Министерство науки и высшего образования
Российской Федерации

УНИВЕРСИТЕТ ИТМО

Некоммерческое партнёрство ПРИОР Северо-Запад

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ВЫЧИСЛИТЕЛЬНЫЕ ОНТОЛОГИИ

Выпуск 7

Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г.

Сборник научных трудов

Санкт-Петербург

2024

УДК 800 (075.3)
ББК 81.1
К63

К63 **Компьютерная лингвистика и вычислительные онтологии. Выпуск 7** (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023). Санкт-Петербург, 26–28 июня 2023 г. Сборник научных трудов — СПб.: Университет ИТМО, 2024. — 100 с.

ISSN 2541-9781
ISBN 978-5-7577- 0710-5

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

Издание адресовано научным работникам, преподавателям, аспирантам и магистрантам, интересующимся междисциплинарными проблемами обработки естественного языка, представления знаний, разработки информационных систем.

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

 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 Минцифры России, Комитета информатизации и связи и Комитета по науке и высшей школе Санкт-Петербурга.

УДК 800 (075.3)
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 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 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 Петербурге. С 2019 года ИТМО самостоятельно присуждает ученые степени кандидата и доктора наук.

ISBN 978-5-7577-0710-5

XXVI Международная объединенная научная конференция **«Интернет и современное общество»** (IMS-2023)

Санкт-Петербург, 26–28 июня 2023 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, с 2016 г. — в рамках Недели технологий информационного общества в Санкт-Петербурге. Объединенная конференция «Интернет и современное общество» в 2023 г. была проведена при поддержке Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, Комитета по науке и высшей школы 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 объединенной, так как научная программа конференции консолидирует серию специализированных международных и российских научных конференций, симпозиумов, семинаров, круглых столов и других мероприятий, посвященных специальным вопросам развития технологий информационного общества. Отдельные специализированные и проблемно-ориентированные мероприятия проводятся в сотрудничестве с партнерскими организациями.

Основу научной программы конференции 2023 года составили международные семинары, включающие сессии на русском и английском языках:

- **Электронное управление** (E-Governance — eGov-2023);
- **Цифровая урбанистика** (Digital City — DCity-2023);
- **Компьютерная лингвистика** (Computational Linguistics — CompLing-2023);
- **Киберпсихология** (Internet Psychology — IntPsy-2023).

Традиционно в программу конференции были включены также сессии научных докладов:

- **Электронное обучение и дистанционные образовательные технологии;**
- **Культурология киберпространства.**

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

- Международный научно-практический симпозиум **«Цифровое здравоохранение и перспективы развития концепции активного долголетия** / Digital Health and Active Aging Development». Симпозиум проводится второй год в сотрудничестве с Хуачжунским университетом науки и технологии, Ухань, Китай (Huazhong University of Science and Technology, Wuhan, China) 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»;
- International Workshop **«Interactive Systems & Information Society Technologies»** (InterSys2023) был проведён в сотрудничестве с партнёрами из Бразилии и Индии: Федеральный университет Параны, Куритиба, Бразилия (Federal University of Paraná, UFPR, Curitiba, Brazil); Институт технологий и науки Бирла, Пилани, кампус в Дубае (Birla Institute of Technology & Science, BITS, Pilani, Dubai Campus);

- Межрегиональный семинар **«Электронное участие в регионах России 2020-2023: состояние и перспективы»** (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»);
- Круглый стол **«Экосистема городских сервисов Санкт-Петербурга: текущее состояние и перспективы развития»** (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» и в сотрудничестве с Санкт-Петербургским информационно-аналитическим центром);
- Научно-практический симпозиум «Этико-правовые аспекты цифровой трансформации»;
- Симпозиум молодых учёных «Киберпространство: перспективные социально-экономические и гуманитарные исследования»;
- Young Scholars' Poster Session «Digital Transformation in Governance and Society» (Young DTGS-2023).

Отбор докладов на конференции и текстов для публикации производится по результатам двойного слепого рецензирования членами программного комитета с использованием международной системы сопровождения научных конференций Easychair.org.

По результатам объединенной конференции IMS-2023 издаются три сборника научных трудов (сериальные издания), сборник тезисов на русском языке и сборник статей на английском языке:

- **Информационное общество: образование, наука, культура и технологии будущего** (ISSN 2587-8557), вып. 7;
- **Государство и граждане в электронной среде** (ISSN 2541-979X), вып. 7;
- **Компьютерная лингвистика и вычислительные онтологии** (ISSN 2541-9781), вып. 7;
- **Интернет и современное общество**: сборник тезисов докладов IMS-2023.

Статьи, представленные для докладов на английском языке и прошедшие рецензирование, включены в сборник, подготовленный совместно с зарубежными партнерами конференции. Сборник публикуется в издательстве Springer (индексация в базе Scopus). В сборник включены также научные статьи, отобранные на конкурсной основе за авторством молодых учёных — участников Young DTGS-2023.

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 года конференция сотрудничает с научным журналом **«International Journal of Open Information Technologies»** (http://injoit.org, ВАК, РИНЦ), издаваемым в МГУ, по формированию специального номера. В 2023 г. такой номер также подготовлен и издан;
- Международный научный электронный журнал **«Культура и технологии»** (http://cat.ifmo.ru/) регулярно публикует лучшие статьи авторов IMS по своей тематике;
- С 2022 года началось партнерство с научным журналом **«Journal on Interactive Systems»** (https://sol.sbc.org.br/journals/index.php/jis), Бразилия.

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

ПРОГРАММНЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Председатель Программного комитета:

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

Заместители председателя Программного комитета:

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

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

Члены Программного комитета:

Алехин А. Н., докт. мед. наук, Российский государственный педагогический университет им. А. И. Герцена

Азарова И. В., канд. филол. наук, Санкт-Петербургский государственный университет

Алексеев А. М., Санкт-Петербургское отделение Математического института им. В. А. Стеклова РАН

Аркатов Д. А., НИУ «Высшая школа экономики» — Санкт-Петербург

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

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

Богдановская И. М., канд. психол. наук, Российский государственный педагогический университет им. А. И. Герцена

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

Бродовская Е. В., докт. политич. наук, Финансовый университет при Правительстве РФ

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

Вяхирева В. В., Нижегородский государственный университет им. Н. И. Лобачевского

Гаврилов А. В., канд. техн. наук, Новосибирский государственный технический университет

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

Голубева А. А., канд. эконом. наук, Санкт-Петербургский государственный университет

Григорьева И. А., докт. социол. наук, Социологический институт РАН — филиал ФНИСЦ РАН

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

Детинко Ю. И., канд. филол. наук, Сибирский федеральный университет

Захаров В. П., канд. филол. наук, Санкт-Петербургский государственный университет

Игнатьев А. В., докт. технич. наук, Волгоградский государственный технический университет

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

Игнатьева С. В., 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

Ковальчук С. В., канд. технич. наук, Университет ИТМО

Коваленко К. И., канд. филол. наук, Европейский университет в Санкт-Петербурге, Институт лингвистических исследований РАН

Колмогорова А. В., докт. филол. наук, НИУ «Высшая школа экономики» — Санкт-Петербург

Королева Н. Н., докт. психол. наук, Российский государственный педагогический университет им. А.И. Герцена

Лактюхина Е. Г., канд. социол. наук, Волгоградский государственный университет

Ларионов И. Ю., канд. филос. наук, Санкт-Петербургский государственный университет

Лукашевич Н. В., докт. технич. наук, Московский государственный университет им. М. В. Ломоносова

Магировская О. В., докт. филол. наук, доцент, Сибирский федеральный университет

Prof. Jingdong MA, PhD, Huazhong University of Science and Technology, China
Krissia MENEZES, Federal University of Paraná, Brazil
Dr. Raja MUTHALAGU, Birla Institute of Technology and Science Pilani, UAE
Radka NACHEVA, PhD, University of Economics, Bulgaria
Júlia Bathke ORTIZ, Federal University of Paraná, Brazil
Dr. Pranav M PAWAR, Birla Institute of Technology and Science Pilani, UAE
Roberto PEREIRA, PhD, Federal University of Paraná, Brazil
Dr. Tamizharasan PERIYASAMY, Birla Institute of Technology and Science Pilani, UAE
Alisson Andrey PUSKA, Federal University of Paraná, Brazil
Elakkiya R, PhD, Birla Institute of Technology and Science Pilani, UAE
Gustavo Yuji SATO, Federal University of Paraná, Brazil
Olga SCRIVNER, PhD, Rose-Hulman Institute of Technology, USA
Zicheng WANG, PhD, Hunan Agricultural University, China
Wei ZHANG, PhD, Huazhong University of Science and Technology, China
Zhaozi ZHAO, Huazhong University of Science and Technology, China

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

Председатель оргкомитета:

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

Заместитель председателя оргкомитета:

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

Члены оргкомитета:

Бакаев М. А., Новосибирский государственный технический университет
Болгов Р. В., Санкт-Петербургский государственный университет
Григорьева И. А., Социологический институт РАН — филиал ФНИСЦ РАН
Захаров В. П., Санкт-Петербургский государственный университет
Кабанов Ю. А., НИУ «Высшая школа экономики» — Санкт-Петербург
Королева Н. Н., Российский государственный педагогический университет им. А. И. Герцена
Метелева А. С., Университет ИТМО (информационный менеджер конференции)
Микляева А. В., Российский государственный педагогический университет им. А. И. Герцена
Митягин С. А., Университет ИТМО
Низомутдинов Б.А., Университет ИТМО, НП ПРИОР Северо-Запад
Парыгин Д. С., Волгоградский государственный технический университет
Перов В. Ю., Санкт-Петербургский государственный университет
Толстикова И.И., Университет ИТМО, Социологический институт РАН — филиал ФНИСЦ РАН
Чижик А. В., Санкт-Петербургский государственный университет, Университет ИТМО
Чугунов А. В., Университет ИТМО, НП ПРИОР Северо-Запад (ученый секретарь конференции)

От редколлегии

В последние годы автоматизированная обработка текста и речи проникла во все сферы жизни, а разработка соответствующих систем и приложений стала одной из фундаментальных задач современного информационного общества. Вместе с этими тенденциями поменялась и структура компьютерной лингвистики как научной области, которая призвана обеспечить технологии и подходы, необходимые человечеству для человеко-компьютерной коммуникации и обработки текстовых данных. Определяя рамки компьютерной лингвистики в настоящее время, мы говорим об области науки, которая объединяет знания из лингвистики, математики и информатики. Вооружившись математическими закономерностями и возможностями статистики, она продолжает изучать фундаментальные закономерности языка, но вместе с тем даёт и много возможностей для прикладных исследований, так как изучает проблемы обработки естественного языка компьютерными методами и разрабатывает методы и алгоритмы для автоматического анализа, понимания и генерации текста.

Важность компьютерной лингвистики в современном мире трудно переоценить. Мы часто говорим: век информационных технологий окутал нас в цифровые данные. Это действительно так, и наряду с количественными данными огромную роль играют тексты — как поддающееся анализу воплощение речи и языка, которые являются базисом взаимодействия индивидов на макро- и микроуровнях. Они окружают нас повсюду: историки исследуют метрические книги, медики пытаются автоматизировать запись пациентов на первичный приём, юристы разрабатывают анализаторы документов, в социальных сетях каждую минуту появляется новый пост и комментарии к нему. Необходимость обработки и анализа больших объёмов текстов, а также развитие искусственного интеллекта как концепта, определяющего многие методологии взаимодействия и использования текстовых данных, продолжают развиваться с каждым днём. В этом процессе компьютерная лингвистика оказывается местом структурирования разрозненных исследований.

При этом, являясь одной из самых востребованных и динамически развивающихся наук, компьютерная лингвистика претерпевает значительные изменения: меняются её подходы и методы, это находит отражение и в новой терминологии, и в новом понимании места компьютерной лингвистики в системе наук. Один из таких обобщающих взглядов является включение компьютерной лингвистики в понятие «искусственный интеллект». Стоит также отметить, что использование лингвистических технологий стало элементом профессиональной деятельности самых разных специалистов. Поэтому ещё одна функция семинара, появившаяся в новом времени, — обеспечить и стимулировать диалог между лингвистами и специалистами, относящимися к самым разным направлениям научного знания, так как обмен идеями и попытка понять методологии друг друга даёт возможность переместить область междисциплинарных исследований на качественно новый уровень.

Статьи, публикуемые в данном сборнике, представляют собой изложение докладов русскоязычной секции семинара «Компьютерная лингвистика», который состоялся 28 июня 2023 года в рамках XXVI Международной объединённой конференции «Интернет и современное общество 2023». С полной программой семинара можно ознакомиться на сайте конференции по адресу ims.itmo.ru.

Темы, представленные авторами, имеют как теоретическое, так и прикладное значение и отражают широкий спектр исследований, а также многогранность задач в области автоматической обработки текста.

Это уже двенадцатый семинар по компьютерной лингвистике в составе данной конференции, первый состоялся в 2008 г. Считаем нелишним напомнить, что у истоков этого направления конференции стоял Валерий Шлёмович Рубашкин, являющийся классиком онтологической семантики и представителем междисциплинарного подхода к построению интеллектуальных информационных систем. Руководители семинара и

редакторы сборника надеются, что они продолжают дело Валерия Шлёмовича, один (В. Захаров) как классический компьютерный лингвист, другая (А. Чижик) как представитель нового, междисциплинарного, этапа в развитии компьютерной лингвистики, в рамках которого научная область является частью науки о данных (data science).

За эти годы семинар стал важной платформой для представления и обсуждения современного состояния и результатов в области компьютерных технологий и автоматизированной обработки естественного языка.

Итак, краткий обзор докладов.

Дмитрий Мельничук и Анастасия Носкина, представители Саратовского национального исследовательского государственного университета имени Н. Г. Чернышевского, описали в своей статье «Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке» результаты сравнения основных NLP-моделей (mBART, T5 и GPT-3), которые в своей основе имеют архитектуру трансформеров. Модели были использованы для суммаризации научных статей на русском языке. Авторы приводят в статье описание эксперимента, характеристики данных, использованных для тестов, а также значения метрик.

В статье «Функционирование устойчивой модели <X от слова Y> в современном интернет-пространстве» Юлия Локалина, Санкт-Петербургский государственный университет, рассматривает особенности функционирования устойчивой языковой модели, которая способствует появлению в языке таких конструкций, как <от слова совсем>, <от слова вообще> и т.п. В результате проведённого исследования были выявлены пунктуационные, синтаксические и др. особенности употребления конструкции. На основе собранных данных была построена шкала интенсификации, которая собрала все возможные варианты использования конструкции в интернете.

Команда разработчиков из Университета ИТМО (М. Егоров, Д. Погребной, М. Якубова, А. Кривошапкина, А. Чижик) представила в статье «Поддержка модели превентивной медицины: модуль обработки естественного языка для дистанционного взаимодействия "клиника-пациент"» созданный ими языковой модуль для выделения симптомов и классификации болезней. Модуль создан на языке Python. Авторы описывают особенности обучающих данных (использовалась прямая речь пациентов, описывающих свои диагнозы), сравнивают успешность методов векторизации текстов. В статье дан обзор архитектуры модуля и сравнение метрик качества при различных вариантах обучения моделей, включённых в состав системы.

Екатерина Татур и Екатерина Клименко, представляющие Санкт-Петербургский государственный университет, в статье «Выявление скрытых закономерностей в реакции общества на бренд: анализ привлекательности названия методами машинного обучения» описывают результаты работы над прикладной областью анализа текстов. Учёные собрали набор данных с разметками успешности названий у брендов, которые предложили разметить респондентам (специалистам по пиару и рекламе). Далее авторы приводят результаты анализа данных: закономерности, которые удалось найти. Данная работа крайне интересна с точки зрения применения классических методов анализа текста с целью дальнейшего использования в рамках автоматизации деятельности пиар-специалистов.

Анна Чижик (Университет ИТМО и Санкт-Петербургский государственный университет) в статье «Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей» представила результаты анализа успешности моделей векторизации применительно к коротким текстам. В работе сравниваются три актуальных на данный момент подхода к созданию векторного представления: учёт веса слова в документе (TF-IDF), использование дистрибутивной семантики при создании векторов слов (Word2Vec) и векторизация целых предложений (Laser). В статье приводятся размышления над метриками качества, описываются данные, использованные для тестирования моделей.

Леонард Ходоровский представляет в статье «Сведения, информация и информационная коммуникация» свои размышления на тему заявленных категорий. Учёный анализирует соотношение понятий «сведения» и «информация», являющихся элементами процесса «информационная коммуникация». В работе даётся определение понятия «сведения о свойствах сущности» как обозначение неоднородности реального или вымышленного мира и неравномерности протекания процессов в этом мире, характеризующих эту сущность. Идейно эта работа связана с ещё одним важным докладом, прозвучавшем на семинаре, — «К конструктивному определению свойств информации», который был представлен Николаем Максимовым и Александром Лебедевым (Национальный исследовательский ядерный университет «МИФИ»). Их исследование посвящено анализу и обоснованию свойств информации. Авторы, основываясь на понимании информации, как особой формы материи, вводят фундаментальные, прагматические и атрибутивные свойства.

Статья «К идентификации ситуативных ролей сущностей в контексте задачи семантического информационного поиска», представленная коллективом Национального исследовательского ядерного университета «МИФИ» (А. Гаврилкина, Н. В. Максимов, О. Голицына), раскрывает подходы к идентификации семантических ролей сущностей в контексте задачи семантического информационного поиска. Авторы дают анализ различных определений ролей. Также в статье предлагается подход к назначению ролей в соответствии с классами онтологии отношений, построенной на основе расширенной функциональной модели.

В статье Анны Быковой (СПбГУ) освещаются возможности применения методов машинного и глубокого обучения к оценке эмоциональной окраски текста постов, содержащих эмодзи, из социальной сети «ВКонтакте» (Оценка эмоциональной окраски постов социальной сети «ВКонтакте», включающих эмодзи, методами машинного и глубокого обучения). Автором описывается несбалансированный набор данных с текстом постов, размеченный по 15 классам, учитывающим эмоциональную и тональную составляющие в тексте. На полученном наборе данных она проводила эксперименты с использованием 6 методов классического машинного обучения, их ансамблей с мажоритарным и мягким голосованием и тремя нейросетевыми методами.

Семинар продемонстрировал, что динамика развития компьютерной лингвистики ставит все новые и новые проблемы, намечает новые рубежи, связанные с её междисциплинарностью, выходом за пределы собственно лингвистики в широкую область информационных технологий и искусственного интеллекта. Можно смело утверждать, что, в конечном счёте, компьютерная лингвистика играет важную роль в современном обществе, помогая нам преодолевать языковые барьеры, обрабатывать и анализировать огромные объёмы текстовой информации и развивать искусственный интеллект. Она улучшает наши возможности в общении, обмене информацией и нахождении нужных данных, что делает её одной из самых важных областей развития наших технологий.

Редакторы сборника
В. П. Захаров, А. В. Чижик

Оценка эмоциональной окраски постов социальной сети «ВКонтакте», включающих эмодзи, методами машинного и глубокого обучения

А. П. Быкова

Санкт-Петербургский государственный университет

st098553@student.spbu.ru

Аннотация

В данной работе исследуются возможности применения методов машинного и глубокого обучения к оценке эмоциональной окраски текста постов, содержащих эмо́дзи, из социальной сети «ВКонтакте». Описывается несбалансированный набор данных с текстом постов, размеченный по 15 классам, учитывающим эмоциональную и тональную составляющие в тексте. На полученном наборе данных проводятся эксперименты с использованием 6 методов классического машинного обучения, их ансамблей с мажоритарным и мягким голосованием и 3 нейросетевых методов. Лучший результат по метрикам качества классификации получился для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмо́дзи: F1-мера macro = 69.70%, F1-мера weighted = 82.06% и для рекуррентной нейросети GRU на 15 эпохах обучения: F1-мера macro = 48.77%, F1-мера weighted = 83.74%.

Ключевые слова: анализ эмоций, эмоциональная окраска текста, эмо́дзи, машинное обучение, нейронные сети

Библиографическая ссылка: Быкова А. П. Оценка эмоциональной окраски постов социальной сети «ВКонтакте», включающих эмо́дзи, методами машинного и глубокого обучения // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 12–20. DOI: 10.17586/2541-9781-2024-7-12–20

1. Введение

Уже не одно десятилетие исследователи достаточно много внимания уделяют анализу тональности текста и речи. Результаты анализа эмоциональной окраски текстов имеют множество практических применений, например, в различных приложениях при работе с клиентами, в политологии при работе с политическими окрашенными текстами, здравоохранении. Изучается потенциал анализа эмоций для выявления и предотвращения различных форм онлайн-злоупотреблений, например, запугивания пользователей. Кроме того, растёт интерес к тому, как эмоции передаются в разных языках и культурах, и как это может повлиять на оценку эмоциональной окраски различной информации [1].

Оценка эмоциональной окраски текста может быть полезна во многих областях, например, для того, чтобы понять какое настроение выражено в тексте. Эта информация может использоваться для анализа мнений, анализа отзывов клиентов, мониторинга социальных сетей. Понимая эмоции, выраженные в тексте, организации могут лучше учитывать потребности и предпочтения своих клиентов. Понимание эмоций также можно использовать в личном общении, чтобы оценить настроение человека и отреагировать

соответствующим образом. В данном исследовании оценка эмоциональной окраски текста постов в социальной сети «ВКонтакте» проводится методами машинного и глубокого обучения. Для разметки постов использовалась автоматическая разметка на основании встречающихся в этих постах эмóдзи.

Эмóдзи — это цифровые изображения или значки, которые используются в текстовых сообщениях в различных социальных сетях, в том числе «ВКонтакте». Язык эмóдзи своего рода графический язык, где вместо слов используются сочетания картинок. Впервые эмóдзи появились в Японии и распространились по всему миру. В настоящее время использование эмóдзи достаточно популярно и доступно в самых разных стилях и дизайнах. Популярность эмóдзи обусловлена тем, что они могут передавать эмоции и добавлять контекст к текстовому общению. В некоторых случаях эмóдзи помогают преодолевать языковые барьеры и делают общение более доступным среди людей, которые владеют разными языками.

2. Подходы к анализу эмоциональной окраски текста

Анализ тональности текста — одно из направлений в компьютерной лингвистике, в рамках которого решается задача выявления мнения автора текста по поводу того, что обсуждается в тексте.

Тональность текста можно рассматривать как с точки зрения автора текста, так и с точки зрения того, кто читает и воспринимает этот текст. Поскольку в данном исследовании эмóдзи являются маркером для разметки, а эмóдзи проставляет сам автор текста, то в этом исследовании тональность и эмоциональная окраска текста рассматривается с точки зрения автора этого текста.

В целом, анализ эмоциональной окраски текста подразумевает собой применение методов, с помощью которых можно определить, к какому классу относится тот или иной текст. В основном используются алгоритмы на основе словарей и правил [2; 3] и методы на основе машинного обучения. Также существуют комбинированные методы, в которых словари оценочной лексики являются компонентом модели машинного обучения [4].

