университет, студент.

Белкин Антон Дмитриевич, Санкт-Петербургский политехнический университет Петра Великого, студент, ORCID 0009-0006-3147-6986.

Болсуновская Марина Владимировна, кандидат технических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, доцент, ORCID 0000-0001-6650-6491.

Букреева Людмила Александровна, Санкт-Петербургский государственный университет, студент.

Выборная Вероника Витальевна, Санкт-Петербургский государственный университет, студент, ORCID 0009-0008-0041-9705.

Голубев Ростислав Васильевич, Санкт-Петербургский государственный университет, студент.

Гончарова Алина Максимовна, Санкт-Петербургский государственный университет, студент, ORCID 0009-0003-3542-2801.

Гусяцкая Полина Андреевна, Санкт-Петербургский государственный университет, студент.

Зернова Алиса Кирилловна, Санкт-Петербургский государственный университет, студент.

Коган Марина Самуиловна, кандидат технических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, доцент, ORCID 0000-0002-7519-2161.

Литвинова Анна Артемовна, Санкт-Петербургский государственный университет, студент.

Макеев Кирилл Владимирович, Санкт-Петербургский государственный университет, студент.

Митрофанова Ольга Александровна, кандидат филологических наук, доцент, Санкт-Петербургский государственный университет, доцент, ORCID 0000-0002-3008-5514.

Павликова Владислава Станиславовна, Санкт-Петербургский государственный университет, студент.

Плюснина Елизавета Алексеевна, Санкт-Петербургский государственный университет, студент.

Родина Анна Андреевна, Санкт-Петербургский государственный университет, студент, ORCID 0009-0002-4730-1946.

Сологуб Полина Юрьевна, Санкт-Петербургский государственный университет, студент.

Сухан Даниил Дмитриевич, Санкт-Петербургский государственный университет, студент, Just AI, разработчик.

Трошина Александра Валерьевна, Санкт-Петербургский государственный университет, студент.

Уткина Александра Алексеевна, Санкт-Петербургский государственный университет, студент.

Авторский указатель

Содержание

Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г.) Сборник научных трудов. — СПб.: Университет ИТМО, 2024. — 83 с.

Компьютерная лингвистика и вычислительные онтологии

Выпуск 8

Сборник научных трудов

Под редакцией А. В. Чижик
Дизайн обложки С. Н. Ушаков
Оригинал-макет А. С. Метелева, Ю. В. Байкеева
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н. Ф. Гусарова
Подписано к печати 20.12.2024
Заказ № 4789 от 20.12.2024
Тираж 50 экз.

Университет ИТМО. 197101, Санкт-Петербург,
Кронверкский пр., 49, лит. А.