Для многих задач автоматической обработки текста необходимы специально размеченные текстовые данные, например, для автоматического распознавания в тексте иронии или сарказма [5].

Большую популярность в задачах анализа тональности приобрели методы машинного обучения. С начала 2000-х годов широко применяются классические методы машинного обучения, такие как логистическая регрессия, метод опорных векторов, наивный байесовский классификатор [6]. Также широко применяются классические нейронные сети, например, рекуррентные нейронные сети и свёрточные нейронные сети [7]. В 2019 году появились новые подходы к анализу текста на основе нейросетевой архитектуры трансформер, такие как модель BERT [8]. Использование архитектуры серии BERT для различных задач автоматической обработки текста привело к росту качества решений этих задач, в том числе и в задачах анализа тональности.

Первоначально модель BERT обучалась на многоязычных текстовых данных, затем в ряде исследований было выявлено, что дообучение BERT на данных конкретного естественного языка может дать лучшие результаты решения задач для этого языка. Так, например, в работе [9] исследователи описывают модель RuBERT, которая дообучена на модели BERT для русского языка.

3. Сбор и разметка данных

Набор данных создавался самостоятельно из постов социальной сети «ВКонтакте». Данные взяты из 100 наиболее популярных сообществ «ВКонтакте» на 5 февраля 2023 года.

Статистика по самым популярным сообществам взята с сайта «TOPPOST». Выбор постов из социальной сети «ВКонтакте» в качестве материала исследования обусловлен тем, что данная социальная сеть является популярной платформой, которой пользуются русскоязычные пользователи. В постах пользователи выражают собственное мнение и открыто взаимодействуют посредством различных реакций (лайки, комментарии, репосты). В социальной сети чаще происходит неформальное эмоционально окрашенное общение, поэтому текст постов можно использовать для оценки эмоциональной окраски текста.

API (Application Programming Interface) «ВКонтакте» представлен в открытом доступе, с помощью открытых методов был написан скрипт для скачивания текста постов.

В качестве маркеров для разметки текста использовались эмóдзи, которые встречаются в постах. Разметка на основе эмóдзи является ограничением данного исследования, поскольку такой разметки может быть недостаточно для точной классификации текста по эмоциям и тональности, особенно в том случае, если эмóдзи использовались авторами постов неоднозначно. Для пояснения значений эмóдзи и их классификации использовались карточки с описанием с сайта «Смайлики Эмóдзи».

Собран словарь эмóдзи, которые встречаются в скачанных постах, состоящий из 146 эмóдзи, входящие в тематическую группу Smileys & Emotion.

Полученные 146 эмóдзи распределены по классам. В этих классах учитывается эмоциональная составляющая и тональная составляющая, поскольку однозначно категоризировать эмоции достаточно сложно, например, для такой эмоции, как удивление, может присутствовать как положительная, так и отрицательная тональность (см. табл. 1).

Таблица 1. Классы эмоций и тональности

№	Эмоция	Настроение (тональность)
1	улыбка (smile)	позитивное/негативное (positive/negative)
2	нет эмоции (no_emotion)	нейтральное/скептическое (neutral/skeptical)
3	удовольствие	позитивное (positive)
4	нет эмоции (no_emotion)	позитивное/негативное (positive/negative)
5	грусть (sadness)	негативное (negative)
6	страх (feat)	негативное (negative)
7	стыд (shame)	негативное (negative)
8	гнев (anger)	негативное (negative)
9	отвращение (disgust)	негативное (negative)
10	удивление (surprise)	позитивное/негативное (positive/negative)
11	отвращение (disgust)	нейтральное/скептическое (neutral/skeptical)
12	удивление (surprise)	негативное (negative)
13	нет эмоции (no_emotion)	негативное (negative)
14	грусть (sadness)	позитивное/негативное (positive/negative)
15	испуг (fear)	позитивное/негативное (positive/negative)

Самое большое количество эмóдзи 26 из 146 относится к классу joy positive (удовольствие позитивное настроение).

Для оценки эмоциональной окраски проведена автоматическая разметка данных на основе использованных в тексте эмóдзи. Для обучения и оценки алгоритмов машинного обучения были выбраны посты с 1 эмóдзи и длиной поста не более 11 токенов вместе с эмóдзи, получилось 9220 постов. В дальнейшем планируется исследовать тексты с другими параметрами по количеству эмóдзи и количеству токенов в посте.

Среди выбранных данных больше всего постов с эмóдзи из класса smile positive/negative (улыбка позитивное/негативное настроение), полученный набор данных является несбалансированным. Для эффективной работы с данными необходима их предобработка. Для этого из текста постов удалены id пользователей и групп, текст переведён в нижний регистр. Полученные 9220 постов автоматически размечены по 15 выделенным классам.

4. Эксперименты с моделями машинного и глубокого обучения

Выбор метода для анализа эмоциональной окраски текста зависит от требований решаемой задачи и характера набора данных. Для того, чтобы узнать, какой метод больше всего подходит для данного исследования, был проведён ряд экспериментов.

Тестовая выборка данных составляла 20% из общего числа постов. Для оценки эмоциональной окраски размеченных постов использовались классические методы машинного обучения из пакета scikit-learn. Для предобработки постов использовался лемматизатор pymorphy2.

Эксперименты проводились для текста с пунктуацией и с эмо́дзи, для текста без пунктуации и без эмо́дзи, для лемматизированного с помощью pymorphy2 текста с пунктуацией и с эмо́дзи.

Использовались представления слов в виде мешка слов (Bag of Words) [10], предобученные плотные векторные представления слов для русского языка из библиотеки Navec и Word2Vec [11] для классических методов машинного обучения, таких как, наивный байесовский классификатор (GaussianNB), логистическая регрессия (Logistic Regression), метод опорных векторов (SVM), градиентный бустинг (Gradient Boosting), случайный лес (Random Forest), классификатор дерева решений (DecisionTreeClassifier). Также проведены эксперименты с использованием ансамблей классификаторов с мажоритарным и мягким голосованием с помощью VotingClassifier.

Для экспериментов использовались нейросетевые модели: одномерная свёрточная нейросеть CNN, рекуррентная нейросеть LSTM (Long Short-Term Memory) и рекуррентная нейросеть GRU (Gated Recurrent Units).

5. Результаты оценки эмоциональной окраски текста постов

Для оценки качества классификации использовались метрики: F1-мера по макроусреднению (macro) и F1-мера по взвешенному усреднению (weighted). Выбор данных метрик для оценки эмоциональной окраски текста постов обусловлен тем, что полученный набор данных не является сбалансированным.

По F1-мере лучший результат получился для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмо́дзи, F1-мера macro равна 69.70%, F1-мера weighted равна 82.06%. Полученные результаты представлены в таблице 2 (лучшие результаты выделены жирным шрифтом).

Таблица 2. Результаты оценки эмоциональной окраски текста для классических методов машинного обучения

Модель	F1 macro, %	F1 weighted, %
1	2	3
BoW + Logistic Regression (текст с пунктуацией и с эмо́дзи)	51.74	78.97
BoW + SVC (текст с пунктуацией и с эмо́дзи)	40.26	72.21
BoW + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	65.10	80.54
BoW + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	63.79	79.45
BoW + GaussianNB (текст с пунктуацией и с эмо́дзи)	28.49	62.17
BoW + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	65.27	81.48
BoW + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	64.64	81.34
BoW + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	67.99	82.02
Navec + Logistic Regression (текст с пунктуацией и с эмо́дзи)	11.31	49.51
Navec + SVC (текст с пунктуацией и с эмо́дзи)	6.78	50.25
Navec + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	11.02	51.75

Продолжение таблицы 2

1	2	3
Navec + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	10.21	46.40
Navec + GaussianNB (текст с пунктуацией и с эмо́дзи)	1.15	4.34
Navec + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	8.19	49.25
Navec + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	10.50	51.68
Navec + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	11.02	52.07
Word2Vec + Logistic Regression (текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + SVC (текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	23.64	62.53
Word2Vec + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	15.22	54.54
Word2Vec + GaussianNB (текст с пунктуацией и с эмо́дзи)	1.63	7.60
Word2Vec + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	15.55	57.24
Word2Vec + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	11.76	54.77
Word2Vec + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	13.21	57.14
BoW + Logistic Regression (текст без пунктуации и без эмо́дзи)	8.33	52.19
BoW + SVC (текст без пунктуации и без эмо́дзи)	6.99	50.60
BoW + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	15.36	53.06
BoW + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	13.42	50.57
BoW + GaussianNB (текст без пунктуации и без эмо́дзи)	10.86	39.11
BoW + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	12.58	51.35
BoW + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	14.38	52.34
BoW + VotingClassifier (soft) (текст без пунктуации и без эмо́дзи)	14.80	53.75
Navec + Logistic Regression (текст без пунктуации и без эмо́дзи)	12.28	50.47
Navec + SVC (текст без пунктуации и без эмо́дзи)	7.67	50.36
Navec + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	10.57	51.80
Navec + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	10.63	44.41
Navec + GaussianNB (текст без пунктуации и без эмо́дзи)	1.05	1.05
Navec + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	9.12	49.62
Navec + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	11.41	51.92
Navec + VotingClassifier (soft) (текст без пунктуации и без эмо́дзи)	11.89	52.67
Word2Vec + Logistic Regression (текст без пунктуации и без эмо́дзи)	5.17	48.99
Word2Vec + SVC (текст без пунктуации и без эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	9.71	51.74
Word2Vec + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	9.83	45.48
Word2Vec + GaussianNB (текст без пунктуации и без эмо́дзи)	0.03	0.01
Word2Vec + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	8.58	49.09
Word2Vec + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	8.78	50.66
Word2Vec + VotingClassifier (soft) (текст без пунктуации и без	8.91	50.82
BoW + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	51.49	78.64

Продолжение таблицы 2

1	2	3
BoW + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	40.16	72.11
BoW + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	66.77	81.43
BoW + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	66.30	79.28
BoW + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	27.26	60.74
BoW + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	67.01	81.74
BoW + VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	63.35	81.32
BoW + VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	**69.70**	**82.06**
Navec + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	11.39	49.15
Navec + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	6.59	50.04
Navec + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	11.14	51.98
Navec + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	9.89	46.18
Navec + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	1.14	4.83
Navec + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	7.35	49.20
Navec + VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	10.50	51.60
Navec + VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	10.91	51.73
Word2Vec + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	25.85	64.19
Word2Vec + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	18.91	55.79
Word2Vec + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	1.40	6.56
Word2Vec + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	18.77	56.97
Word2Vec +VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	13.93	55.47
Word2Vec +VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	18.09	59.61

Также для экспериментов использовались нейросетевые модели: одномерная свёрточная нейросеть CNN, рекуррентная нейросеть LSTM (Long Short-Term Memory) и рекуррентная нейросеть GRU (Gated Recurrent Units).

Лучший результат среди использованных нейросетевых моделей показала рекуррентная нейросеть GRU на 15 эпохах обучения: F1-мера macro равна 48.77%, F1-мера weighted равна 83.74%. Полученные результаты представлены в таблице 3 (лучшие результаты выделены

жирным шрифтом).

Таблица 3. Результаты оценки эмоциональной окраски текста для нейросетевых методов

Модель	F1 macro, %	F1 weighted, %
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=5)	17.42	71.54
Рекуррентная нейросеть LSTM (токенизатор Keras, optimizer='adam', epochs=5)	11.85	62.66
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=5)	28.29	78.42
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=10)	40.20	81.86
Рекуррентная нейросеть LSTM (токенизатор Keras, optimizer='adam', epochs=10)	23.44	78.39
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=10)	43.15	83.85
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=15)	27.77	78.96
LSTM (токенизатор Keras, optimizer='adam', epochs=15)	29.34	79.81
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=15)	**48.77**	**83.74**

Получили, что в случае макроусреднения, т. е. когда всем классам даётся одинаковый вес, независимо от их количества в наборе данных, лучший результат F1-мера macro = 69.70% для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмóдзи. В случае же взвешенного усреднения, т. е. когда вес классам даётся согласно количеству объектов в этих классах, лучший результат F1-мера weighted = 83.74% для модели рекуррентной нейросети GRU на 15 эпохах обучения.

В дальнейшем планируется сравнить полученные результаты по метрикам качества классификации F1-меры с результатами работы модели RuBERT, которая позволяет работать с текстами на русском языке и имеет качественные распределённые векторные вложения (embeddings) текстов.

6. Заключение

В данной статье представлена оценка эмоциональной окраски постов из социальной сети «ВКонтакте», описан процесс получения, обработки и использования полученного набора данных. Приводятся результаты экспериментов с использованием методов машинного и глубокого обучения с оценкой работы методов по метрикам качества классификации. По оценке качества классификации текста постов лучший результат по метрике F1-мера macro = 69.70% показала модель BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмóдзи. Лучший результат по метрике качества классификации F1-мера weighted получен для модели рекуррентной нейросети GRU F1-мера weighted = 83.74%.

Поскольку эксперты не размечали полученные данные, а использовалась автоматическая разметка постов на основании, встречающихся в этих постах эмóдзи, в дальнейшем планируется провести экспертную оценку полученной автоматической разметки постов по выделенным классам.

Также планируется провести эксперименты на сбалансированных данных. В будущем можно продолжить исследование с использованием текстов с другими параметрами по количеству эмóдзи и токенов в тексте.

Литература

[1] Calvo R. A., D'Mello S. Affect detection: An interdisciplinary review of models, methods and their applications // IEEE Transactions on affective computing. 2010. Vol. 1 (1). P. 18–37.
[2] Кузнецова Е. С., Лукашевич Н. В., Четверкин И. И. Тестирование правил для системы анализа тональности // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции Диалог 2013. М.: Изд-во РГГУ, 2013. Вып. 12. Т. 2. С. 71–80.
[3] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1171–1176.
[4] Kiritchenko S., Zhu X., Mohammad S. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. 2014. Vol. 50. P. 723–762.
[5] Joshi A., Bhattacharyya P., Carman M. Automatic sarcasm detection: A survey // ACM Computing Surveys (CSUR). 2017. Vol. 50 (5). P. 1–22.
[6] Pang B., Lee L., Vaithyanathan S. Thumb up? Sentiment Classification using Machine Learning Techniques // Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2002. 2002. P. 79–86.
[7] Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey // Wiley Reviews: Data Mining and Knowledge Discovery. 2018. Vol. 8 (4).
[8] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.
[9] Куратов Ю., Архипов М. Адаптация глубоких двунаправленный многоязычных моделей на основе архитектуры Transformer для русского языка // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции Диалог 2019. М.: Изд-во РГГУ, 2019. Вып. 19 (25). С. 333–339.
[10] HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation // PloS one. 2020. Vol. 15 (5).
[11] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th international conference on neural information processing systems. 2013. Vol. 2. P. 3111–3119.

Evaluation of Emotionality of Posts with Emojis in the VKontakte Social Network Using Machine and Deep Learning Methods

Anna P. Bykova

Saint Petersburg State University

In this paper possibilities of applying machine and deep learning methods to emotionality evaluation of posts text with emojis from the VKontakte social network are investigated. An unbalanced data set with posts text is described, the data set is annotated by 15 classes. In these classes emotional and tonal components of text are taken into account. Experiments are conducted on the obtained data set with using 6 classical machine learning methods, their ensembles with hard and soft voting, and 3 neural network methods. The best result by classification quality metrics was obtained for the Bag of words + VotingClassifier ensemble method with soft voting on lemmatized text with punctuation and emoji: F1-macro measure = 69.70%, F1-weighted measure = 82.06% and for the recurrent neural network GRU on 15 epochs of training: F1-measure macro = 48.77%, F1-measure weighted = 83.74%.

Keywords: emotion analysis, evaluation of emotionality in text, emoji, machine learning, neural networks

Reference for citation: Bykova A.P. Evaluation of Emotionality of Posts with Emojis in the VKontakte Social Network Using Machine and Deep Learning Methods // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 12–20. DOI: 10.17586/2541-9781-2024-7-12–20

Reference

[1] Calvo R. A., D'Mello S. Affect detection: An interdisciplinary review of models, methods and their applications // IEEE Transactions on affective computing. 2010. Vol. 1 (1). P. 18–37.

[2] Kuznecova E. S., Lukashevich N. V., Chetverkin I.I. Testirovanie pravil dlya sistemy analiza tonal'nosti // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam mezhdunarodnoj konferencii Dialog 2013. M.: Izd-vo RGGU, 2013. Vyp. 12. T. 2. S. 71–80. (in Russian)[3] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1171–1176.

[4] Kiritchenko S., Zhu X., Mohammad S. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. 2014. Vol. 50. P. 723–762.

[5] Joshi A., Bhattacharyya P., Carman M. Automatic sarcasm detection: A survey // ACM Computing Surveys (CSUR). 2017. Vol. 50 (5). P. 1–22.

[6] Pang B., Lee L., Vaithyanathan S. Thumb up? Sentiment Classification using Machine Learning Techniques // Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2002. 2002. P. 79–86.

[7] Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey // Wiley Reviews: Data Mining and Knowledge Discovery. 2018. Vol. 8 (4).

[8] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.

[9] Kuratov Yu., Arhipov M. Adaptaciya glubokih dvunapravlennyj mnogoyazychnyh modelej na osnove arhitektury Transformer dlya russkogo yazyka // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam mezhdunarodnoj konferencii Dialog 2019. M.: Izd-vo RGGU, 2019. Vyp. 19 (25). S. 333–339. (in Russian)

[10] HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation // PloS one. 2020. Vol. 15 (5). DOI: 10.1371/journal.pone.0232525.

[11] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th international conference on neural information processing systems. 2013. Vol. 2. P. 3111–3119.

К идентификации ситуативных ролей сущностей в контексте задачи семантического информационного поиска

А. С. Гаврилкина, Н. В. Максимов, О. Л. Голицына

Национальный исследовательский ядерный университет «МИФИ»

asgavrilkina@yandex.ru, nv-maks@yandex.ru, olgolitsina@yandex.ru

Аннотация

В статье рассматриваются подходы к идентификации семантических ролей сущностей в контексте задачи семантического информационного поиска. Проанализированы различные определения ролей. Предложен подход к назначению ролей в соответствии с классами онтологии отношений, построенной на основе расширенной функциональной модели. Проведён эксперимент по оценке качества автоматической идентификации ролей Причина, Основание, Условие.

Ключевые слова: семантические роли, информационный поиск, автоматическая обработка текста, полнотекстовое индексирование

Библиографическая ссылка: Гаврилкина А. С., Максимов Н. В., Голицына О. Л. К идентификации ситуативных ролей сущностей в контексте задачи семантического информационного поиска // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 21–31. DOI: 10.17586/2541-9781-2024-7-21–31

1. Введение

Целесообразность использования указателей роли для повышения точности отбора документов при информационном поиске была рассмотрены и реализована ещё на заре становления автоматизированных информационно-поисковых систем (см., например, [1, 2]). Этот очевидно целесообразный подход, однако, не получил развития и не нашёл в последующем практического применения не в последнюю очередь вследствие неоднозначности подходов к унификации типологии связей и усложнения языка запросов, а, соответственно, возникновения трудностей в использовании. Косвенным, но в итоге определяющим фактором были сравнительно малые объёмы информации и слабые вычислительные возможности. Отметим, что «минимальной основой для построения исчисления семантики является иерархия классов сущностей, отношений и свойств, базирующаяся на общей системе характеристических признаков. Однако обзор существующих решений показывает, что нет строгой и всеобщей классификации, хотя есть ряд вполне самодостаточных решений» [3].

Развитие ИТ и ВТ обеспечило существенное увеличение возможностей по обработке информации, в частности, текстов документов, а онтологический подход позволил реализовать глубинный семантический поиск [4]. В этом случае поисковые образы представляют полномасштабную концептуальную модель объекта, отражая существо его функционирования — это не только понятийный состав (традиционные ключевые слова), но все именованные сущности и ситуативные отношения. Построение основывается на выделении сущностей и связей между ними при помощи шаблонов, связывающих сущности

и отношения с их вербальными формами представления [5]. Выражение имён сущностей и отношений на знаковом уровне позволяет индексировать элементарный факт, как триплет — последовательность знаков, в которой представлены не только имена, но и типы сущностей и отношений. Таким образом, могут быть построены как традиционные индексы (по ключевым словам), так и индексы, представляющие семантические связи. Наличие таких индексов позволяет в рамках традиционной теоретико-множественной модели информационного поиска (и средствами традиционного дескрипторного ИПЯ) реализовать отбор документов с учётом имманентных и ситуативных отношений между сущностями. ИПЯ при этом попадает в семантической классификации в класс языков, имеющих средства выражения (и различения) и имманентных, и ситуативных отношений.

При этом связи между сущностями (указывающие на взаимодействие/соотношение объектов действительности) могут быть идентифицированы как через отношения, которые выражаются преимущественно глагольными и предложными конструкциями, так и через роли, выделяемые на основе имён сущностей (то есть на понятийном уровне присутствует условное разделение на отношения и роли). Отметим, что понятие роли вводится и в языке онтологического моделирования Gellish стандарта ISO-15926 [6].

Таким образом, роли могут использоваться как параметр для поиска (квалификатор в условии отбора), позволяющий выделить множество имён сущностей с общими для связанных с определенным классом предикатов свойствами, или в качестве одного из параметров при поиске на графах — при построении отображения графа, например, на основе функциональной модели, или при задании аспектной проекции [4].

В настоящей работе будут рассмотрены различные определения ролей и описан подход к автоматической идентификации ролей сущностей в тексте в соответствии с классификацией отношений, построенной на основе расширенной функциональной модели.

2. К определениям понятия «роль»

Рассмотрим некоторые определения ролей.

Роль — это функциональное назначение, которое объект (нечто, anything) выполняет в данной ассоциации по отношению к связи, или наоборот, связь по отношению к объекту [7].

В [8] семантическая роль имени при предикате определяется как часть семантики предиката, которая отражает общие свойства участников определенных групп ситуаций.

В [9] семантические роли рассматриваются как классы, к которым сводятся лексические партиципанты. Предполагается, что таких классов должно быть небольшое обозримое число.

Исходя из приведённых определений можно сказать, что у сущности существует роль относительно предиката, ситуации или фрейма, которая отражает основные свойства для группы. Здесь семантический фрейм, согласно [10], представляет собой совокупность фактов, определяющих «характерные черты, атрибуты и функции денотата, а также его характерные взаимодействия с вещами, обязательно или типично с ним связанными». В лингвистике фрейм ассоциируется с определенным предикатом (а чаще несколькими близкими по смыслу предикатами), при котором выражаются участники фрейма (аргументы) [11].

В настоящей работе роль рассматривается как проекция отношения на объект, причём отношения могут быть как связями между взаимодействующими объектами, так и между объектами и обстоятельствами. То есть роль выступает как форма объективизации взаимосвязи. Это особенно заметно, например, в случае отношения «вход», когда «поглощение» входного объекта для обработки является скорее событием, чем действием, и его очевидно удобнее идентифицировать ролью «исходный материал» [12].

Сущность может быть соотнесена с несколькими ролями, так как количество ролей зависит непосредственно от количества отношений, в которых она состоит. Например, подлежащему, как главному члену предложения, являющемся наиболее очевидным

вариантом для выбора в пару «сущность — отношение» может быть назначено несколько ролей, так как оно связывается с обстоятельствами и дополнениями и таким образом получает роль для идентификации в каждом отношении. То есть роль представляет основные свойства сущности в контексте конкретного отношения. И наоборот, идентификатора роли должно быть достаточно, чтобы сказать, каким типом отношения связана сущность.

3. Типизация отношений и ролей

В последние десятилетия активно решается задача автоматической разметки ролей, которая заключается в автоматической идентификации участников ситуаций, обозначаемых предикатами (глаголами, существительными, прилагательными и т. д.), и разметка способа их выражения — вне зависимости от того, связаны ли обозначающие участников лексические единицы с предикатом синтаксически или нет. [11]

Определение ролевой структуры высказывания позволяет соотносить похожие по смыслу предложения независимо от их синтаксических структур. Ролевые структуры высказываний используются во многих прикладных задачах: в вопросно-ответном поиске, машинном переводе, оценке смысловой близости фрагментов текстов на естественном языке и др. [13].

Однако, учитывая многообразие предикатов в языке, в том числе со схожими в некоторых аспектах значениями, требуется определить признаки для обобщения.

С точки зрения [14] подходы к определению ролей можно разделить согласно уровню обобщения: при максимальном обобщении роли сводятся к двум — Actor и Undergoer, где роль Actor включает роли агенса, экспериенцера, инструмета и др., а Undergoer — роли пациенса, темы, приёмника и т.д.; с противоположной стороны спектра находятся предикатно-специфичные роли, то есть уникальные для каждого предиката; все наборы ролей с некоторым обобщением находятся в середине спектра.

К наборам ролей из середины спектра можно отнести предложенный Ч. Филмором классический набор ролей, который он считал универсальным для всех языков: агентив, объектив, датив, инструменталис, фактив, локатив [15].

При решении задачи семантической разметки ролей в большинстве случаев используются предикатно-специфичные роли, что связано с популярностью методов машинного обучения и ограниченностью размеченных корпусов для их применения. Наиболее известны изначально англоязычные аннотированные семантическими ролями корпуса текстов PropBank [16] и FrameNet [17], которые позже были сопоставлены и продублированы на других языках, первый из которых ориентирован на аннотацию глаголов, а второй — на заполнение фреймов (в которых предикаты могут быть сгруппированы в типовых ситуациях). Для русского языка был создан FrameBank, основанный на концепции фреймов, но всё же сосредоточенный на описании конкретных глаголов и их окружения (то есть анализируется возможная сочетаемость глагола с предлогами и назначаются роли участникам ситуации). Однако хотя FrameNet и FrameBank и допускают разные уровни обобщения, классификационные признаки для группировки фреймов и ролей чётко не обозначены.

Рассмотрим более подробно корпус FrameBank [11], который рассматривается создателями с одной стороны, как ресурс для тренировки систем искусственного интеллекта, с другой — как инструмент для изучения конструкционных свойств русской лексики.

FrameBank включает словарь лексических конструкций и корпус примеров их употребления. В шаблонах лексических конструкций указываются морфосинтаксические характеристики элементов конструкции, синтаксический ранг участника, роль (экспликация) участника, лексико-семантические ограничения на заполнение слота конструкции, статус участника: обязательный или факультативный, буква, маркирующая

участника в кратком паттерне. Ядро системы FrameBank составляют 2200 частотных русских глаголов. В корпусной части ресурса FrameBank представлено приблизительно по 100 примеров из НКРЯ на каждый глагол. В работах [13, 18] описывается использование FrameBank для автоматической разметки ролей с применением машинного обучения.

В FrameBank на основе ролей согласно [19] выделены 88 базовых экспликаций семантических ролей, которые разделены на следующие группы [11]:

- блок Агенса;
- блок Пациенса;
- блок Экспериенцера;
- блок Инструмента;
- блок Адресата;
- блок обстоятельственных характеристик (группы Места, Времени, Параметров, Признаков, Причины, Цели);
- группа посессивных ролей;
- группа Источников и Ресурсов.

В настоящей работе к классификации отношений применяется подход на основе расширенной функциональной модели. Акт процесса (реализация операции функциональной обработки) целенаправленной и управляемой деятельности (познавательной, производственной и т.д.) рассматривается с точки зрения теории систем и может быть представлен средствами функционального моделирования [12].

Для типизации ситуативных (функциональных) отношений (и соотнесённых с ними ролей) используется ориентированная на научно-техническую сферу онтология отношений, основанная на трехуровневой иерархической классификации. Первый уровень отражает соотношение реальность/модель, второй — комбинации соотношений отдельного (часть) и агрегатного (целое), третий уровень построен по признаку формы проявления отношения — действие-ориентированные, объект-ориентированные и результат-ориентированные. Листьями иерархического дерева являются классы отношений, обладающие комбинацией свойств верхних уровней. Каждому классу поставлено в соответствие множество лингвистических конструкций, которые выражают его семантику. Классы нижнего уровня открыты для пополнения и содержимое их может зависеть от вида/жанра текстового массива, подлежащего обработке. Для удобства восприятия и эксплуатации онтология отношений представлена в виде иерархически упорядоченной таксономии. [5]

На верхнем уровне классификации находятся 27 классов отношений, среди которых Создание, Воздействие, Изменение, Идентификация, Отождествление, Зависимость-связь, Размещение, Разъединение и т.д. [3]. Каждый класс отношений соотносится с набором ролей, назначаемых его подклассам, например, сущностям, связанным отношениями из класса Зависимость-связь, могут быть назначены следующие роли: Вход, Выход, Цель, Результат, Причина, Следствие, Ресурс, Ограничение, Средство, Инструмент, Управление, Условие, Помеха, Основание, Назначение.

Так как в данной работе роли рассматриваются как свойства связи и соотнесены с классами отношений, то для идентификации ролей используются классифицированные лингвистические конструкции отношений (преимущественно глаголы, отглагольные части речи и предлоги), при помощи которых отношение может быть представлено в тексте, их морфологические характеристики (возвратность, наклонение глаголов, вид причастий), а также место сущности в триплете. Уровни обобщения ролей соответствуют уровням таксономии отношений, но при этом есть основания полагать, что они могут группироваться иначе, чем классы отношений.

Кроме того, следует учитывать и взаимозависимость сущностей и отношений, т. е. возможность совершения действия того или иного характера над объектами определённых классов. Например, некоторые действия можно выполнить только с одушевлёнными объектами или только с абстрактными и т.д. При этом часто встречается и перенос

буквальных значений с одушевлённого на неодушевлённое, например, для «дышать», «есть», «жить».

Как отмечалось выше для идентификации ролей использовалась онтология отношений, ориентированная на научно-техническую сферу. Поэтому для сравнительной оценки адекватности идентификации ролей сопоставим результаты с получаемыми на основе подхода FrameBank:

1) в тексте *Земля представляет собой сфероид* у сущности *Земля* определена роль Идентифицируемое, связанная с классом отношений Идентификация, в то время как в FrameBank — это Статус в блоке Инструмент, который является частью Блока обстоятельственных характеристик;
2) в предложении *Билет стоил тысячу рублей* у сущности «тысяча рублей» определена роль Отождествляемое, связанная с классом отношений Отождествление, в FrameBank — Цена в блоке Параметры, который является частью Блока обстоятельственных характеристик;
3) в результате обработки предложения *Он разломал кусок хлеба на 4 части* будет построено два триплета *<Он> разломал <кусок хлеба>, <кусок хлеба> разломал на <4 части>*, где для сущности *кусок хлеба* определена роль Разъединяемое, для сущности *4 части* — Части разъединённого, соотнесённые в с классом отношений Разъединение, во FrameBank сущности *4 части* назначается роль Результат в блоке Пациенс.

Таким образом, можно сделать вывод, что в естественно-научной классификации отношений некоторые классы более детализированы (например, класс отношений Изменение включает подклассы Изменение состояния, Изменение формы, Изменение состава и пр. и соответствующие роли) или ключевой признак сущности, по которому определена роль, отличается, что связано с задачей и предметной областью. Для разметки научно-технических текстов FrameBank избыточен, кроме того, в словаре лексических конструкций не были обнаружены некоторые характерные для научно-технических текстов производные предлоги (в частности «с целью», «по причине», «в соответствии с», «в условиях», «в связи с», «вследствие»). Однако для обработки художественных текстов множество лингвистических конструкций отношений классификации может быть пополнено предикатами из FrameBank.

4. Эксперимент по оценке качества идентификации ролей

Для оценки качества идентификации ролей использовались 19 текстов общим объёмом 150 Кб (19 426 слов), после обработки которых с помощью опытного образца сервиса визуального онтологического анализа научно-технических текстов [20] было получено 3619 имён сущностей (уникальных в рамках каждого текста) и 4234 триплетов с отношениями из текста.

Для экспертной оценки были выбраны роли Причина, Основание, Условие, связанные с подклассами класса Зависимость-связь: Быть причиной, Быть основанием, Быть условием соответственно. В классификации лингвистических конструкций отношений для идентификации роли Причина предназначено 20 лингвистических конструкций отношений, Основание — 22, Условие — 16, из которых использовалось 12, 7 и 15, соответственно. В эксперименте идентифицировалась роль только первой сущности. В результате для 40 сущностей была определена роль Причина, для 62 — роль Основание, для 59 — роль Условие. Применённый метод автоматической обработки текста допускает формирование некорректных триплетов, однако при оценке были рассмотрены только триплеты с корректно определёнными отношением и сущностью, которой назначена роль.

Таблица 1. Количество и процентное соотношение сущностей с корректно и некорректно определёнными ролями Причина, Основание, Условие

	Причина	**Основание**	**Условие**
Количество имён сущностей, у которых корректно определена роль	40 (100 %)	48 (77%)	58 (98%)
Количество имён сущностей, у которых некорректно определена роль	0	14 (23%)	1 (2%)

Были проанализированы случаи некорректного определения ролей и выявлены следующие причины:

- многозначность предлога «по», по которому идентифицировано 93% сущностей с ролью Основание: в сочетаниях «по мнению экспертов», «по ветру», «по плану» и других случаях, когда предлог «по» может быть заменён «согласно» или «в соответствии с», назначение роли считалось корректным в отличие от случаев «завод по обогащению урана», «программе по развитию процессов памяти» и прочих, где речь идёт о теме или предмете деятельности;
- возможность формирования триплетов, в которых некорректно связаны имена сущностей, которые в качестве членов предложения идентифицируются как дополнения.

Первая проблема, может быть, в некоторой мере решена пополнением множества классифицированных лингвистических конструкций отношений распространёнными сочетаниями с предлогами (преимущественно для повышения точности идентификации при обработке текстов художественного стиля), вторая — введением ограничений на падежи главных слов в словосочетаниях, связываемых определёнными отношениями.

В таблицах 2-4 приведены примеры триплетов с корректно определённой ролью сущности и фрагменты текста, в результате обработки которых они составлены. Имена сущностей и отношений в триплетах приведены в нормальной форме (пословная нормализация). Код указывает на класс отношения в таксономии отношений[1]. В правой колонке курсивом выделена лингвистическая конструкция отношения, указывающая на роль сущности, полужирным — сущность, которой присваивается роль.

Таблица 2. Примеры триплетов с отношением из класса Быть причиной и сущностями с ролью Причина

Триплет, включающий сущность с ролью Причина	**Фрагмент текста**
<поддержание межличностный отношение> из-за 12315[из-за] <**особенность мировосприятия**>	Чрезмерное пребывание за компьютером в ущерб всему остальному, трудности в установлении и поддержании межличностных отношений *из-за* **особенностей мировосприятия** и познавательных процессов, обусловленных взаимодействием с виртуальной реальностью компьютерных технологий, приводят к упрощению системы отношений «человек - человек» у подростков, включённых в компьютеризированную деятельность

[1] Упоминающиеся в работе коды соотносятся с классами отношений следующим образом: 12315 — Быть причиной, 1231E — Быть основанием, 1231C — Быть условием, 1121 — Передача, 11212 — Излучение, 1133 — Явление, факт, 1211 — Изменение, 12113 — Изменение состояния, 1213 — Усложнение, 1232 — Зависимость-соотношение, 1322 — Определение, 2111 — Присоединение, 22212 — Локативность в пространстве, 2223 — Обладание.

Продолжение таблицы 2

Триплет, включающий сущность с ролью Причина	**Фрагмент текста**
<характеристика> обладать в связи с 12315[в связь с] |2223[обладать] <**опосредованностью компьютер**>	Все названные формы общения *в связи с* **его опосредованностью компьютером** обладают такой характеристикой, как анонимность, которая имеет целый ряд последствий
<словарь русский язык> пополняться за счёт 12315[за счёт] |2111[пополнять] <**англицизм**>	Наблюдается картина, где словарь русского языка, в основном, пополняется *за счёт* **англицизмов**
<**дефицит вод**> вызывать 12315[вызывать] <снижение работоспособность>	**Дефицит воды** в организме *вызывает* снижение работоспособности, потеря воды в количестве 10% от массы тела приводит к нарушению обмена веществ, потеря 15-20% смертельна при температуре воздуха 30°С, а потеря 25% абсолютно смертельна
<актуальность водоснабжение населённый пункт> кардинально возрастать в случай 12315[в случай] |1211[возрастать] <**разрушение инфраструктура жизнеобеспечение население**>	Следовательно, *в случае* **разрушения инфраструктуры жизнеобеспечения населения** актуальность водоснабжения населённых пунктов кардинально возрастает
<феномен профессиональный деятельность> в результат 12315[в результат] <**стресс**>	В новой редакции Международной классификации болезней (МКБ-11), разработанной ВОЗ (вступит в силу с 1 января 2022 года), выгорание официально зарегистрировано как феномен профессиональной деятельности *в результате* **стресса**
<применение технология реверсивный обучение> являться по причина 12315[по причина] |1133[являть] <**переворачивание традиционный модель обучение**>	Применение технологии реверсивного обучения при обучении иностранным языкам, на наш взгляд, является эффективной технологией обучения *по причине* **переворачивания традиционной модели обучения**, вследствие которого домашнее задание выполняется в аудитории, тогда как классная работа выносится на самостоятельное изучение
<они> могут восприниматься вследствие 11212[воспринимать]|12315[вследствие] <**распад связь время**>	Массовым сознанием *вследствие* **распада связи времён** они могут восприниматься как новые

Класс отношений Быть причиной содержит преимущественно производные предлоги, вследствие однозначности которых точность определения роли Причина высока.

Таблица 3. Примеры триплетов с отношением из класса Быть основанием и сущностями с ролью Основание

Триплет, включающий сущность с ролью Основание	**Фрагмент текста**
<агрегирование> всегда основываться на 1231Е[основывать на] <**абстрагирование**>	Агрегирование всегда *основывается на* **абстрагировании**, т. е. отвлечении от несущественных моментов и выделении наиболее значимых, существенных, типичных черт, закономерностей экономических процессов и явлений
<вершина эволюция> благодаря 1231Е[благодаря] <**память**>	Человек поднялся на «вершину эволюции», *благодаря* своей **памяти** и её постоянному совершенствованию
<память> определить по 1322[определить]|1231Е[по] <**мнение Р.С. Немова**>	*По* **мнению Р. С. Немова**, память можно определить, как способность к получению, хранению и воспроизведению жизненного опыта

Продолжение таблицы 3

Триплет, включающий сущность с ролью Основание	**Фрагмент текста**
<сводный результат> быть получить следовать по 1231E[по]|1121[получить]|8[получить]| 12315[следовать] <**все четыре методика**>	В процессе проведённого исследования (*по* **всем четырём методикам**) были получены следующие сводные результаты
<неопытный команда> быть полностью поразить по 1231E[по] <**ветер**>	В Бискайском заливе судно столкнулось с сильнейшими штормами и шло *по* **ветру**, а неопытная команда была полностью поражена морской болезнью
<психологический основа конструирование различный вариант УПО студент> на основа 1231E[на основа] <**компонент**>	В данной статье обсуждается вопрос о психологической основе конструирования различных вариантов УПО студентов *на основе* **компонентов** в его структуре
<жизнь> развиваться по 1231E[по]|1213[развивать] <**единый закон эволюция**>	Жизнь на них развивается *по* **единому закону эволюции**, восходя ко все более сложным формам и видам, достигая постепенно своего высшего уровня, разумной жизни

Класс отношений Быть основанием включает многозначный предлог «по», из-за которого требуется дополнительный анализ окружения для более точного определения роли присоединяемой им сущности.

Таблица 4. Примеры триплетов с отношением из класса Быть условием и сущностями с ролью Условие

Триплет, включающий сущность с ролью Условие	**Фрагмент текста**
<вероятность применение> подтверждать в условие 1231C[в условие]|1232[подтверждать] <**рост геополитический напряжение международный отношенье**>	*В условиях* **роста геополитического напряжения международных отношений** в настоящее время подтверждает вероятность применения в современной войне оружия массового поражения
<физиологический потребность> зависеть от 1231C[зависеть от] <**возраст**>	Физиологическая потребность в воде *зависит от* **возраста**, характера работы, пищи, профессии, климата и т.д.
<название> обусловить 1231C[обусловить] <**особенность ритмика электроэнцефалограмма**>	Существует две фазы сна: медленная и быстрая. Эти названия *обусловлены* **особенностями ритмики электроэнцефалограммы** (ЭЭГ) во время сна медленной активностью в ФМС и более быстрой в ФБС
<мозг человек> переключаться при 12113[переключать]|1231C[при]|22212[при] <**недостаток сон**>	*При* **недостатке сна** мозг человека переключается на более примитивные формы деятельности и не в состоянии нормально управлять эмоциями, эмоциональные зоны мозга становятся реактивными
<изменение> определяться в рамка 1231C[в рамка]|1322[определять] <**обогащающей модели**>	В связи с этим эффективность обучения студентов *в рамках* «**обогащающей модели**» определяется изменениями в их интеллекте как в специфической форме организации индивидуального познавательного (ментального) опыта, обеспечивающей возможность эффективного включения в учебно-познавательную деятельность студентов, так и в учебно-познавательном опыте как результате этой деятельности

5. Заключение

В данной работе оценивалась адекватность автоматического назначения ролей, определенных в соответствии с классами отношений, применительно к онто-графовому представлению содержания текста.

Использование для идентификации отношений и сущностей классификации с явно заданными характеристическими признаками делает возможным количественно определять семантическую близость между сущностями и поисковыми образами, а использование ролей сущностей повышает точность поиска и позволяет при поиске обращаться к ролям сущностей как к самостоятельному поисковому атрибуту.

Проведённый эксперимент на текстах разного стиля показал минимальную точность 77% и максимальную — 100% (для ролей Основание и Причина, соответственно) и, таким образом, подтверждается целесообразность использования предложенного подхода для автоматизированного индексирования и поиска научно-технической литературы. Однако для тексов художественного стиля множество лингвистических конструкций для идентификации отношений и ролей должно быть расширено.

Литература

[1] Скороходько Э. Ф. Лингвистические проблемы обработки теистов в автоматизированных информационно-поисковых системах // Вопросы информационной теории и практики. 1974. № 25. С. 5–120.

[2] СИНТОЛ // Сборник переводов по вопросам информационной теории и практики. М.: ВИНИТИ, 1968. С. 36–47; 50–52; 66–72; 76–80.

[3] Максимов Н. В., Гаврилкина А. С., Андронова В. В., Тазиева И. А. Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 2018. № 11. С. 32–42.

[4] Максимов Н. В., Голицына О. Л. От семантического к когнитивному информационному поиску. Основные положения и модели глубинного семантического поиска // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 2022. № 6. С. 1–16.

[5] Голицына О. Л., Гаврилкина А. С. Об одном подходе к выделению имён сущностей и связей в задаче построения семантического поискового образа // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2021. № 3. С. 17–26.

[6] Van Renssen A. Gellish: A Generic Extensible Ontological Language. Delft: Delft University Press, 2005. 238 с.

[7] Kozaki K., Sunagawa E., Kitamura Y., Mizoguchi R. Role Representation Model Using OWL and SWRL // Roles'07. Proceedings of the 2nd Workshop on Roles and Relationships in Object Oriented Programming, Multiagent Systems, and Ontologies. Berlin, 2007. С. 39–46.

[8] Плунгян В. А. Основные синтаксические граммемы имени // Общая морфология: Введение в проблематику: Учебное пособие. Изд. 2-е, исправленное. М.: Едиториал УРСС, 2003. 384 с.

[9] Семантическая роль как элемент метаязыков общей и специальной типологии // 40 лет Санкт-Петербургской типологической школе. М.: Знак, 2004. С. 233–252.

[10] Allan K. Natural language semantics. Oxford, UK; Malden, Mass.: Blackwell, 2001. 529 p.

[11] Кашкин Е. В., Ляшевская О. Н. Типы информации о лексических конструкциях в системе ФреймБанк // Труды института русского языка им. В.В. Виноградова. 2015. Т. 6. С. 464–556.

[12] Максимов Н. В. Методологические основы онтологического моделирования документальной информации // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 2018. №3. С. 6–22.

[13] Шелманов А. О., Каменская М. А. Обучение анализатора для определения ролевых структур высказываний в текстах на русском языке на автоматически размеченном корпусе // Труды Института системного анализа Российской академии наук. 2017. Т. 67, №. 2. С. 104–120.
[14] Van Valin R. D. Generalized semantic roles and the syntax-semantics interface //Empirical issues in formal syntax and semantics. 1999. Vol. 2. P. 373–389.
[15] Fillmore Ch. The case for case. // Universals in linguistic theory / Ed by E.Bach and R.T.Harms. N. Y., Chicago, San Francisco, 1968.
[16] Kingsbury P. R., Palmer M. From TreeBank to PropBank // LREC. 2002. P. 1989-1993.
[17] Fillmore C. J. et al. FrameNet in action: The case of attaching // International journal of lexicography. 2003. Vol. 16 (3). P. 297–332.
[18] Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З. Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря. М., 2010.
[19] Автоматическая разметка семантических ролей в русском языке: автореферат дис. ... кандидата филологических наук: 10.02.21 / Кузнецов Илья Олегович; [Место защиты: Моск. гос. ун-т им. М.В. Ломоносова]. М., 2016. 25 с.
[20] Свидетельство о государственной регистрации программы для ЭВМ № 2021610648, 15.01.2021 / Авторы: Максимов Н.В., Голицына О.Л., Монанков К.В., Гаврилкина А.С.

On the Identification of Situational Roles of Entities in the Context of the Semantic Information Retrieval Task

Anastasiia S. Gavrilkina, Nikolay V. Maksimov, Olga L. Golitsina

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)

The article discusses approaches to identifying the semantic roles of entities in the context of the task of semantic information retrieval. Various definitions of roles are analyzed. An approach is proposed to the distribution of roles in accordance with the classes of the ontology of relations built on the basis of an extended functional model. An experiment was carried out to assess the quality of automatic determination of the roles Problem, Foundation, Condition.

Keywords: semantic roles, information retrieval, automatic text processing, full-text indexing

Reference for citation: Gavrilkina A.S., Maksimov N.V., Golitsina O.L. On the Identification of Situational Roles of Entities in the Context of the of Semantic Information Retrieval Task // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). - St. Petersburg: ITMO University, 2024. P. 21–31. DOI: 10.17586/2541-9781-2024-7-21–31

Reference

[1] Skorohod'ko E. F. Lingvisticheskie problemy obrabotki teistov v avtomatizirovannyh informacionno-poiskovyh sistemah // Voprosy informacionnoj teorii i praktiki. 1974. № 25. S. 5–120. (in Russian).
[2] SINTOL // Sbornik perevodov po voprosam informacionnoj teorii i praktiki. M.: VINITI, 1968. S. 36–47; 50–52; 66–72; 76–80. (in Russian).
[3] Maksimov N. V., Gavrilkina A. S., Andronova V. V., Tazieva I. A. Sistematizaciya i identifikaciya semanticheskih otnoshenij v ontologiyah nauchno-tekhnicheskih predmetnyh

oblastej // Nauchno-tekhnicheskaya informaciya. Ser. 2: Informacionnye processy i sistemy. 2018. № 11. S. 32–42. (in Russian)
[4] Maksimov N. V., Golicyna O. L. Ot semanticheskogo k kognitivnomu informacionnomu poisku. Osnovnye polozheniya i modeli glubinnogo semanticheskogo poiska // Nauchno-tekhnicheskaya informaciya. Ser. 2: Informacionnye processy i sistemy. 2022. № 6. S. 1–16. (in Russian)
[5] Golicyna O. L., Gavrilkina A. S. Ob odnom podhode k vydeleniyu imyon sushchnostej i svyazej v zadache postroeniya semanticheskogo poiskovogo obraza // Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2021. № 3. S. 17–26. (in Russian)
[6] Van Renssen A. Gellish: A Generic Extensible Ontological Language. Delft: Delft University Press, 2005. 238 p.
[7] Kozaki K., Sunagawa E., Kitamura Y., Mizoguchi R. Role Representation Model Using OWL and SWRL // Roles'07. Proceedings of the 2nd Workshop on Roles and Relationships in Object Oriented Programming, Multiagent Systems, and Ontologies. Berlin, 2007. P. 39–46.
[8] Plungyan V. A. Osnovnye sintaksicheskie grammemy imeni // Obshchaya morfologiya: Vvedenie v problematiku: Uchebnoe posobie. Izd. 2-e, ispravlennoe. M.: Editorial URSS, 2003. 384 s. (in Russian)
[9] Semanticheskaya rol' kak element metayazykov obshchej i special'noj tipologii // 40 let Sankt-Peterburgskoj tipologicheskoj shkole. M.: Znak, 2004. S. 233–252. (in Russian)
[10] Allan K. Natural language semantics. Oxford, UK; Malden, Mass.: Blackwell, 2001. 529 p.
[11] Kashkin E. V., Lyashevskaya O. N. Tipy informacii o leksicheskih konstrukciyah v sisteme FrejmBank // Trudy instituta russkogo yazyka im. V.V. Vinogradova. 2015. T. 6. S. 464–556. (in Russian)
[12] Maksimov N. V. Metodologicheskie osnovy ontologicheskogo modelirovaniya dokumental'noj informacii // Nauchno-tekhnicheskaya informaciya. Ser. 2: Informacionnye processy i sistemy. 2018. №3. S. 6–22. (in Russian)
[13] Shelmanov A. O., Kamenskaya M. A. Obuchenie analizatora dlya opredeleniya rolevyh struktur vyskazyvanij v tekstah na russkom yazyke na avtomaticheski razmechennom korpuse // Trudy Instituta sistemnogo analiza Rossijskoj akademii nauk. 2017. T. 67, №. 2. S. 104–120. (in Russian)
[14] Van Valin R. D. Generalized semantic roles and the syntax-semantics interface //Empirical issues in formal syntax and semantics. 1999. Vol. 2. P. 373–389.
[15] Fillmore Ch. The case for case. // Universals in linguistic theory. Ed by E.Bach and R. T. Harms. N. Y., Chicago, San Francisco, 1968.
[16] Kingsbury P. R., Palmer M. From TreeBank to PropBank // LREC. 2002. P. 1989–1993.
[17] Fillmore C. J. et al. FrameNet in action: The case of attaching // International journal of lexicography. 2003. Vol. 16. № 3. P. 297–332.
[18] Apresyan Yu. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z. Teoreticheskie problemy russkogo sintaksisa: vzaimodejstvie grammatiki i slovarya. M., 2010. (in Russian)
[19] Avtomaticheskaya razmetka semanticheskih rolej v russkom yazyke: avtoreferat dis. ... kandidata filologicheskih nauk: 10.02.21 / Kuznecov Il'ya Olegovich; [Mesto zashchity: Mosk. gos. un-t im. M. V. Lomonosova]. M., 2016. 25 s. (in Russian)
[20] Svidetel'stvo o gosudarstvennoj registracii programmy dlya EVM № 2021610648, 15.01.2021 / Avtory: Maksimov N. V., Golicyna O. L., Monankov K. V., Gavrilkina A. S. (in Russian)

Функционирование устойчивой модели <X от слова Y> в современном интернет-пространстве

Ю. С. Локалина

Санкт-Петербургский государственный университет

lokalina13@mail.ru

Аннотация

В работе рассматриваются особенности функционирования устойчивой модели <X от слова Y>, которая способствует появлению в языке таких конструкций, как *<от слова совсем>*, *<от слова вообще>* и под. Такие фразеологизированные выражения произошли от свободного сочетания *от слова*, которое обычно употребляется как отсылка к этимологии слова. Компоненты, выступающие в роли такой «отсылки» (*совсем, вообще* и др.), являются усилительными единицами, так называемыми интенсификаторами.

Это достаточно новое в языке явление порождает омонимию в поисковых сервисах: появляется большое количество вариантов с ключевым сочетанием *от слова*: как отсылка к этимологии, так и другие случаи (конструкции, свойственные разговорной речи).

В результате проведённого исследования были выявлены пунктуационные, синтаксические и др. особенности употребления конструкции *<от слова совсем>*; проанализированы подобные конструкции, построенные по модели *<X от слова Y>*. Эта модель функционирует настолько свободно, что на месте X и Y могут стоять любые части речи, из-за чего поисковый запрос «*от слова*» «выдаёт» большое количество употреблений с чем-либо, однако не ищет случаи с использованием графики и кавычек. Кроме этого, на основе собранных данных была построена шкала интенсификации, которая собрала все возможные варианты использования конструкции в Интернете. Полученные результаты могут быть использованы в дальнейших исследованиях, связанных с интенсифицирующими конструкциями, языковыми корпусами и поисковыми сервисами.

Ключевые слова: типические конструкции, устойчивая модель, интенсификатор, омонимия, конструкция

Библиографическая ссылка: Локалина Ю. С. Функционирование устойчивой модели <X от слова Y> в современном интернет-пространстве // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 32–41. DOI: 10.17586/2541-9781-2024-7-32–41

Особенностью современной разговорной речи является тенденция к гиперболизации. На это обратила внимание ещё А. Вежбицкая, которая отметила, что гиперболизованность русской речи как средство «выражения любых оценок, как положительных, так и отрицательных» [1, с. 84], является яркой чертой современного языка, использующего интенсифицирующие средства. Действительно, часто в разговорной речи носителей языка, в рекламе, в Интернете и в других источниках можно услышать такие словосочетания, как *суперпитательный батончик, грандиозный скандал* и т. п.

Анализ этого явления позволяет говорить о вхождении в активную лексику носителей языка целого ряда особых единиц — интенсификаторов. Интенсификаторами могут быть как отдельные слова, так и конструкции, которые и являются объектом внимания в настоящем исследовании.

В последнее время в русском языке появилось множество конструкций типа *<от слова совсем>*, *<от слова вообще>* и т. п. Причиной их появления как раз и служит тенденция к интенсификации.

По мнению П. А. Леканта, под интенсификацией понимается ряд функционально-семантических операций, включающих подчёркивание (акцентирование), усиление, полноту, градацию (обычно высокую её степень), обобщение и оценку [2, с. 58]. Лексические и фразеологизированные интенсификаторы — это языковые единицы, план содержания которых включает вышеупомянутые функционально-семантические операции. Ср., например, исследование функционирования интенсификатора *самый* в русской устной речи [3], а также специальную работу о семантической типологии интенсификаторов в русском языке [4].

Ещё одной причиной появления подобных конструкций является существование в русском языке такой устойчивой модели, как *<X от слова Y>*. Она выполняет функцию объяснения и верификации (путём апелляции к источнику). Это характерно для научного дискурса, для отсылок к этимологическому источнику. Однако данная конструкция стала развиваться и по другому направлению, в результате чего в языке появились такие употребления [5, с. 74]:

- *«Успех» от слова «успеть»;*
- *Обожаю снимать красивых женщин, от слова «женственность»;*
- *Возможности подключить безлимитный НЕТ, от слова НИКАКОЙ <...>*.

По этим примерам можно судить, что рассматриваемая конструкция имеет как отрицательную, так и положительную коннотацию. Кроме того, она «допускает использование слов, которые вряд ли возможны в одной предикативной конструкции» [5, с. 74]:

- *Идя в кино (к слову, зал был полон от слова «целиком»), я не знала, что это лишь одна из частей;*
- *Твоё мнение интересно от слова нисколько.*

Интересно, что такие «креативные» конструкции переняли черты, особенности конструкции, положившей им начало, — *от слова совсем*.

Такое развитие и «успех» конструкции *<X от слова Y>* можно объяснить системой языка, в которой в принципе возможна реализация типических конструкций.

Возможность функционирования в языке типических конструкций на основе устойчивой модели *<X от слова Y>* стала, в числе прочего, причиной появления омонимии в поисковых сервисах: Google и Яндекс.

Ввод в поисковую строку Google, Яндекс «от слова» предлагает, кроме сайтов с отсылками к конструкциям *<от слова совсем>* и *<от слова вообще>*, сайты с объяснением этимологии слов *совсем*, *вообще*. Яндекс [6] «советует» определить значение предлога *от*. Кроме этого, поисковый сервис понимает *от слова* как предлог с существительным во множественном числе и предлагает найти слова с предлогом *от*.

Только поиск по подкорпусам Национального корпуса русского языка (НКРЯ) показывает разные употребления: в основном подкорпусе предлагаются варианты с этимологией, а подкорпус «Социальные сети» предлагает контексты использования конструкций *<от слова совсем>* и *<от слова вообще>*.

Важно, что одна из этих конструкций (*<от слова совсем>*) вошла в речь носителей языка примерно 17 лет назад. По мнению С. С. Белоусова, первое её употребление зафиксировано в 2005 году [5, с. 72-77].

Анализ таких единиц позволяет заметить, что наибольшую известность приобрела конструкция <*от слова совсем*>, а все остальные являются только производными от неё. Это могут подтвердить данные сайта Trends.google.ru [7] (см. рис. 1).

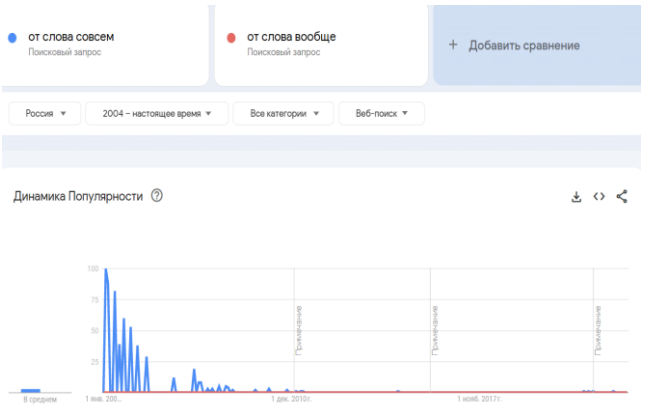

Рис. 1. Популярность конструкции <*от слова совсем*>

На рисунке 1 в сравнении показана популярность двух конструкций, одна из которых (<*от слова совсем*>) имеет гораздо больше запросов, чем другая (<*от слова вообще*>).

Исследование конструкции <*от слова совсем*> показало, что есть члены предложения, которые активно взаимодействуют с ней. По результатам анализа собранного материала, данная конструкция чаще всего усиливает сказуемое, выраженное глаголом (*Кадры с фронтальным светом* ***не получаются от слова «совсем»***). Далее результаты разнятся. «Второе место» занимают обстоятельства меры и степени (*Они абсолютно* ***чёрствые, от слова совсем***), а также наиболее редкие случаи — существительные и прилагательные, выступающие в роли сказуемого (*Он* ***хорошенький от слова совсем***).

На основе таких результатов можно сделать вывод: сказуемые, выраженные глаголами, наиболее часто поддаются интенсификации (уточним, что функция *конструкции <от слова совсем>* в предложении — обстоятельство меры и степени); все остальные случаи — вариации, связанные с данной конструкцией. Это можно представить так: X — сказуемое, Y — <*от слова совсем*>. В связи с этим можно обозначить схематично как XY и другие варианты — ZY, QY и т. д., где первые буквы — объект усиления. Можно встретить также их сочетания: XZY, XQY и т. д. Приведём примеры:

- *Ты же* ***никогда не могла бегать*** *#отсловасовсем;*
- *Взрослых* ***не уважает вообще****, ну просто,* ***от слова совсем****;*
- *Его* ***не происходило вообще****, от слова совсем. Просто от слова совсем;*
- *Тут мне уже* ***ничего не хотелось****, от слова «СОВСЕМ».*

Нередко в одном предложении встречаются и наречие, и сказуемое, взаимодействующие с конструкцией. Чтобы узнать, семантику сказуемого или наречия хотят усилить носители языка, был проведён опрос. В процессе его разработки было решено строить его следующим образом: от более очевидных и простых случаев — к более сложным (спорным).

Среди 40 респондентов, носителей языка, большая часть — женщины (82,5 %), остальные — мужчины (17,5 %). По возрасту респонденты распределились следующим образом:

- до 20 лет: 45 %;
- 21-30 лет: 40 %;
- 31-45 лет: 12,5 %;
- 46 и более: 2,5 %.

Важно, что не все респонденты имеют филологическое образование, а именно 52,5 %, остальные — 47,5 %, т. е. в этом отношении участники опроса разделились примерно пополам.

Как показал опрос, 62,5 % респондентов знают о существовании конструкции *от слова совсем*, а 37,5 % — нет. Это хорошие показатели, результатом которых стало большое количество разных мнений.

Источниками, откуда респонденты узнали об исследуемой конструкции, были Интернет и друзья/знакомые (см. рис. 2).

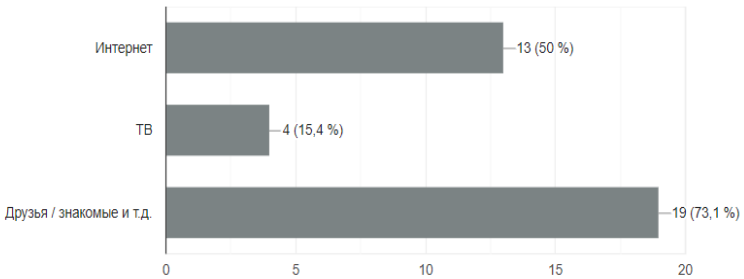

Рис. 2. Источник информации о конструкции

Любопытно, что, несмотря на популярность исследуемой конструкции, используют её все же далеко не все носители языка. На вопрос «Используете ли Вы конструкцию или хештег?» были даны следующие ответы (см. рис. 3).

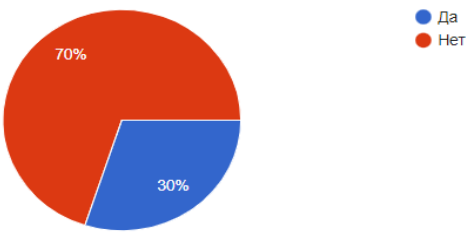

Рис. 3. Частота использования конструкции или хештега

Респондентам было предложено 7 контекстов, в которых они должны найти так называемое «определяемое слово», к которому относится конструкция *от слова совсем*.

Результаты распределились следующим образом:

1) *На пляже отсутствуют лежаки #отСловаСовсем.*

В первом контексте респонденты определили «главным словом» сказуемое. Это самый простой и очевидный случай (см. рис. 4).

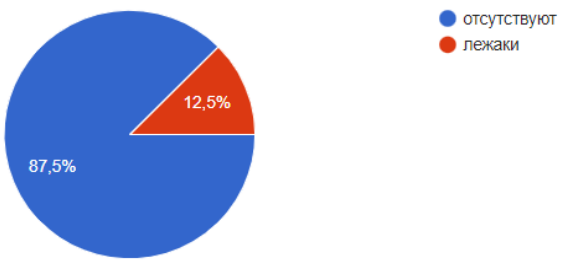

Рис. 4. Первый контекст

2) *Всё работает иначе. #отсловасовсем*

В контексте (2) респонденты «отдали большинство голосов» за наречие *иначе*, т. е. за обстоятельство меры и степени. Этот контекст также можно отнести к наиболее простым и очевидным (см. рис. 5).

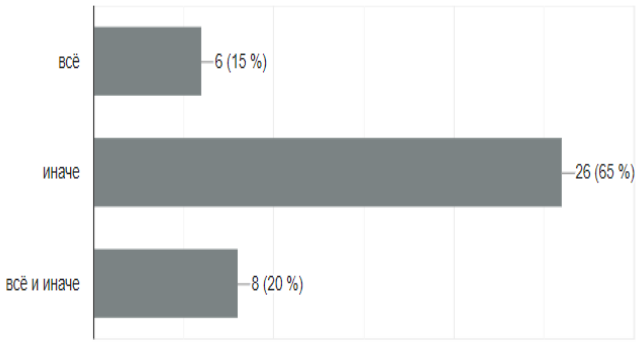

Рис. 5. Второй контекст

3) *Столько вдохновения, когда не получается вообще #отсловасовсем.*

В третьем примере предпочтение почти в равной мере было отдано и сказуемому, и обстоятельству меры и степени. Результаты опроса распределились так, как показано на рис. 6.

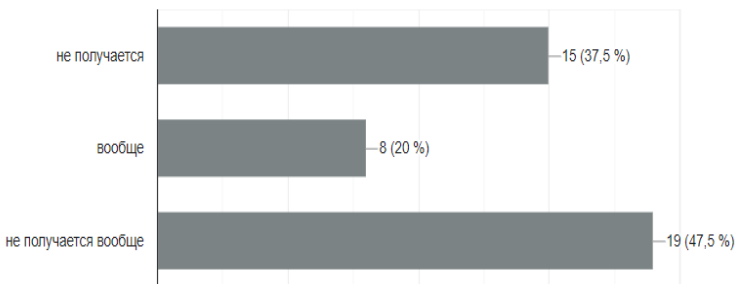

Рис. 6. Третий контекст

4) *Вещи ваще не успевают сносится, надоесть и тем более отработать вложенные в них кровные #отсловасовсем.*

В четвёртом, более сложном, предложении большая часть респондентов «проголосовала» за ряд сказуемых (см. рис. 7).

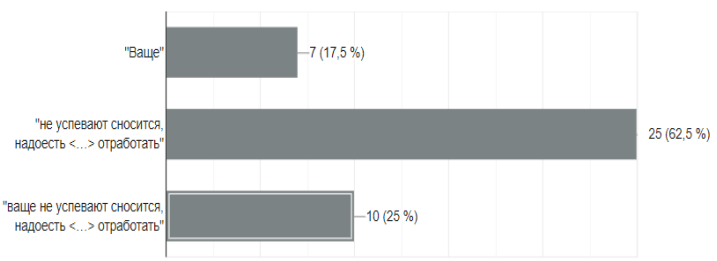

Рис. 7. Четвёртый контекст

5) *Я вообще никак не ориентируюсь в модных брендах от слова совсем.*

В пятом предложении «главным словом» вновь стало сказуемое (см. рис. 8).

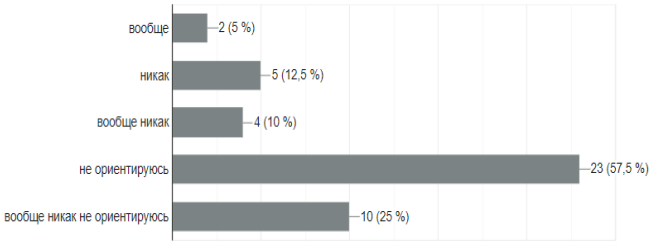

Рис. 8. Пятый контекст

6) *Их в принципе не существует ОтСловаСовсем.*

Шестой контекст показал такие же результаты, как предыдущие два, т. е. «главное слово» здесь — сказуемое (см. рис. 9).

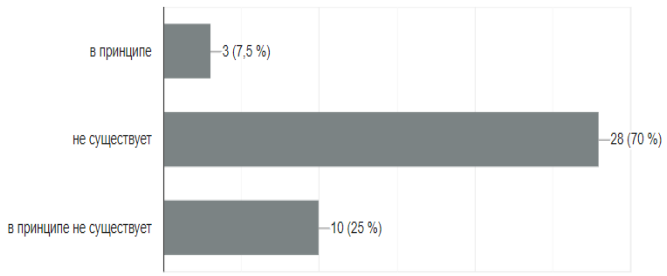

Рис. 9. Шестой контекст

7) *Но никак от слова совсем не заставить меня написать.*

В седьмом предложении носители языка определили «главным» наречие *никак* — обстоятельство меры и степени, однако и сказуемое занимает «лидирующую» позицию (см. рис. 10).

Рис. 10. Седьмой контекст

8) *Её вообще нигде не было от слова совсем.*

В 8-ом предложении мнения респондентов интересно разделились: «лидирует» наречие *нигде*, далее следуют сказуемое и наречие *вообще* в разных вариациях (см. рис. 11).

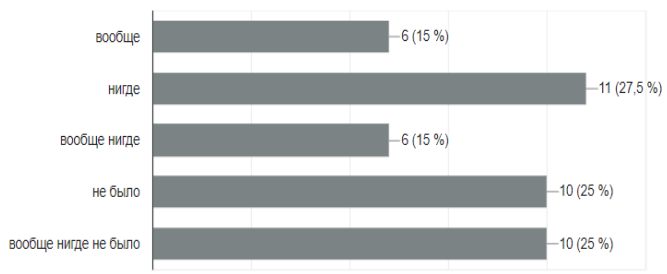

Рис. 11. Восьмой контекст

Полученные данные позволяют сделать ряд выводов. Как уже было сказано, конструкция *от слова совсем* усиливает семантику сказуемого в предложении. Это показал и проведённый опрос. Конечно, в ответах респондентов были и другие варианты, однако они являются достаточно редкими. А так как *от слова совсем* выполняет функцию обстоятельства меры и степени, то усиливает эта конструкция чаще всего сказуемое или обстоятельства меры и степени в предложении, т. е. слова, обозначающие степень проявления действия, состояния или свойства.

Кроме этого, важно отметить, что в материале исследования не было отмечено случаев, когда конструкция и взаимодействующее с ней слово были бы расположены друг от друга на расстоянии более, чем в 2-3 слова. Это можно объяснить тесной связью между интенсификатором, роль которого выполняет *<от слова совсем>,* и усиливаемым словом (сказуемым, обстоятельством меры и степени и т. д.). Они не могут находиться далеко друг от друга, поскольку тогда будет нарушена логика предложения. Приведём примеры:

- *Раньше я **не умела** отдыхать #ОтСловаСовсем#;*
- *Первые 7 лет своей жизни я **ни слова не знала** на русском #отсловасовсем;*
- *Ну **нет** интереса и времени от слова совсем;*
- *Я **не спала** сегодня от слова «совсем»;*
- *Я **не разбираюсь** в советской литературе от слова совсем.*

На основе пользовательского подкорпуса, в который вошло 200 контекстов из социальных сетей (Telegram, ВКонтакте, Одноклассники, Живой Журнал) и из Национального корпуса русского языка, была создана шкала интенсификации. Она собрала все возможные случаи употребления конструкции *от слова совсем* (см. рис. 12).

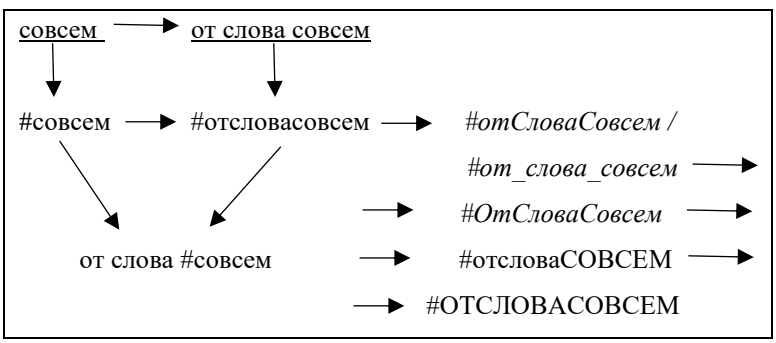

Рис. 12. Шкала интенсификации

Хештеги в схеме расположены согласно результатам опроса респондентов, которые определили, какой из хештегов с разным написанием (строчными или прописными буквами) больше усиливает значение, а какой — меньше. Двадцати респондентам был предложен список хештегов, а также шкала с цифрами от 1 до 6 (по количеству разных хештегов), которые означают степень интенсификации (1 — самое слабое, 6 — самое сильное). Для опроса специально были выбраны студентки филологического факультета в возрасте от 20 до 25 лет, которые являются активными пользователями Интернета и понимают специфику употребления хештегов. Обобщённые результаты проведённого опроса (в виде серии диаграмм) представлены на рис. 13.

Также нужно отметить особенность обособления рассматриваемой конструкции на письме. Носители языка решают эту проблему по-разному, поэтому появились варианты написания с использованием запятой, отделяющей конструкцию от главного предложения:

- *Никакого отёка, от слова совсем;*

с использованием точки, в результате чего образуется парцелляция:

- *Вы знаете, я раньше не умела благодарить. От слова совсем;*

а также встречается и многоточие, интонационно выделяющее конструкцию:

- *Это был потрясающий отдых… от слова совсем.*

Из представленных данных можно сделать вывод, что одного способа обособления конструкции пока не существует.

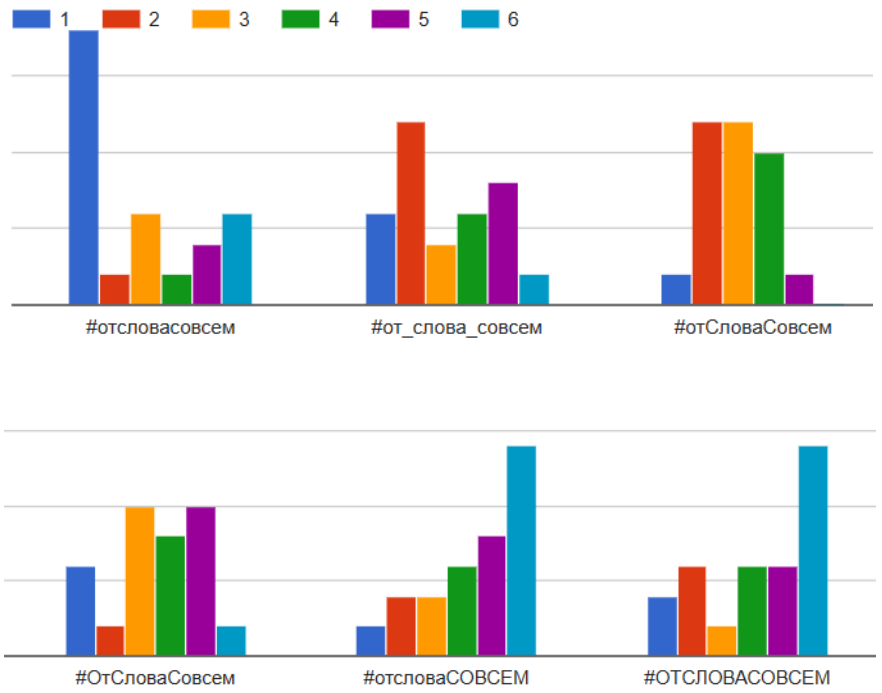

Рис. 13. Результаты опроса

Особенность использования конструкции как хештега — графика. На основе собранных контекстов были выделены следующие варианты написания #отсловасовсем:

- *использование только прописных букв: #ОТСЛОВАСОВСЕМ;*
- *выделение отдельных начальных букв слов: #отСловаСовсем;*
- *выделение всех начальных букв слов: #ОтСловаСовсем;*
- *использование прописных букв для выделения нужного слова: #отсловаСОВСЕМ.*

Таким способом пользователи сети часто прибегают к выражению эмоций. Интересно, что, набирая весь текст прописными буквами и создавая графически иллюзию изменения интонации, носители языка выражают широкий спектр эмоций и чувств, например, гнев, раздражение, недовольство и т. д. Рассмотрим примеры:

- *Раньше я не умела отдыхать #ОтСловаСовсем;*
- *Потому что от психологического насилия законы России не защищают. #отсловаСОВСЕМ.*

Примеры показывают, что носитель языка пытается сделать акцент на отдельных словах, создавая тем самым логическое ударение в высказывании. Такое явление можно назвать «двойной интенсификацией».

В результате проведённого исследования были выявлены пунктуационные, синтаксические и др. особенности употребления конструкции *<от слова совсем>*; проанализированы подобные конструкции, построенные по модели *<X от слова Y>*. Эта модель, перенявшая особенности использования *<от слова совсем>*, функционирует настолько свободно, что на месте X и Y могут стоять любые части речи. Из-за этого

поисковый запрос «*от слова*» «выдаёт» огромное количество употреблений с чем-либо, однако не ищет конструкции с использованием графики и кавычек, в которые часто пользователи Интернета «заключают» усиливающий компонент (совсем, вообще и т. д.).

Все полученные данные могут быть использованы в дальнейших исследованиях, связанных с интенсифицирующими конструкциями, языковыми корпусами и поисковыми сервисами.

Таким образом, работа с материалом современной русской разговорной речи — одна из важнейших задач лингвистики и компьютерных наук.

Литература

[1] Вежбицкая А. Язык. Культура. Познание / Пер. с англ. / Отв. ред. М. А. Кронгауз; вступ. ст. Е. В. Падучевой. М.: Русские словари, 1996. 416 с.

[2] Лекант П. А. Субъективная аналитическая категория интенсива в русском языке // Аналитизм в лексико-грамматической системе русского языка. Монография / П. А. Лекант (ред.). М.: МГОУ, 2011. С. 130–136.

[3] Сунь Сяоли. Самый как слово-интенсификатор в современном русском языке: модели употребления // Мир русского слова. № 3, 2021. С. 53–60.

[4] Сандакова М. В. К вопросу о семантической типологии интенсификаторов // Русский язык в России и за рубежом: изучение активных процессов в языке и речи. Сб. научных статей / Отв. ред. Л. В. Рацибурская. Н. Новгород: Изд-во ННГУ им. Н. И. Лобачевского, 2021. С. 283–289.

[5] Белоусов С. С. От слова совсем как грамматическая конструкция // Учёные записки Петрозаводского гос. ун-та. 2016, № 7–1 (160). С. 72–77.

[6] Yandex.ru [сайт]. URL: https://yandex.ru/search/?text=от+слова&clid=2233626&search_source=dzen_desktop_safe&lr=2 (дата обращения: 17.06.2023).

[7] TRENDS.GOOGLE.RU [сайт]. URL: https://trends.google.ru/trends/explore?date=all&geo=RU&q=от%20слова%20совсем,от%20слова%20вообще&hl=ru (дата обращения: 04.04.2023).

Functioning of the Sustainable Model <X from the word Y> in the Modern Internet Space

Julia. S. Lokalina

Saint Petersburg State University

The paper considers the peculiarities of the functioning of the stable model <X ot slova Y>, which contributes to the appearance in the language of such constructions as <ot slova sovsem>, <ot slova voobshche> and so on. Such phraseologized expressions originated from the free combination ot slova, which is usually used as a reference to the etymology of the word. The components acting as such a «reference» (sovsem, voobshche, etc.) are intensifying units, so-called intensifiers. They indicate a different degree of manifestation of some feature and contribute to the functioning of the construction as an intensifier in this or that text.

This rather new phenomenon in the language gives rise to homonymy in search services: a large number of variants with the key combination from the word appear: both reference to etymology and other possible cases (constructions peculiar to colloquial speech).
As a result of this research, punctuation, syntactic and other peculiarities of the use of the construction <ot slova sovsem> were revealed; similar constructions built according to the model <X ot slova Y> were analyzed. This model functions so freely that X and Y can be replaced by any parts of speech, so the search query «from the word» gives a huge number of variants of usage with something, but does not search for usage using graphics and quotation marks. In addition, an intensification scale was constructed based on the collected data, which collected all possible uses of the construction on the Internet. All this can be used in further research related to intensifying constructions, language corpora and search services.

Keywords: typical constructions, stable model, intensifier, homonymy, construction

Reference for citation: Lokalina Ju. S. Functioning of the Sustainable Model <X from the word Y> in the Modern Internet Space // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). - St. Petersburg: ITMO University, 2024. P. 32–41. DOI: 10.17586/2541-9781-2024-7-32–41

Reference

[1] Vezhbickaya A. Yazyk. Kul'tura. Poznanie / Per. s angl. / Otv. red. M. A. Krongauz; vstup. st. E. V. Paduchevoj. M.: Pucskie slovari, 1996. 416 s. (in Russian)

[2] Lekant P. A. Sub"ektivnaya analiticheskaya kategoriya intensiva v russkom yazyke // Analitizm v leksiko-grammaticheskoj sisteme russkogo yazyka. Monografiya / P. A. Lekant (red.). M.: MGOU, 2011. S. 130–136. (in Russian)

[3] Sun' Syaoli. Samyj kak slovo-intensifikator v sovremennom russkom yazyke: modeli upotrebleniya // Mir russkogo slova. № 3, 2021. S. 53–60 (in Russian).

[4] Sandakova M. V. K voprosu o semanticheskoj tipologii intensifikatorov // Russkij yazyk v Rossii i za rubezhom: izuchenie aktivnyh processov v yazyke i rechi. Sb. nauchnyh statej / Otv. red. L. V. Raciburskaya. N. Novgorod: Izd-vo NNGU im. N. I. Lobachevskogo, 2021. S. 283–289. (in Russian)

[5] Belousov S. S. Ot slova sovsem kak grammaticheskaya konstrukciya // Uchenye zapiski Petrozavodskogo gos. un-ta. 2016, № 7-1 (160). S. 72–77. (in Russian)

[6] Yandex.ru. URL: https://yandex.ru/search/?text=от+слова&clid=2233626&search_source=dzen_desktop_safe&lr=2 (access date: 17.06.2023).

[7] TRENDS.GOOGLE.RU. URL: https://trends.google.ru/trends/explore?date=all&geo=RU&q=от%20слова%20совсем,от%20слова%20вообще&hl=ru (access date: 04.04.2023).

К конструктивному определению свойств информации

Н. В. Максимов, А. А. Лебедев

Национальный исследовательский ядерный университет «МИФИ»

nv-maks@yandex.ru, lebedevalex@live.ru

Аннотация

Работа посвящена исследованию и обоснованию свойств информации. Информация рассматривается как всеобщее свойство материального мира быть определенным (существовать и изменяться в соответствии с законами природы), быть определяемым (воспринимаемым и идентифицируемым) и быть определяющим (способным изменять состояние целевого объекта). Основываясь на понимании информации, как особой формы материи (связей или зависимостей объектов, явлений или мыслительных процессов), вводятся фундаментальные, прагматические и атрибутивные свойства. Показано, что рассматриваемые свойства могут иметь общее фундаментальное происхождение, что позволяет конструктивно определить подходы к оценке информации.

Ключевые слова: фундаментальные свойства информации, атрибутивные свойства информации, прагматические свойства информации, свободная информация, связанная информация, информационные взаимодействия

Библиографическая ссылка: Максимов Н. В., Лебедев А. А. К конструктивному определению свойств информации // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 42–53. DOI: 10.17586/2541-9781-2024-7-42–53

1. Введение

Понятие информации присутствует во многих предметных областях. И хотя в каждом случае это понятие определяется по-разному, тем не менее, для соответствующей области знаний оно практически всегда бывает достаточно конструктивным. Однако, различие в подходах к определению затрудняет задачу исследования свойств и структуры информации. Например, физические теории микро- и макромира, практически не используют понятие «информация», хотя сами физические теории являются результатом процесса познания (т. е. информационного процесса), не только осуществляемого в физической среде, но и приводящего к изменению этой среды.

Роль информации фундаментальна: обладая особыми свойствами, она обеспечивает, наряду с энергией и материей, управляемое развитие самой среды, в которой, в том числе, живёт человек. Информация присутствует везде. Но при этом она многолика, и поэтому важно иметь адекватное представление об информации как об *объекте*, действующем и используемом не только в сфере информационных технологий, но и в сфере разумной деятельности человека. Только в этом случае можно объективно исследовать её структуру и свойства. Кроме того, информация (по крайней мере свободная) предопределяет использование для её фиксации не только носителя, но и некоторого языка, возможности которого должны быть адекватны не только семантике конкретного сообщения, но и природе и структуре информации как таковой.

2. О сущности информации[1]

В самом общем смысле можно сказать, что информация есть некоторая *особая* форма связей или зависимостей объектов, явлений или мыслительных процессов. Информация, определяемая в [3] как устойчивые (определенное время) неоднородности произвольной физической природы, — это не собственно части среды, находящиеся в различающихся состояниях, а их характер, т. е. это ***образ различий***, который фиксируется, что отражается определением информации, сформулированном Г. Кастлером применительно к биологическим системам: «Информация есть запомненный выбор одного варианта из множества возможных и равноправных.» [4], что уже позволяет «объяснить» процессы генерации информации, как макрообъекта.

Информация как абстракция — это понятие, относящееся к классу закономерностей материального мира и, в том числе, его отражения в человеческом сознании. Но, с другой стороны, информация как физическая сущность так или иначе отождествляется с сигналами, физическими объектами и их взаимодействиями, которые наблюдаются в живых или неживых, искусственных и естественных системах, могут быть измерены и преобразованы. Информация в этом случае выполняет роль «рабочего тела», которое можно обрабатывать и хранить. Это «тело» существует во времени/пространстве, состоит из упорядоченных элементов (имеет структуру), взаимодействует с объектами реальности и с себе подобными информационными объектами. Соответственно, информационные взаимодействия — это не воздействие некой одной «информации» на другую, а физические преобразования, схема которых (соотношение составляющих, порядок, условия и т.п.) имеет некоторые специфические особенности, наиболее характерные из которых отражает данное в [5] обобщение феномена физической сущности информации: “*Информация,* в широком понимании этого термина, представляет собой объективное свойство реальности, которое проявляется в *неоднородности* (асимметрии) распределения материи и энергии в пространстве и в *неравномерности* протекания во времени всех процессов, происходящих в мире живой и неживой природы, а также в человеческом обществе и сознании. … В данном случае речь идёт о физической сущности так называемой «первичной», или «связанной» информации (по терминологии Л. Бриллюэна), которая порождается неоднородностью материальных или же энергетических объектов реального мира. Ведь именно эта информация является первоосновой для формирования так называемой «вторичной» информации, которая представляет собой некоторое «отражение» первичной информации и может быть отчуждена от своего первоисточника”.

С точки зрения формы существования (как макрообъекта) особенностью сущности «информация» является то, что она объединяет в себе содержание и форму. Причём это явление *одновременное* и *однопричинное* [6]. Содержание представляется языком (в физической основе которого знаковая система), обеспечивающим построение и использование смысловых, ценностных компонентов, а форма реализуется другим языком — знаковым представлением на физическом носителе. По существу, в одной структуре объединяются два типа описания, произведённые на языках, не имеющих между собой прямых смысловых связей. При этом содержание логически не зависит от его физического описания, что вполне соответствует принципу изофункционализма систем А. Тьюринга о возможности воспроизведения системы с данным набором функций на разных элементных базах (*принцип инвариантности информации по отношению к физическим свойствам её носителя)*. Отметим, что кодированные зависимости становятся средствами и элементами развития и самоорганизации: над ними, в свою очередь, могут формироваться новые кодовые зависимости более высокого уровня.

[1] Более подробно см. [1, 2].

Информация, как операбельный макрообъект — информационный объект (ИнфОб), определяется в [1] как образ оригинала, обладающий способностью к информационному взаимодействию с другими образами или оригиналами в некоторой области (в поле, в предметной области), связанной с оригиналом. Такой образ, точнее ИнфОб, — это *состояние физического носителя*[2], которое изменяется/измеряется и фиксируется (запись/чтение) по правилам языка предметной области (ПрО), причём разнообразие состояний должно быть не меньше разнообразия состояний оригинала в этой области (принцип необходимого разнообразия Эшби).

При этом любые ИнфОб'ы от физического сигнала до данных в вычислительной среде и текста в социуме имеют семантическую природу. Семантика сигнала определяется законами природы, семантика данных — структурой (логическими-физическими схемами БД), семантика текста — категориально-понятийной системой ПрО. Семантика (точнее смысл [7]) ИнфОб'а — это концептуальный образ его содержания (тоже ИнфОб), определяемый относительно некоторой ПрО, аспекта. Это значение, полученное отображением содержания (величины ИнфОб'а) на семантическое поле (спецификацию, в частности в виде онтологии ПрО или отдельного аспекта).

3. Свойства информации

Разнообразие определений понятия и сущности «информация», соответствует двум следующим концепциям. *Атрибутивная концепция* рассматривает информацию как фундаментальную естественно-научную категорию, как неотъемлемое свойство материи и энергии. *Функционально-кибернетическая* — как неотъемлемый элемент и/или функцию управляемых или самоуправляемых систем (технических, биологических, социальных), связанных с понятиями «целеполагание» и «сознание». Отметим, что эти концепции не являются взаимоисключающими: их соотношение скорее отражает взаимную связь материи и сознания. Отметим также, что свойства информации в большинстве публикаций вводятся и классифицируются именно в рамках этих концепций, при этом определяются содержательно и по-разному в зависимости от области знаний. Но безусловно понятно, что свойства информации как неэлементарной сущности (части наблюдаемой действительности) объективны, поскольку являются следствием закономерностей организации материи. С другой стороны, информация, как «рабочее тело» имеет собственные свойства — атрибутивные, а также свойства, проявляющиеся в результате взаимодействия с окружающей средой — так называемые прагматические свойства.

4. Фундаментальные свойства информации

Исходя из физической и логической природы информации, можно определить следующие основные фундаментальные (определяющие форму существования и особенности взаимодействия) свойства[3] информации и особенности информационных взаимодействий.

1. Информационные объекты и взаимодействия являются объективной и закономерной действительностью, существующей наряду с фундаментальными элементарными частицами и взаимодействиями. Информационные объекты и взаимодействия реализуются, в конце концов, в той же «элементной базе» фундаментальных частиц и взаимодействий, причём по отношению к ним выступают как надстройка в виде законов, констант и параметров порядка. Информация — это преходящее или зафиксированное в виде ИнфОб состояние, отражающее соотношения отображаемых сущностей в некотором

[2] Вследствие этого в обыденном понимании информация часто отождествляется с носителем.

[3] Отметим, что это феноменологические свойства — проявление неоднородности материи (в т.ч. сознания) и неравновесности взаимодействий, которые в свою очередь определяют форму существования информации и особенности информационных взаимодействий.

пространстве[4] отображения. То есть информация — это, можно сказать, сущность второго порядка сложности, предполагающая неатомарность формы существования — предполагающая наличие других объектов, составляющих или состояний (что соответствует понятию «свойство», рассматриваемому как то, что *проявляется* при взаимодействии, т. е. является вторичным, производным). Это действительно «не материя и не энергия», а их относительно устойчивые взаимозависимости[5], что, собственно, и обеспечивает возникновение неоднородностей и зависимостей, и которые, будучи зафиксированными в виде свойств, моделей, функций, законов, и будут отождествляться с традиционно понимаемой информацией.

2. Информация имеет двойственность состояния. Информационный объект до взаимодействия — это некоторое цельное неделимое образование, во время взаимодействия — это макрообразование, структура «квантов», которые также могут быть составляющими элементами или комбинациями и других, существующих или гипотетических объектов (т. е. значение информации - воспринимаемого ИнфОБ'а, зависит не только от его содержания, но и от воспринимающей стороны). Или, по аналогии с физикой, информация в процессах хранения и передачи проявляет свойства макрообъекта, а в процессах информационного взаимодействия с другими информационными объектами — волновые[6].

3. Информация имеет двойственную природу своего проявления: с одной стороны — это объект, который можно обрабатывать, а с другой стороны — это «сила», приводящая к изменениям в результате взаимодействия с другими объектами. Это переход от свойства элемента к элементу, обладающему свойством.

4. Информация возникает, когда:

1) есть некоторое количество физических неоднородностей, различимых в выбранном пространстве, проявляющихся при взаимодействии как свойства, события, т. е. есть появляется связанная информация;

2) существуют (как различимые неоднородности абстрактной природы) категории «пространства», «сходства», «различия», «взаимодействия», а также категории, свойственные деятельности, такие как «цель», «результат», «организация» и т.д., что обеспечивает появление свободной информации.

5. Сопряжённость информации (в случае документальной информации говорят, что она имеет семантическую природу).

Для связанной информации — это данность (ограничения, отражаемые законами), определяющая развитие (следующее состояние) в процессах реализации физических взаимодействий. Это не собственно отдельные части реальности (макрообъекты), находящиеся в различающихся состояниях, а характер неоднородности состояний. При

---

[4] Понятие «состояние» предполагает наличие объекта, имеющего это состояние. Для информации в привычном понимании («Информация — это сведения …») — это носитель информации, а в случае природы — физические частицы и их состояния, *играющие роль* информации и некоторым образом (обычно, в виде свойств) представляющие состояния, и соотношения физической реальности.

[5] Устойчивость (что выражается в наблюдаемой закономерности) взаимозависимостей и характера неоднородности определяется тем, что «…существуют законы более общие, чем физические, — законы информатики. Законы, определяющие, ограничивающие физические явления и процессы, законы, предшествующие физическим законам. Это, прежде всего, закон простоты сложных систем, закон сохранения неопределённости (информации), закон конечности информационных систем, закон необходимого разнообразия Эшби, теорема Геделя, закон Онсагера.» [8]. То есть в этом случае «информация, как всеобщее свойство материи» [9] будет первичной, изначальной — она *будет определять будущее* состояние материи. Но эта «изначальность» относительна: трансформации перманентны, а «первичность» здесь может приниматься только в субъективной предрасположенности считать сигнал (т. е. ИнфОб) причиной только потому, что его энергоматериальные характеристики сравнительно меньше энергоматериальности «результата».

[6] Для информации свойственны и такие, не рассматриваемые в данной статье, явления как дифракция и интерференция. (см., например, [10, 11].

этом включение наблюдателя и измерений, производимых в соответствии с определенной теорией, порождает свободную информацию — образ реальности в виде совокупности параметров и уравнений, отражающих различия состояний/взаимосвязей: формируется «физическая картина мира», представленная законами природы, открываемыми человеком.

Для свободной информации — информации в коммуникациях (как кода), описания, представляющие информацию — это разнообразие, которое некоторый объект содержит в себе о разнообразии другого объекта. Причём соответствующие данные, сигналы всегда связаны с контекстом — знанием, целями, ситуацией приемника, что определяет и форму её существования.

6. Действенность информации.

Причём для информационных взаимодействий данное свойство характеризуется следующими особенностями.

В информационных взаимодействиях участвуют имеющие природу образа ИнфОб'ы, которые обладают способностью избирательного (возможно параметрически управляемого) взаимодействия как с другим ИнфОб'ами, так и с оригиналами. Это позволяет реализовать эмерджентные или энергоэффективные замещающие процессы в предметной области.

Сигнальный характер — это основа для реализации «переключающего воздействия», для которого характерно, что энергетика порождения сигнала (информации) настолько мала, что соответствующее изменение состояния информационной среды не влечет существенного изменения состояния. ПрО, и наоборот — воздействие энергетически слабого сигнала приводит к существенному изменению состояния. ПрО, т. е. предметная область имеет естественные или искусственные точки бифуркации.[7]

Нелинейность (необратимость), обусловленная дискретным характером процесса отображения из одного пространства в другое, когда происходит редукция свойств оригинала при генерации образа, обусловленная выбором конкретного отображения.

Для реализации взаимодействия среда должна иметь механизмы сопряжения — интерфейсы, т. е. для реализации взаимодействия объекты должны иметь соответствующую общность природы (проявляющейся как общность свойств). Основой для этого является то, что разнообразие состава и структур, а также взаимодействий и соотношений в ПрО ограничено и определяется действующими законами Природы. Аналогично, категории и понятия (как чисто информационные объекты) неминуемо связаны со свойствами ПрО и, так или иначе, их отражают (положены в основу построения этих абстрактных объектов). Соответственно и образы, и оригиналы одинаково наблюдаются посредством обнаружения и измерения свойств в результате действия той же системы законов.

5. Атрибутивные свойства информации

Следствием выше представленных фундаментальных свойств являются следующие, характерные прежде всего для свободной информации, атрибутивно-функциональные свойства, отражающие способность информации к взаимодействию и преобразованиям.

1. Эмерджентность (наличие у системы свойств, не присущих её компонентам по отдельности). Эмерджентность связанной информации проявляется в ограничениях на возможные состояния, структуру и поведение некоторого множества элементов, последовательное или параллельно-последовательное взаимодействие которых и образуют систему (целостный объект, объединяющий элементы). Например, для кучи щебня, в

[7] Для таких точек ветвления характерно то, что незначительные изменения параметра, переход его через некоторый порог приводит к качественной перестройке процесса, новому состоянию равновесия. Это один из важнейших адаптационных механизмов, существующих в природе. Она предполагает перестройку организации — переход к новой структуре [12].

частности, объем или структура и есть эмерджентные свойства. Эмерджентность свободной информации определяется (является следствием) того, что представляющий эту информацию ИнфОб соотносится с другими объектами, выбираемыми в зависимости от цели, ситуации и т.п. И собственно этот выбор породит то или иное эмерджентное свойство (новый смысл). Примечательно что, представляющий информацию ИнфОб, как совокупность элементов (знакового уровня) также обладает связанной информацией, характеризующей уже зависимости языка и, в какой-то степени, ПрО. (Конструктивная модель возникновения эмерджентности информации, в частности, приведена в [13]).

Например, смысл предложения далеко не всегда сводится к смыслу слов, из которых оно состоит. В предложении как целом появляется нечто новое. Другим примером может являться совокупность научных фактов, которая при их систематизации позволяет обнаружить значительно больше важных свойств исследуемого объекта.

2. Неассоциативность и некоммутативность по существу является следствием свойства эмерджентности.

Для случая связанной информации очевидно в следствие того последовательность взаимодействия *неоднородных* элементов приведёт к разным результатам в зависимости от того какими свойствами будет обладать очередной «присоединяемый» элемент. Для той же кучи щебня, в частности, объем и структура (эмерджентные свойства) будут разными, если в начале будут использованы крупные фракции, а потом мелкие, или наоборот.

Для свободной информации получаемый (точнее, формируемый, понимаемый) смысл безусловно зависит от порядка получения элементов: каждое сочетание слов/выражений порождает смысл (эмерджентное свойство), который, в свою очередь будет определять контекст восприятия следующих элементов.

Например, если в случае подключения потребителем какого-либо электроприбора к сети, первой поступит команда «Включить в розетку», а потом — «Проверить, что сеть постоянного тока напряжением 12 вольт», то скорее всего прибор придёт в негодность, поскольку человека в быту в основном окружают высоковольтные сети переменного тока 220 вольт.

3. Куммулятивность информации отражает накопительный[8] характер действенности информации. То есть это свойство, которое обусловлено тем, что любое взаимодействие направлено. Направленность определяется либо действием соответствующего этому преобразованию закона (для связанной информации), либо (для свободной информации) выбором, обусловленным целями процесса. То есть это можно представить как сложение векторов, представляющих отображения этих ИнфОб'ов на предметную область.

Куммулятивность проявляется (т. е. реализуется посредством преемственности-наследования, концентрации информации) в двух следующих аспектах:

- временном — через *преемственность* информации, когда очередное информационное сообщение, отражающее состояние конкретной предметной области, так или иначе, включает содержание более ранних;
- смысловом — через *концентрацию* информации возможность представить объект сообщениями с разной детальности описания.

Преемственность информации, как форма кумулятивности, непосредственно связана с преемственностью в развитии науки, техники, производства и других сфер деятельности человека. Свойство преемственности проявляется в историческом, отраслевом и межотраслевом аспектах.

Историческая преемственность хорошо иллюстрируется словами И. Ньютона. На вопрос, как ему удалось сделать великие открытия, он ответил: «Я видел дальше других,

[8] Здесь *накопление* как *увеличение информации*, что надо отличать от *накопления источников информации,* поскольку росту объёмов публикаций присуща большая избыточность: кроме новых сведений, они зачастую содержат и дублирующие. Т. е. накопление скорее относится к носителям информации (кодам, представляющим информацию).

потому что стоял на плечах гигантов». Но преемственность здесь не означает простое восприятие прошлого, того, что накоплено в процессе исторического развития науки, а подразумевает его критическое освоение и переработку.

Отраслевая и межотраслевая преемственность информации отражает преемственность в развитии системы «наука — техника — производство», которая и ведёт к синтезу нового знания. С информационной точки зрения эта преемственность заключается в том, что объёмы информации, циркулирующей в сфере науки, превышают те объёмы, которые используются в технике, а информация в сфере техники больше её объёмов, используемых в производстве. Преемственность в развитии системы «наука-техника-производство» отражает также и закономерность опережающего характера развития науки перед техникой и техники перед производством.

Концентрация информации достигается в процессе, который иногда называют *свёртыванием* (путём идентификации, агрегирования или обобщения). Это свойство отражает закономерность развития научного и других видов человеческого знания. Формами концентрации информации являются, например, законы и категории науки, пословицы, поговорки и т.п.

Концентрация в сфере информационной деятельности проявляется, по крайней мере, в трех формах: документационной, фактографической и теоретическо-концептуальной [14].

Первая - документационная форма, где объектом является *документ,* реализуется путём идентификации, в результате чего появляется образ, представленный, главным образом внешними идентификационными признаками документа (автор, заголовок, выходные данные и т. п.) или его содержания. Такой семантический образ формируется, в частности, путём *реферирования*, которое предполагает извлечение из документа основных положений содержания и их представление в виде реферата. Другой путь — *классифицирование* и *индексирование*, в результате которого на некотором специальном, обычно искусственном, языке отражаются с большей или меньшей степенью глубины и полноты тематические или фактографические признаки содержания документов. Все эти виды концентрации информации предполагают неизбежные потери информации, поскольку в этом случае отражаются только основные аспекты содержания документов.

Вторая форма — фактографическая, ориентирована не на отдельный документ, а на совокупность фактов или документов по определенной теме или проблеме. В качестве таких форм могут выступать реферативные обзоры, фактографические информационные картотеки, тематические подборки и т. д.

Третья, высшая форма — теоретико-концептуальная концентрация информации реализуется преобразованием, которое позволяет представить смысл на более высоком уровне обобщения. Примером являются аналитические формы представления математических зависимостей, законы, теоремы, позволяющие выводить различные следствия и т. д.

То есть свойство концентрации проявляется в том, что по мере накопления (в отрасли, группе отраслей науки или техники и т. д.) определенных единиц информации они имеют тенденцию к объединению в более информационно-ёмкие формы.

Рассеяние информации означает, что информация, которая была бы полезной для решения данной проблемы, может оказаться в документах, относящихся к другим предметным областям.

Для ИнфОб’ов, которые представляют (фиксируют с помощью языка и обеспечивают распространение информации), рассеяние является следствием свойств понятийно-знаковой системы и предметной области (как системы в контексте Общей теории систем). Предметные области могут иметь (и имеют) разные системные основания: по-разному выделяют и соотносят объекты, связи и свойства в разных ПрО, что предопределяет существование «предметного рассеяния». Аналогично, понятийные системы могут быть разными по составу и структуре и по-разному отражать ПрО, что предопределяет существование «семантического рассеяния». И, соответственно, можно использовать

разные слова для выражения одного и того же смысла, что предопределяет лексическое рассеяние.

Следствием свойства рассеяния является *дублирование информации*, что в свою очередь предопределяет рассредоточение информации, полезной для некоторой ПрО по документам, относящимся к разным предметным областям.

4. Полипотентность (межотраслевой характер по [14]) — это возможность использования информации, созданной для решения какой-либо одной задачи, для решения других, в том числе других отраслях знаний/деятельности.

5. Гранулированность информации, которая в информационных технологиях отражается следующей иерархией:

- сигнал — это различаемое/воспринимаемое изменение (или измерение состояния) наблюдаемого физического объекта, где переносчиком (собственно ИнфОб'ом) является энергия;
- данные — определенные состояния некоторого определенного носителя (его части) зафиксированные на нем и считываемые с него, где переносчиком является сигнал, приводящий к соответствующему изменению состояния носителя («определённость» означает, что данные - это не только «состояние носителя», но и метаданные, указывающие метод кодирования, т. е. соответствие «сигнал-состояние», а также метод записи/чтения);
- информация — определенная некоторым контекстом совокупность данных, т. е. <агрегаты данных+метаинформация>, где переносчиком являются данные;
- знания — опредмеченная информация, т. е. связанная (отражённая и конкретизированная) с наличным знанием и/или практическим опытом, зафиксированная в индивидуальном сознании (неявные знания) или в виде сообщения (явные, обобществлённые знания), где переносчиком является информация (по существу, обобществлённые знания обретают роль данных, отражающих некоторую конкретную данность познания, и которые будут использоваться как потенциально полезная информация, т. е. она возможно будет связываться с другими контекстами и другой информацией, и таким образом будет использована для построения новых знаний).

Заметим, что в области социальных коммуникаций этому соответствуют гранулы *явление — факт (сведения) — сообщение — знание (обобщение, объяснение)*. В области онтологического представления этому соответствуют гранулы *объекты (сущности, отношения, свойства) — элементарный факт — ситуативный факт — завершённый факт*, которые в свою очередь в области рациональной деятельности соответствуют гранулам *производственный базис — операция — процесс (действие) — целесообразная работа (деятельность)*. А в области обработки данных — это гранулы *элементная база - данные (элементарные типы) — структуры данных — объекты — программы*.

Отметим также, что физическим носителем (базовым переносчиком, элементной базой) в любом случае является физическая среда, в которой реализуются природные явления, в частности, рассматриваемые как изменения состояния носителя, т. е. неоднородности.

6. Информирование «по ассоциации» — это условное связывание объектов деятельности: по смежности, сходству, контрасту, а также по некоторым смысловым схемам (например, ассоциации «вид — род», «часть — целое» и «причина — следствие»).

6. Прагматические свойства информации

Для свободной информации, для которой характерна целенаправленность её получения и использования, исходя из обобщённой схемы ОД-ИД, где информация формируется, сохраняется, трансформируется и используется, можно определить следующие прагматические (внешние относительно неё) свойства, имеющие относительный характер,

и которые используются и, обычно, количественно или качественно оцениваются в процессе целенаправленной деятельности.

1. Транслируемость и тиражируемость — свойство, отражающее возможность существования в разных формах и экземплярности, как следствие свойства независимости от носителя.

2. Актуальность — свойство, отражающее соответствие задачам, значимым (актуальным) для ПрО в настоящее время. Может определяться через отображение информации на образ задачи.

3. Новизна — свойство, отражающее наличие семантических блоков / связей, отсутствующих или иных по отношению к текущему состоянию ПрО. Степень новизны может определяется через отображение информации на образ ПрО.

4. Старение — свойство, обусловленное: (1) появлением новых решений той же задачи — созданием нового знания; (2) рассогласованием с семантикой ПрО и/или языка в следствие их развития — возникает новое представление того же знания.

Старение информации может носить как абсолютный, так и относительный характер. Устаревшей считается и информация, которая с появлением новой информации оказалась недостоверной. Относительный характер старения информации можно рассматривать с точки зрения её новизны не только по временным параметрам, но и по отношению к совокупному или индивидуальному знанию. Если, например, в физике открыта новая частица, то информация об этом будет новой и для физики, и для любого, кто об этом прочитает или услышит. В то же время сообщение об известных ранее частицах в физике (по отношению к совокупному знанию этой науки) будет считаться относительно устаревшим, а по отношению к отдельному индивидуальному знанию конкретного человека, например, для школьника, может быть новым.

Таким образом, при оценке старения информации следует учитывать связь между собственно информацией и её потребителем.

5. Объективность/субъективность (степень зависимости от информационного состояния приемника информации, чьего-либо мнения) — определяется «отражателем»: *объективная* информация, получаемая отображением оригинала на ПрО, *субъективная* — на образ ПрО, имеющийся у субъекта.

6. Полнота — свойство, которое может оцениваться (1) долей присутствия семантических элементов (конструктивных фактов) в образе темы; (2) долей конструктивных (отвечающих цели) аспектов по отношению к образу темы/цели.

7. Точность — свойство, которое может оцениваться (1) детальностью (уровнем общности/детальности представления семантических элементов; (2) долей релевантных сообщений (соответствующих теме/цели) в подборке.

8. Специфичность — свойство, которое может оцениваться степенью моно / политематичности сообщения / выборки.

9. Достоверность — свойство, которое может оцениваться (1) подтверждаемостью — смысловым соответствием другими ИнфОб’ами; (2) не противоречивостью с ожидаемым результатами.

10. Адекватность — свойство, которое может оцениваться уровнем полноты / точности / достоверности, необходимый и достаточный для достижения цели.

11. Полезность, ценность — свойство, которое может оцениваться (1) как «прибыль» от снижения стоимости реализации (иначе, стоимость повторного изобретения); (2) как «стоимостью» потерь от передачи конкурентам.

12. Значимость — качественный показатель полезности информации для достижения цели.

7. Заключение

Таким образом, можно констатировать[9], что информация — это проявление всеобщего свойства материального мира быть *определенным* (существовать и изменяться в соответствии с законами природы), быть *определяемым* (воспринимаемым и идентифицируемым) и быть *определяющим* (способным изменять состояние некоторого целевого объекта).

Информация, как объективная реальность, имеет вполне определенные свойства. Атрибутивные свойства — «собственные», внутренние свойства информации являются следствием фундаментальных её свойств, которые непосредственно обусловлены законами природы и информатики. При этом прагматические свойства имеют чисто информационную природу — это информация — соотнесение данных, представляющих (в зависимости от типа свойства) содержание, с теми или иными ИнфОб'ами, представляющими внешний объект — цель, задачу, процесс и т.п.

Информация наравне с материей и энергией — это основа и средство организации и изменений. В целенаправленной деятельности человека информация (точнее, работа с информацией, реализация её свойства действенности) — это средство *замещения* основной (в физической среде) деятельности человека. С точки зрения времени можно сказать, что назначение информации — воспроизводство вида/знаний в настоящем и связь с прошлым и/или с будущим. При этом рост знаний реализуется посредством преемственности-наследования, концентрации и рассеяния информации. Важность концентрации информации в процессе развития человеческого знания обобщается словами Н. Н. Моисеева: «Любое человеческое знание начинается с накопления фактов, с помощью наблюдения или направленного эксперимента. Но не превращённая в систему, река новых знаний не утолит жажду. Пока хаос новых фактов не структурирован, пока человек не может окинуть взглядом явления в целом, он не может эти знания использовать для практики. Поэтому второй этап — это переработка информации, представление её в такой форме, которая уже может быть «переварена» человеком. Ну а третий этап — это возвращение к практике, использование знаний для тех целей, ради которых они были созданы» [16].

Литература

[1] Максимов Н. В. Информация и знания: природа, концептуальная модель // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2010. № 7. С. 1–10.

[2] Лебедев А. А., Максимов Н. В. Аналогии в физике и обработке информации //Научно-техническая информация. Серия 2: Информационные процессы и системы. 2020. № 10. С. 1–11.

[3] Гуревич И. М. О физической информатике: предпосылки и основные результаты. М.: ЛЕНАНД, 2014. 160 с.

[4] Quastler H. The emergence of biological organization. New Haven: Yale University Press, 1964. 83 с.

[5] Колин К. К. Природа информации и философские основы информатики // Открытое образование. 2005. № 2. С. 43–51.

[6] Лекторский В. А. и др. Информационный подход в междисциплинарной перспективе (материалы «круглого стола») // Вопросы философии. 2010. № 2. С. 84–112.

[7] Лебедев А. А., Максимов Н. В. Об одном подходе к определению семантической информации // Компьютерная лингвистика и вычислительные онтологии. 2022. Вып. 6. С. 30–40.

[9] Расширяя и уточняя формулировку, приведённую в [15].

[8] Гуревич И. М. Законы информатики — основа строения и познания сложных систем. 2-е изд. М.: ТОРУС ПРЕСС, 2007. 400 с.
[9] Гуревич И. М., Урсул А. Д. Информация — всеобщее свойство материи: характеристики, оценки, ограничения. 2-е изд. М.: КД «ЛИБРОКОМ», 2013. 312 с.
[10] Лебедев А. А., Максимов Н. В., Смирнова Е. В. Семантический сдвиг термина: анализ зависимостей и квантомеханическая модель. // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 2016. № 2. С. 14–22.
[11] Lebedev A. A., Maksimov N. V., Smirnova E. V. The quantum-mechanical approach to construction of quantitative assessments of some documentary information properties (on example of nuclear knowledge) // Journal of Physics: Conference Series. IOP Publishing, 2017. Vol. 781. № 1. Art. 012059.
[12] Моисеев Н. Н. Человек. Среда. Общество. 2-е. изд. М.: ЛЕНАНД, 2021. 248 с.
[13] Varley T. F., Hoel E. Emergence as the conversion of information: A unifying theory //Philosophical Transactions of the Royal Society A.2022. Т. 380. №. 2227. С. 20210150. URL: https://arxiv.org/pdf/2104.13368.pdf (дата обращения: 16.04.2023).
[14] Муранивский Т. В. Теоретические основы научно-технической информации. М.: МГИАИ, 1982. 160 с.
[15] Коллендер Б. Информация об информации. URL: http://www.elektron2000.com/kollender_0225.html (дата обращения: 16.04.2023).
[16] Моисеев Н. И. Предисловие //Число и мысль. М.: Знание, 1977. 176 с.

Towards a Constructive Definition of Properties of Information

Nikolay V. Maksimov, Alexander A. Lebedev

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)

The work is devoted to the nature and substantiation of the properties of information. Information is considered as a universal property of the material world to be defined (to exist and change in accordance with the laws of nature), to be definable (perceived and identifiable) and to be determinative (able to change the state of the target object). Based on the nature of information accepted by the authors as a special form of matter (connections or dependencies of objects, phenomena or thought processes), fundamental, pragmatic and attributive properties are introduced. It is shown that the properties under consideration may have a common fundamental origin, which makes it possible to constructively define approaches to evaluating information.

Keywords: fundamental properties of information, attributive properties of information, pragmatic properties of information, free information, related information, information interactions

Reference for citation: Maksimov N. V., Lebedev A. A. Towards a Constructive Definition of Properties of Information // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 42–53. DOI: 10.17586/2541-9781-2024-7-42–53

Reference

[1] Maksimov N. V. Informaciya i znaniya: priroda, konceptual'naya model' // Nauchnotekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2010. № 7. S. 1–10. (in Russian)

[2] Lebedev A. A., Maksimov N. V. Analogii v fizike i obrabotke informacii //Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2020. № 10. S. 1–11. (in Russian)
[3] Gurevich I. M. O fizicheskoj informatike: predposylki i osnovnye rezul'taty. M.: LENAND, 2014. 160 s. (in Russian)
[4] Quastler H. The emergence of biological organization. New Haven: Yale University Press, 1964. 83 p.
[5] Kolin K. K. Priroda informacii i filosofskie osnovy informatiki // Otkrytoe obrazovanie. 2005. № 2. S. 43–51. (in Russian)
[6] Lektorskij V. A. i dr. Informacionnyj podhod v mezhdisciplinarnoj perspektive (materialy «kruglogo stola») // Voprosy filosofii. 2010. № 2. S. 84–112. (in Russian)
[7] Lebedev A. A., Maksimov N. V. Ob odnom podhode k opredeleniyu semanticheskoj informacii // Komp'yuternaya lingvistika i vychislitel'nye ontologii. 2022. Vyp. 6. S. 30–40. (in Russian)
[8] Gurevich I. M. Zakony informatiki — osnova stroeniya i poznaniya slozhnyh sistem. 2-e izd. M.: TORUS PRESS, 2007. 400 s. (in Russian)
[9] Gurevich I. M., Ursul A. D. Informaciya — vseobshchee svojstvo materii: harakteristiki, ocenki, ogranicheniya. 2-e izd. M.: KD «LIBROKOM», 2013. 312 s. (in Russian)
[10] Lebedev A. A., Maksimov N. V., Smirnova E. V. Semanticheskij sdvig termina: analiz zavisimostej i kvantomekhanicheskaya model'. // Nauchno-tekhnicheskaya informaciya. Ser. 2: Informacionnye processy i sistemy. 2016. № 2. S. 14–22. (in Russian)
[11] Lebedev A. A., Maksimov N. V., Smirnova E. V. The quantum-mechanical approach to construction of quantitative assessments of some documentary information properties (on example of nuclear knowledge) //Journal of Physics: Conference Series. IOP Publishing, 2017. Vol. 781. №. 1. Art. 012059.
[12] Moiseev N. N. Chelovek. Sreda. Obshchestvo. 2-e. izd. M.: LENAND, 2021. 248 s. (in Russian)
[13] Varley T. F., Hoel E. Emergence as the conversion of information: A unifying theory // Philosophical Transactions of the Royal Society A. 2022. Vol. 380. №. 2227. Art. 20210150. URL: https://arxiv.org/pdf/2104.13368.pdf (access date: 16.04.2023).
[14] Muranivskij T. V. Teoreticheskie osnovy nauchno-tekhnicheskoj informacii. M.: MGIAI, 1982. 160 s. (in Russian)
[15] Kollender B. Informaciya ob informacii. URL: http://www.elektron2000.com/kollender_0225.html (access date: 16.04.2023). (in Russian)
[16] Moiseev N. I. Predislovie //Chislo i mysl'. M.: Znanie, 1977. 176 s. (in Russian)

Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке

Д. В. Мельничук, А. В. Носкина

Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского

melnichukdv@sgu.ru, noskinaav@sgu.ru

Аннотация

В данном исследовании сравниваются основные NLP-модели, такие, как mBART, T5 и GPT-3, которые в своей основе имеют архитектуру трансформеров, т. е. механизм «внимания», кодирующий, декодирующий и нормализующий слои. Данные предобученные модели на задаче суммаризации русского текста были использованы для суммаризации научных статей на русском языке. Для выявления лучшей модели на данном классе задач в исследовании был использован набор данных, включающий в себя тексты научных статей и соответствующие им авторские аннотации на русском языке. Далее, стандартными для задачи суммаризации статистическими метриками, такими, как семейство метрик ROUGE (ROUGE-1, ROUGE-2 и ROUGE-L), а также BLEU и Perplexity, находилась наиболее эффективная модель в рамках поставленной задачи, т. е. сравнивались по отдельности сгенерированные варианты аннотаций с авторской. Полученные результаты имеют практическую ценность, так как суммаризация текста является важной задачей в области обработки естественного языка.

Ключевые слова: NLP, суммаризация, mBART, T5, GPT-3

Библиографическая ссылка: Мельничук Д. В., Носкина А. В. Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 54–59. DOI: 10.17586/2541-9781-2024-7-54–59

1. Введение

Основной целью данной работы является ответ на вопрос: «Какая из NLP-моделей суммаризации (Natural Language Processing, NLP — Обработка текстов на естественном языке) наиболее оптимально работает в контексте академической литературы на русском языке»? Под суммаризацией текста понимается процесс автоматического сокращения объёма исходного текста путём извлечения наиболее важных и существенных идей, фактов и информации, а также представления в форме краткого и сводного текста, который сохраняет основные аспекты исходного материала.

Для сравнения эффективности разных типов предобученных (pre-trained) NLP-моделей использовался набор статей из открытой научной электронной библиотеки CyberLeninka, из части массива доступных данных, были использованы тексты (text) научных статей и соответствующие аннотации (annotation) их авторов на русском языке. Всего 825 статей, среди которых область наук и тип журнала брались в случайном порядке.

2. Модели и данные

Для исследования были выделены наиболее популярные, по версии ресурса HuggingFace Hub, открытые NLP-модели суммаризации текста, обученные на одном и том же корпусе новостных текстов на русском языке (Gazeta) [5].

Языковая модель GPT-3 (Generative Pre-trained Transformer) использует механизмы трансформеров для анализа контекста и генерации последовательностей слов, учитывая вероятность каждого следующего слова на основе предыдущих слов в тексте. Модель также способна выполнить различные задачи, такие, как ответы на вопросы, перевод текстов на другие языки и создание текстовых статей. В нашем исследовании была использована GPT-3 модель, обученная под задачу суммаризации, под кодовым названием модели на ресурсе HuggingFace Hub: RuGPT3MediumSumGazeta [6].

Модель T5 (Text-to-Text Transfer Transformer) также использует архитектуру трансформеров и обучается на задачах преобразования текста в текст. На вход подаётся задание и исходный текст, а затем генерируется выходной текст, решающий поставленную задачу. Модель обучается на широком спектре задач, включая машинный перевод, генерацию текста, ответы на вопросы, классификацию текста и многое другое. Для нашего исследования была использована базовая модель RuT5Base, обученная на новостных текстах на русском языке (Gazeta) под задачу суммаризации: RuT5SumGazeta [7].

Модель mBART (multilingual Bidirectional and Auto-Regressive Transformer) использует технологию мультиязычного перевода, обученную на большом количестве текстов на разных языках. Каждый язык представлен в виде уникального кода, и модель может работать с несколькими языками одновременно. При обучении модели mBART используется подход обучения с подкреплением, который позволяет модели улучшать свой перевод по мере того, как она получает обратную связь. Архитектура трансформеров позволяет данной модели учитывать контекст и зависимости между словами в предложении. Аналогичным образом использована базовая модель mBART, обученная на новостных текстах на русском языке (Gazeta) под задачу суммаризации: MBARTRuSumGazeta [8].

3. Метрики

Для оценки и сравнения языковых моделей используются два подхода.

Первый подход — это внешняя оценка (External evaluation), при которой оценивание модели происходит за счёт решения с её помощью задачи, на которую она рассчитана, и дальнейший анализ итоговых показателей потерь/точности, а также является лучшим подходом к оцениванию моделей, так как это единственный способ реально оценить, как разные модели справляются с интересующей нас задачей. Однако реализация данного подхода может потребовать больших вычислительных мощностей, его применение может оказаться медленным, так как для этого нужно обучение всей анализируемой системы (BLEU, ROUGE — это внешняя оценка).

Второй же подход - это внутренняя оценка (Internal evaluation), которая производит оценку самих языковых моделей, без учёта конкретных задач, для решения которых их планируется использовать; она является не столь информативной для понимания качества работы модели на конкретной задаче, как внешняя, но, если необходимо провести итоговую оценку модели, то данный подход может быть весьма эффективным для быстрого сравнения моделей (Perplexity — это внутренняя оценка).

В данной работе были использованы метрики: BLEU, семейство метрик ROUGE и Perplexity.

Метрика BLEU (Bilingual Evaluation Understudy) — это алгоритм оценки качества машинной генерации текста (в том числе перевода), основанный на сравнении выходных текстов, т. е. сгенерированных (predictions) с известными, эталонными (references) текстами. Сам подход заключается в сравнении двух вариантов текста, по совпадению слов и их

расположению, также это называют схожести n-грамм (последовательности n слов). В итоге получается количественная оценка соответствия между результатом работы NLP-модели и результатом работы человека: чем ближе машинная генерация к исходному тексту человека, тем он лучше - такова основная идея BLEU. Метрика BLEU включает корректировки весов, такие, как фактор бонуса на основе bi-граммов и сглаживание на основе ковариационной матрицы предложений, чтобы справиться с некоторыми из проблем данного подхода [3].

Пусть C — множество слов сгенерированного текста, R — множество слов эталонного текста, соответственно c_i и r_i — это i-е слова этих множеств (списков). Пусть n — максимальная длина n-грамм, которые мы рассматриваем. Тогда BLEU оценивает качество сгенерированного текста с путём вычисления взвешенного гармонического среднего точности n-грамм:

$$\text{BLEU} = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \cdot BP$$

$$p_n = \frac{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)_{r_i}}{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)}$$

$$\text{BP} = \begin{cases} 1, & if\ c > r \\ e^{(1-r/c)}, & if\ c \leq r, \end{cases}$$

- p_n— точность n-грамм (последовательности n слов) в сгенерированном тексте;
- N — количество предложений в корпусе, c_i — i-я сгенерированная фраза, r_i — эталонный перевод для i-й кандидатской фразы;
- $Count(n-gram)_{r_i}$ — количество вхождений n-грамм в r_i;
- $Count(n-gram)$ — количество вхождений n-грамм в c_i;
- BP— штрафной фактор.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) — это набор показателей (семейство метрик) для оценки автоматического суммирования текстов (в том числе машинного перевода), основанный на сравнении n-грамм сгенерированного (predictions) текста с n-граммами эталонных (references) текстов. Основная идея метрики ROUGE заключается в сравнении двух текстов и подсчёте базовых единиц (n-грамм, т. е. последовательностей слов и количества пар слов). В результате получаем количественную оценку работы NLP-модели, которая показывает, насколько сгенерированный текст совпадает с текстом, составленным человеком (экспертом). В отличие от BLEU, ROUGE использует как полноту (recall), так и точность (precision) для сравнения сгенерированных текстов с эталонными текстами, составленными человеком [2].

В ROUGE-1 сравниваются единицы (слова) между сгенерированным и эталонным текстами. В ROUGE-2 сравниваются последовательности из двух слов, взятых из сгенерированного и эталонного текста. В ряде источников ROUGE-1 и ROUGE-2 могут обозначаться общей записью ROUGE-N. ROUGE-L, в свою очередь, не сравнивает n-граммы, а обрабатывает тексты и ищет самую длинную последовательность (LCS), которая является общей для двух текстов, а затем измеряет её длину.

Пусть S — сгенерированный текст, G — эталонный текст, соответственно s_i и g_i — это i-е слова в S и G. ROUGE-N оценивает качество генерации из S путём вычисления точности совпадения слов в S с G, подсчитывает количество совпадающих (co-occurrences) n-грамм (для ROUGE-1 это одно слово, для ROUGE-2 это последовательность из двух слов), найденных как в выходных данных модели, так и в эталоне, а затем делит это число на общее количество n-грамм в S:

$$\text{ROUGE} - 1 = \frac{\sum_i Count_{natch}(s_i)}{\sum_i Count(s_i - 1, s_i)}$$

$$\text{ROUGE} - 2 = \frac{\sum_i Count_{match}(s_i - 1, s_i)}{\sum_i Count(s_i - 1, s_i)}$$

$$\text{ROUGE} - \text{L} = \frac{LCS(S,G)}{max(|S|,|G|)},$$

- *S* — сгенерированный текст, *G* — эталонный текст, s_i — i-е слово в *S*;
- $Count_{match}(s_i)$ — число вхождений слова s_i в обоих текстах *S* и *G;*
- $Count(s_i)$ —общее число n-грамм s_i в *S;*
- *LCS(S, G)* — самая длинная общая последовательность слов в *S* и *G;*
- *max(|S|, |G|)* — максимальное значение между количеством слов в *S* и *G*.

Метрика Perplexity в языковых моделях используется для оценки того, насколько хорошо модель может предсказать следующее слово в тексте. Для хорошей NLP-модели метрика Perplexity будет давать высокие вероятности синтаксически корректным предложениям, а предложениям некорректным (или очень редко встречающимся) — низкие вероятности. При условии, что набор данных состоит из корректных предложений, лучшей моделью будет та, которая назначит наивысшую вероятность этому тестовому набору, что означает то, что модель обладает хорошим пониманием того, как устроен язык [4].

$$\text{P(W)} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i|w_{i-1,\dots,}w_{i-n+1})}$$

$$\text{Perplexity(W)} = \sqrt[N]{\frac{1}{P(w_i, w_2, \dots, w_N)}},$$

- W — набор слов в предложении;
- $p(w_i|w_{i-1})$ — вероятность того, что слово i будет следовать за словом i — 1.

4. Методология

Методология исследования выглядит следующим образом: изначально выделяются данные, включающие в себя авторскую аннотацию и три варианта автоматически сгенерированных для каждого отдельно взятого текста статьи NLP-моделями суммаризаций, а затем проводится сравнение на близость сгенерированных текстов с исходным текстом авторской аннотации.

В результате для каждой исходной статьи формируется оценка по пяти метрикам для трех NLP-моделей. Далее, имея данные результаты, находится среднее по каждому показателю, что и является итоговой оценкой эффективности работы данных NLP-моделей на задаче суммаризации.

5. Результаты и выводы

В результате, на задаче суммаризации академических текстов на русском языке, наилучшим образом проявила себя модель T5, которая показала наибольшую эффективность на основе статистических метрик. Данный результат может быть обусловлен тем, что модель T5 обеспечивает лучшую производительность и точность на задаче суммаризации текста, благодаря своей более общей и гибкой архитектуре, а также улучшенным параметрам и настройкам, в отличие от mBART и GPT-3.

Таблица. Результаты исследования на всем объёме данных

Модель	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity
mBART	9.1	**28.3**	13.3	27.0	39.84
T5	**10.1**	24.2	**13.4**	**27.6**	**30.7**
GPT-3	4.3	21.1	6.9	19.7	42.1

В дальнейших исследованиях планируется расширение проверочного набора статей, сравнение большего числа моделей, разделение проверочного датасета на области наук и сравнение результатов дискретно по научным областям.

Исследование выполнено за счёт гранта Российского научного фонда № 22-18-00153 «Образ СССР в исторической памяти: исследование медиастратегий воспроизводства представлений о прошлом в России и зарубежных странах» (https://rscf.ru/project/22-18-00153/).

Литература

[1] Gusev I. Dataset for Automatic Summarization of Russian News // Artificial Intelligence and Natural Language. AINL 2020 / Filchenkov A., Kauttonen J., Pivovarova L. (eds). Communications in Computer and Information Science. Vol. 1292. 2020. P. 122–134.

[2] Lin C. ROUGE: a package for automatic evaluation of summaries // Text Summarization Branches Out / Association for Computational Linguistics. Barcelona. 2004. P. 74–81.

[3] Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation // 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311–318.

[4] Nallapati R., Zhai F., Zhou B. SummaRuNNer: a recurrent neural network-based sequence model for extractive summarization of documents // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligenc. 2017. P. 3075–3081.

[5] Gusev I. Gazeta - Dataset for Automatic Summarization of Russian News // Hugging Face. 2021. URL: https://huggingface.co/datasets/IlyaGusev/gazeta (дата обращения: 17.05.2023).

[6] Gusev I. RuGPT3MediumSumGazeta — Model for abstractive summarization for Russian based on rugpt3medium // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta (дата обращения: 17.05.2023).

[7] Gusev I. RuT5SumGazeta — Model for abstractive summarization for Russian based on rut5-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (дата обращения: 17.05.2023).

[8] Gusev I. MBARTRuSumGazeta — Model for abstractive summarization for Russian based on rumbart-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta (дата обращения: 17.05.2023).

Comparison of NLP-models on the Task of Summarizing Academic Texts in Russian Language

Dmitriy V. Melnichuk, Anastasia V. Noskina

Saratov State University

This study compares major NLP models such as mBART, T5 and GPT-3, which have at their core a transformer architecture, i.e., an "attention" mechanism that encodes, decodes and normalizes layers. These pre-trained models on the task of summarizing Russian text, were used to summarize scientific articles in Russian. To identify the best model on this class of tasks, the study used a dataset including the text of scientific articles and their corresponding author's annotations in Russian. Then, using standard statistical metrics, such as the ROUGE family of metrics (ROUGE-1, ROUGE-2 and ROUGE-L), BLEU and Perplexity, the most effective model was found for the task, i.e., the generated annotation variants were compared separately with the author's annotation. The results obtained are of practical value, as text summarization is an important task in the field of natural language processing.

Keywords: NLP, summarization, mBART, T5, GPT-3

Reference for citation: Melnichuk D. V., Noskina A. V. Comparison of NLP-models on the Task of Summarizing Academic Texts in Russian Language // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St.
Petersburg: ITMO University, 2024. P. 54–59. DOI: 10.17586/2541-9781-2024-7-54–59

Reference

[1] Gusev I. Dataset for Automatic Summarization of Russian News // Artificial Intelligence and Natural Language. AINL 2020 / Filchenkov A., Kauttonen J., Pivovarova L. (eds).
Communications in Computer and Information Science. Vol. 1292. 2020. P. 122-134.[2]
Lin C. ROUGE: a package for automatic evaluation of summaries // Text Summarization Branches Out / Association for Computational Linguistics. Barcelona. 2004. P. 74–81.
[3] Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation // 40th Annual Meeting of the Association for Computational Linguistics. 2002. P. 311-318.
[4] Nallapati R., Zhai F., Zhou B. SummaRuNNer: a recurrent neural network-based sequence model for extractive summarization of documents // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligenc. 2017. P. 3075–3081.
[5] Gusev I. Gazeta - Dataset for Automatic Summarization of Russian News // Hugging Face. 2021. URL: https://huggingface.co/datasets/IlyaGusev/gazeta (access date: 17.05.2023).
[6] Gusev I. RuGPT3MediumSumGazeta — Model for abstractive summarization for Russian based on rugpt3medium // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta (access date: 17.05.2023).
[7] Gusev I. RuT5SumGazeta — Model for abstractive summarization for Russian based on rut5-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (access date: 17.05.2023).
[8] Gusev I. MBARTRuSumGazeta — Model for abstractive summarization for Russian based on rumbart-base // Hugging Face. 2021. URL: https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta (access date: 17.05.2023).

Выявление скрытых закономерностей в реакции общества на бренд: анализ привлекательности названия методами машинного обучения

Е. М. Татур, Е. В. Клименко

Санкт-Петербургский государственный университет

katiandkate@gmail.com, ekaterinaklimenko0700@gmail.com

Аннотация

В статье рассматривается проблема выявления скрытых закономерностей в реакции общества на бренды посредством оценки привлекательности названия методами машинного обучения. Авторы проводят анализ большого объёма данных, включающих названия различных торговых марок и информацию об их популярности среди потребителей, с применением математических методов и алгоритмов машинного обучения. В ходе исследования проведены эксперименты со статическими предсказывающими моделями Word2Vec. Результаты исследований доказывают, что разрабатываемый авторами подход позволяет находить названия, удовлетворяющие требованиям запросам сферы маркетинга.

Ключевые слова: классификация текстов, бренд, нейминг

Библиографическая ссылка: Татур Е. М., Клименко Е. В. Выявление скрытых закономерностей в реакции общества на бренд: анализ привлекательности названия методами машинного обучения // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). ─ СПб: Университет ИТМО, 2024. С. 60–66. DOI: 10.17586/2541-9781-2024-7-60–66

1. Введение

Название бренда является одним из возможных способов привлечения внимания клиента к компании. Коммуникативная эффективность названия оказывает прямое влияние на возможность выхода торговой марки на рынок, а также успешность её продвижения. Наименование бренда представляет собой его базовый, наиболее константный атрибут, смена которого крайне нежелательна, так как порождает множество маркетинговых проблем. Маркетинговая значимость названия торговой марки объясняется большим количеством требований, предъявляемых к результату [1]. Имя бренда должно точно отображать заключённую в него идею и ценности компании, ясно выражать его конкурентные преимущества, а также положительно эмоционально ассоциироваться с объектом нейминга. В теории нейминга существуют различные алгоритмы разработки названия торговой марки.

2. Что такое нейминг?

Нейминг — процесс создания названия для компании, продукта, услуги, бренда или любого другого объекта. Он включает в себя анализ целевой аудитории, конкурентов, маркетинговых стратегий и т.д. Цель нейминга — создать уникальное, запоминающееся и привлекательное название, которое отражает суть объекта и привлекает внимание

потенциальных клиентов [2]. Названия могут иметь эмоциональный и символический аспект. Нейминг также может вызывать определенные ассоциации, чувства и атмосферу. Например, название «Coca-Cola» может вызывать ассоциации с безалкогольными напитками и праздниками.

Лингвистическая природа нейминга также связана с грамматическими и синтаксическими аспектами языка. Названия могут создаваться с помощью различных морфологических и синтаксических процессов словообразования, таких как суффиксация, префиксация, композиция и т.д. Например, название «Photography» образовано путём добавления суффикса «-graphy» к слову «photo», а наименование «ReDesign» появилось путём присоединения префикса «-re» к слову «design».

Лингвистическая природа нейминга определяет разнообразие способов формирования неймов, включая следующие:

- вторичная номинация — создание нейма путём использования уже существующего слова;
- словообразовательная модель — создание нейма путём применения продуктивной словообразовательной модели;
- фонетическая модель — создание нейма путём изменения произношения или ударения известного слова;
- синтаксическая модель — создание нейма путём сочетания слов в определенной синтаксической структуре и т.д.

Таким образом, лингвистическая природа именования включает в себя изучение процессов создания имён, их семантики и символического значения, а также их функции в коммуникации.

Алгоритм разработки названия торговой марки может быть заметен при изучении структуры названия. Например, оно может содержать название сферы и местоположение компании или филиала, что говорит о прямом конструировании названия торговой марки. Подобные действия удлиняют название, что может влиять на его запоминаемость или благозвучие.

3. Процесс проведения исследования

В рамках исследования рассматриваются следующие алгоритмы: название, основанное на ключевых словах, связанных с брендом и его ценностями; поиск синонимов, ассоциаций и метафор для расширения ключевых слов; генераторы названий брендов. Также существуют возможности выбрать пустые название, выбранное лишь из принципа не нарушения авторских прав или торговых марок других компаний. Такие примеры рассматриваются отдельно на основе результатов исследования.

Для проведения данного исследования был собран эмпирический материал — названия брендов в количестве 910 с разделением их на сферы бизнеса — 70 брендов в каждой сфере. Разделение на сферы бизнеса сделано с выделением смежных сфер: кафе и рестораны, санатории, базы отдыха, отели и так далее (табл. 1). В ходе исследование сравниваются оценки в смежных сферах.

В рамках исследования наиболее подходящим методом тестирования была выбрана оценочная шкала, поскольку она предоставляла респондентам возможность выставить оценку по различным критериям, что в свою очередь помогало полно и всесторонне оценить названия брендов. Оценка производилась по конкретно заданным показателям, покрывающим основные характеристики, требующие анализа. Были выделены основные критерии, которые выбирались респондентами без зависимости между значениями: соответствие, запоминаемость, благозвучие, понятность, оригинальность, привлекательность, выразительность. Каждый из критериев отвечал за узнаваемость названия бренда. Так, например, соответствие позволяет человеку понять, к какой сфере бизнеса относится компания без дополнительной информации (логотип, слоган и другие).

Таккже в ходе опросов были собраны информация о респондентах: возраст, пол, сфера деятельности.

Таблица 1. Соответствие каждой сфере смежных к ней сфер с учётом синонимичности и вида деятельности

№	Сфера бизнеса	Наличие смежной сферы
1	Косметология	Фармацевтика, Салоны красоты
2	Фармацевтика	Косметология, Взрослые клиники
3	Рекламные агентства	Туристические агентства
4	Транспортные компании	-
5	Салоны красоты (парикмахерские)	Косметология
6	Санатории и базы отдыха	Отели
7	Рестораны	Кафе
8	Отели	Санатории и базы отдыха
9	Взрослые клиники	Фармацевтика
10	Магазины одежды	-
11	Продуктовые магазины	Кафе, Рестораны
12	Туристические агентства	Рекламные агентства
13	Кафе	Рестораны

В процессе исследования использовалась оценочная шкала от «0» до «5», где «0» — несоответствие критерию, а «5» — полное соответствие критерию. В ходе исследования оценки объединялись в три группы специально для того, чтобы учесть погрешность. Оценки «0» и «1» относятся к первой группе, «2» и «3» ко второй и «4» и «5» к третьей. Важно отметить, что высокое соответствие оценок позволяет нейминг выполнять свою основную задачу отражения уникальных отличий, которые попадут в сознание потребителей и создадут нужные ассоциации в необходимые моменты. Поиск таких зависимостей позволит оценивать нейминги на соответствие задачам, которые перед маркетологами ставят компании.

Для проверки принадлежности нейминга или его части к синонимам, ассоциациям и метафорам, связанными со сферой бизнеса, используются статические предсказывающие модели Word2Vec [3; 4; 5]. Далее сравниваются оценки неймингов и пути их создания в смежных сферах. В процессе работы дополнительно использовалась библиотека Gensim. Были разделены все части речи, удалены знаки препинания и стоп-слова. В таблице 2 представлено описание корпусов, на основе которых используются статические предсказывающие модели Word2Vec. В данном случае, под стоп-словами подразумеваются слова, которые использовались для стилизации названия или широкого обозначения местности. Под это обозначение в предложенном контексте попали части наименований, дублирующие названия сайтов компаний: www, ru, ру и так далее. Названия разбивались на части вручную, также сокращённые названия дополнялись.

Таблица 2. Описание лингвистических корпусов

№	Постоянный идентификатор	Корпус	Размер корпуса (слов)	Размерность вектора
1	ruwikiruscorpora_upos_cbow_300_10_2021	НКРЯ и Википедия за ноябрь 2021	1.2 млрд	300
2	ruscorpora_upos_cbow_300_20_2019	НКРЯ	270 млн	300

Синонимы с помощью модели Word2Vec находились и внутри названий самих сфер. В таблице 1 представлены смежные сферы и синонимы. Стоит отметить, что некоторые сферы в дальнейшем можно разделить и изучить их оценки отдельно. Например, базы отдыха и санатории - учреждения этих сфер различаются в своей специфике, что влияет на целевую аудиторию. Санатории подразумевают длительный отдых под присмотром

медицинского персонала, а базы отдыха могут использовать в более короткий срок и с иными целями. Также важно отметить, что смежные сферы у базы отдыха и санаториев различны.

4. Результаты исследования

Рассмотрим пример из раздела аптек: Башфармация заменяется в Башкортостанская фармация. В названиях, состоящих из нескольких слов, производился поиск ближайшего синонима сферы. В результате слова с наличием близкого синонима всегда имели оценки по соответствию из 1 группы. Далее к группе синонимов требуется добавить ассоциации и метафоры, поскольку в нейминга часто используется не синонимичная близость, а культурные ассоциации или транслитерация синонимичных слов.

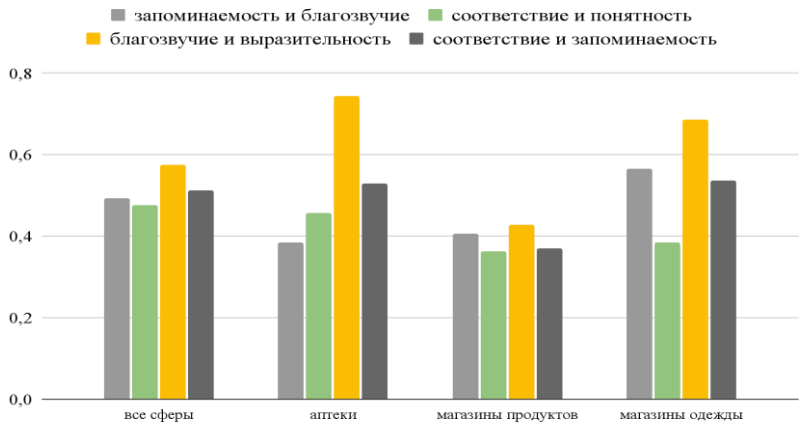

Рис. 1. Сравнение средних значений соответствия оценок запоминаемость и благозвучие, соответствие и понятность, благозвучие и выразительность, соответствие и запоминаемость среди всех сфер, аптек, магазинов продуктов и магазинов одежды

При дальнейшем рассмотрении соответствия значений оценок было выявлено, что для каждой сферы среднее значение двух параметров уникально. Эту тенденцию можно наблюдать на рисунке 1, представляющей соответствия различных оценок как в трех несвязанных сферах, так и среди всех представленных сфер. Рассмотрим среднее значение соответствия оценок запоминаемости и благозвучия после группировки в сферах «Аптеки», «Санатории» (в которые входят санатории и базы отдыха), и «Взрослые поликлиники». Ни одна из них не имеет значение выше среднего, а сфера аптек имеет наименьшее соответствие оценок запоминаемости и благозвучия.

При изучении причин полученного факта было замечено, что сфера аптек имела наибольшее количество стоп-слов. Множество неймингов были составлены из сокращений, которые влияют на благозвучие названий. Среди трех сфер, представленных к сравнению, именно аптеки по среднему показателю имеют большую длину — 14,8 символов, так как некоторые состоят из соединения сокращений, обозначающих регион, в котором работает компания, и ассоциативных слов, связанных со сферой здравоохранения. Самые короткие названия из трех сфер у взрослых поликлиник — средняя длина 10,1 символ.

При детальном рассмотрении на рисунке 2 соответствия высоких оценок запоминаемости и благозвучия, было замечено, что в сфере аптек несмотря на наименьший показатель соответствия двух оценок, большинство из них являются наивысшими, из чего следует зависимость этих показателей в данной сфере. Было также отмечено, что нейминги с соответствующими высокими оценками имеют всего два названия со стоп-словами, среднее значение длины меньше на 1,5 символа — 13,3 символов против 14,8 символов общей средней длины среди неймингов аптек. Среднее значение длины неймингов с

высокими оценками в сферах взрослых поликлиник и санаториев ниже — 11,0 и 9,4 — соответственно.

В сферах, связанных с питанием, — продуктовые магазины, рестораны и кафе, видна обратная зависимость — короткие нейминги чаще имеют высокие оценки запоминаемости и благозвучия. В сфере аптек замечено повышенное использование заимствованных слов из латыни — каждое 5 название содержит транслитерацию, а в сфере санаториев и баз отдыха ниже среднего наличие ближайшего синонима — 3 нейминга из 70. В сфере санаториев также используются синонимы к слову природа — 52 нейминга из 70. На примерах видно, что нейминги строятся не только на основе синонимов — используются разные алгоритмы создания названия бренда.

Также нейминги с высокими оценками запоминаемости и благозвучия были рассмотрены на наличие синонимов. В сферах взрослых поликлиник, аптек и санаториев 72,2% рассматриваемых неймингов имеют близкие синонимы в названиях. В сфере питания всего 16,8% рассматриваемых неймингов имеют близкие синонимы в названиях.

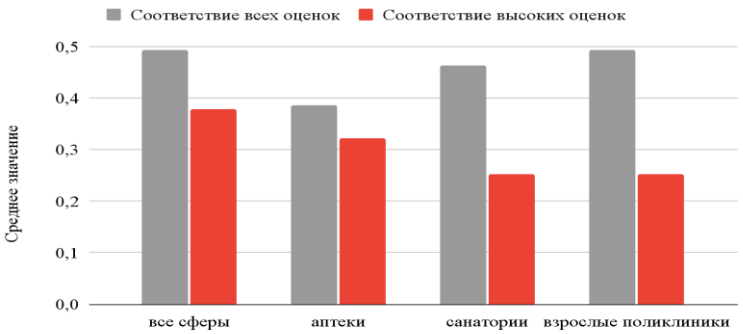

Рис. 2. Сравнение средних значений соответствий оценок запоминаемость и благозвучие во всех сферах, в сфере аптеки, санатории и взрослые поликлиники

Сфера продуктовых магазинов отличается от других сфер усреднёнными низкими показателями, которые демонстрируются на рисунке 2. Средняя длина нейминга магазинов продуктов наименьшая среди всех показателей — 7,8 символов. Среди неймингов с большинством оценок из 3 группы присутствуют лишь названия, состоящие из одного слова. По результатам опроса в сфере продуктовых магазинов большинство неймингов чаще имеют высокую оценку понятности и реже остальных имеют высокую оценку оригинальности.

5. Заключение

В результате эксперимента произведено сравнение значений оценки соответствия нейминга бренда и наличия синонимов в названии, также было проведено сравнение неймингов смежных сфер бизнеса. Рассмотрены основные алгоритмы составления неймингов, выявлены алгоритмы с низкой эффективностью по оценкам запоминаемости и соответствия. Замечено понижение оценок при удлинении названия бренда или внесении нескольких сокращений в сферах продажи продуктов, и повышение оценок при использовании ближайших синонимов в сферах здравоохранения и отдыха.

В дальнейших планах по развитию проекта — создание модели, проверяющей привлекательность и запоминаемость неймингов. Планируется увеличение размера корпуса с 910 слов минимум до 20 тысяч названий с выбранными оценками. Для создания модели также необходимо будет увеличить количество исследуемых сфер, расширить опросы, а также привлечь большее количество респондентов.

Литература

[1] Александрова И. Ю. Проблема нарушения системной реализации психосемантических и психолингвистических требований к названию торговой марки // Вестник университета (Государственный университет управления). 2019. № 2. С. 150–156. DOI: 10.26425/1816-4277-2019-2-150-156.
[2] Елистратов В. С., Пименов П. А. Нейминг: искусство называть. М.: Издательство «Омега-Л», 2014. С. 8–25.
[3] RusVectores [сайт]. URL: https://rusvectores.org/ru/ (дата обращения 11.06.2023).
[4] Тимофеева М. К. Типология семантических отношений, выявляемых посредством инструмента RusVectores // Научный диалог. 2018. № 8. С. 74–87.
[5] Radim Řehurek, models.word2vec — Word2vec embeddings // Gensim, 2019. — URL: https://sysblok.ru/knowhow/word2vec-pokazhi-mne-svoj-kontekst-i-ja-skazhu-kto-ty/ (дата обращения 10.06.2023).
[6] Radim Řehurek, Word2vec Tutorial, 2014 // Rare technologies — URL: https://rare-technologies.com/word2vec-tutorial/ (дата обращения 10.06.2023).

Identification of Hidden Patterns in the Public Reaction to a Brand: Analysis of the Attractiveness of Names Using Machine Learning Methods

Ekaterina M. Tatur, Ekaterina V. Klimenko

Saint-Petersburg State University

The article deals with the problem of identifying hidden patterns in the reaction of society to brands by assessing the attractiveness of the name using machine learning methods. The authors analyze a large amount of data, including the names of various brands and information about their popularity among consumers, using mathematical methods and machine learning algorithms. In the course of the study, experiments were carried out with static predictive Word2Vec models. The results of the research prove that the approach developed by the authors allows finding names that meet the requirements of the marketing sphere.

Keywords: classification of texts, branding, naming

Reference for citation: Tatur E. M., Klimenko E. V. Identification of Hidden Patterns in the Public Reaction to a Brand: Analysis of the Attractiveness of Names Using Machine Learning Methods // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 60–66. DOI: 10.17586/2541-9781-2024-7-60–66

Reference

[1] Aleksandrova I. Yu. Problema narusheniya sistemnoj realizacii psihosemanticheskih i psiholingvisticheskih trebovanij k nazvaniyu torgovoj marki // Vestnik universiteta (Gosudarstvennyj universitet upravleniya). 2019. № 2. Pp. 150–156. DOI:10.26425/1816-4277-2019-2-150-156. (in Russian)
[2] Elistratov V. S., Pimenov P. A. Nejming: iskusstvo nazyvat'. M.: Izdatel'stvo «Omega-L», 2014. Pp. 8–25. (in Russian)
[3] RusVectores. URL: https://rusvectores.org/ru/ (access date: 10.06.2023).

[4] Timofeeva M. K. Tipologiya semanticheskih otnoshenij, vyyavlyaemyh posredstvom instrumenta RusVectores // Nauchnyj dialog. 2018. № 8. Pp. 74–87. (in Russian)
[5] Radim Řehurek, models.word2vec — Word2vec embeddings // Gensim— 2019. — URL: https://sysblok.ru/knowhow/word2vec-pokazhi-mne-svoj-kontekst-i-ja-skazhu-kto-ty/ (access date: 10.06.2023).
[6] Radim Řehurek, Word2vec Tutorial, 2014 // Rare technologies — URL: https://rare-technologies.com/word2vec-tutorial/ (access date: 10.06.2023).

Сведения, информация и информационная коммуникация

Л. А. Ходоровский

lahod@mail.ru

Аннотация

Рассматривается соотношение понятий «сведения» и «информация», являющихся элементами процесса «информационная коммуникация». Дается определение понятия «сведения о свойствах сущности» как обозначение неоднородности реального или вымышленного мира и неравномерности протекания процессов в этом мире, характеризующих эту сущность. Информация о сущности определяется как сведения о ней, влияющие на то, каким образом должна выполняться в системе та деятельность, в которой участвует эта сущность. Информационная коммуникация рассматривается как многошаговый процесс, обеспечивающий преобразование сведений о сущности и передачу их системе, использующей эти сведения. Сведения рассматриваются не как простое отражение результатов преобразования на очередном шаге информационной коммуникации, но как структурированное сообщение о свойствах сущности.

Ключевые слова: сведения, информация, информационная коммуникация, данные, сигнал, преобразование информации, потенциальная информация

Библиографическая ссылка: Ходоровский Л. А., Сведения, информация и информационная коммуникация // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 67–80. DOI: 10.17586/2541-9781-2024-7-67–80

1. Введение

В данной статье рассматривается соотношение между понятиями «сведения» и «информация», а также связывающим эти две сущности процессом «информационная коммуникация».

По определению Н. Винера «Информация — это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему и приспосабливания к нему наших чувств» [1]. А «сведения» по дефиниции «Большого современного толкового словаря русского языка» — это: «1) Известия, сообщения о чем-либо. 2) Факты, данные, характеризующие кого-либо, что-либо. 3) Познания в какой-либо области, осведомленность в чем-либо. 4) Отчет с цифровыми данными». В смысле первых двух толкований понятие «сведения» очевидно подходит на роль «обозначения содержания» чего-либо, что и отражено в ряде дефиниций понятия «информация», например:

- информация первоначально — сведения, передаваемые одними людьми другим людям устным, письменным или каким-либо другим способом (например, с помощью условных сигналов, с использованием технических средств и т. д.) (БСЭ);
- информация — любые сведения, данные, сообщения, передаваемые посредством сигналов («Энциклопедия культурологии»);
- информация — сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации (ГОСТ 7.0-99 «Информационно-библиотечная деятельность. Термины и определения» [2]).

Рассмотрение этих и подобных дефиниций приводит к необходимости ответить на вопросы: что же такое сведения? что такое информация? какие сведения являются информацией? Ряд аспектов, связанных с ответами на эти вопросы, рассматривались нами в одной из работ [3].

Начнем с определения понятия «информация», опираясь на общеупотребительное представление о «сведениях», а потом уточним и представление об этом понятии.

2. Информация

Понятие информации неразрывно связано с понятием «система». Это понятие возникает в связи с обсуждением функционирования высокоорганизованных систем, уровень сложности которых таков, что они способны к целенаправленным действиям, т. е. в стремлении к достижению некоторой цели способны выбирать тот или иной способ действий. К таким системам относятся как системы, участвующие в процессах жизнедеятельности организмов, в психической деятельности человека, так и системы, искусственно сконструированные человеком, а также системы духовной и социальной деятельности человека.

Деятельность систем *неживой* природы может рассматриваться без привлечения понятия информации. А. В. Соколов пишет, что на страницах научно-мировоззренческих книг Стивена Хокинга он не обнаружил понятия информации. «Оказалось, что о Большом взрыве, расширяющейся вселенной, черных дырах, истории времени и других фундаментальных основах Макрокосма можно компетентно рассуждать, не вспоминая об информации как необходимой составляющей физической реальности» [4, с. 169]. Но когда речь идет об объяснении жизнедеятельности *живых* организмов, то, как отмечает Н. Н. Моисеев, «это невозможно без введения в язык термина «информация» … Только законов физики и химии для этого оказывается недостаточно… Информация нужна субъекту (организму) для возможности выбора способа действий при стремлении к достижению некоторой цели» [5, с. 47].

Под системой понимается совокупность элементов некой среды, взаимодействующих между собой, а также с элементами, не входящими в состав системы (внешней средой). Эти взаимодействия будем называть деятельностью системы, а элементы системы, участвующие в этих взаимодействиях, *ресурсами*. Деятельность любой системы подчиняется определенной цели. Речь об информации возникает тогда, когда мы говорим о достаточно сложных системах, называемых системами с целенаправленной деятельностью (СЦД).

В системах с целенаправленной деятельностью для разных значений свойств ресурсов могут быть предусмотрены разные варианты реализации деятельности, ведущие к достижению цели. Та часть системы, которая выполняет те или иные варианты деятельности, называется ее исполнительной подсистемой. Выбором необходимого варианта деятельности исполнительной подсистемы заведует другая подсистема — управляющая. Но для того, чтобы осуществить выбор, управляющая подсистема должна заранее знать, каковы свойства тех или иных ресурсов, участвующих в деятельности системы. Эти-то сведения о ресурсах и являются информацией.

Из вышесказанного вытекает, что между системой и сущностью, участвующей в деятельности системы, существуют два вида связи: непосредственная (ресурсная) и информационная — опосредованные сведения об этой сущности (см. рис.).

Понятие «информация» нынче используется в самых разных ситуациях, однако, нельзя говорить об «информации вообще», всегда говорится про информацию о чем-то. (Д. И. Дубровский: «Любой акт сознания интенционален, это всегда *информация о чем-то*» [6, с. 131].) Поэтому мы далее говорим о дефиниции понятия не «информация вообще», а «информация о сущности». Под сущностью понимается любое «нечто» в материальной или

воображаемой реальности: объекты (вещи, мысленные образы вещей), процессы, явления, отношения.

Рисунок. Схема деятельности системы *S* по манипулированию одним ресурсом

Определение 1. **Информация о некоторой сущности — это сведения об этой сущности, привлекаемые некой системой для влияния на ход той ее деятельности, в которой участвует эта сущность.**

Существенная сторона данной дефиниции: в ней указывается, какие именно сведения являются информаций — это сведения о свойствах ресурса, участвующего в деятельности системы, причем такие, которые влияют на выбор варианта деятельности.

Н. Н. Моисеев отмечает: «Субъект обладает системой рецепторов, способных принять сигнал. Вот их действия и могут быть описаны на языке физики и химии. Но дальше начинается уже малопонятное: организм способен к селекции сигналов, выделению того, что ему «интересно», и игнорированию сигналов, несущих сведения, не влияющие на характер его жизнедеятельности» [5, с. 52]. Видимо, эта селекция сигналов и определяется тем, скоррелирован ли сигнал с каким-либо из ресурсов, которыми манипулирует субъект.

Такое определение информации соответствует представлению об информации как о чем-то двойственном, соотнесенном с двумя разными явлениями (процессами): с одной стороны — это сведения о чем-то, результат отражения (образ) этого «чего-то», а с другой стороны — это нечто, используемое некой системой для влияния на ход ее деятельности, значимое для нее в смысле некоторого критерия.

В ряде дефиниций информации отмечается только отражательный аспект, в других отмечается влияние на систему (объект), например: «Информацией для объекта является сообщение (или множество сообщений), изменяющее его состояние [7]», но не уточняется, какова суть «изменений состояния». В нашем же определении указывается: сведения, непосредственно содержащиеся во входном сигнале, должны отражать характеристики такой сущности, которой манипулирует система, причем такие, которые влияют на выбор варианта деятельности системы.

3. Сведения

Рассмотрение понятия «сведения» опирается на два важнейших положения о свойствах природы.

Первое положение. Объективным свойством реальности является факт неоднородности распределения материи и энергии в пространстве и неравномерности протекания во времени всех процессов, происходящих в мире живой и неживой природы, а также в человеческом обществе и сознании.

Указанные неоднородности и неравномерности определяют структуру реального мира, позволяют выделять в этом мире различные структурные образования: стабильные объекты,

динамические процессы, отношения между ними. Будем называть эти образования сущностями.

Исходя из вышеназванного представления о свойствах реальности, В. М. Глушков в 1963 году дал следующее определение информации [8, с. 36]: «Информация в самом общем её понимании представляет собой *меру* неоднородности распределения материи и энергии в пространстве и во времени, *меру* изменений, которыми сопровождаются все протекающие в мире процессы».

Естественно понимать, что неоднородности и неравномерности могут относиться к миру как материальной, так и воображаемой, субъективной реальности. И поэтому определение может быть распространено на все процессы, протекающие в реальном или воображаемом мире.

Итак, информация определяется как мера чего-то. Естественно, под таким «что-то» понимать какое-либо структурное образование, являющееся сущностью, и свойства этой сущности, т. е. присущие ей особенности неоднородности распределения материи и энергии и неравномерности протекания неких процессов (реальные или воображаемые).

В качестве меры неоднородности/неравномерности может выступать какое-то обозначение, называние, описание этой самой неоднородности/неравномерности. Мера может выражаться количественно (численно), словесно (описательно), вероятностно, изобразительно и др.

Приведенные нами определения и рассуждения соответствуют так называемому функциональному подходу к определению и пониманию понятия «информация», в котором считается, что сведения могут описывать любые понятия живой, неживой, вымышленной реальности, но использоваться эти сведения в качестве информации могут только в системах высокоорганизованной материи, связанных с живой природой.

Существует и другой подход — атрибутивный, который исходит из предположения, что информация присуща любой реальности. Сторонники этого подхода определяют информацию как всеобщее *свойство*: «Информация, в широком понимании этого термина, представляет собой объективное свойство реальности, которое проявляется в неоднородности (асимметрии) распределения материи и энергии в пространстве и в неравномерности протекания во времени всех процессов, происходящих в мире живой и неживой природы, а также в человеческом обществе и сознании» [9]. А также: «Информация — это проявление всеобщего свойства материального мира быть определенным (существовать и изменяться в соответствии с законами природы), быть определяемым (воспринимаемым и идентифицируемым) и быть определяющим (способным изменять состояние некоторого целевого объекта) [10]. Применение такого подхода к отдельным сущностям приводит к тому, что «информацией» называется сама совокупность неоднородностей и неравномерностей, *присущих* этой сущности, образующих «структуру» этой сущности. Эту «информацию» (т. е. собственно «структуру») называют «первичной, или связанной информацией».

Отметим, что первая из процитированных выше дефиниций содержит практически те же слова, что и дефиниция В. М. Глушкова. Однако, в ней отсутствует слово «мера», т. е. *описание* этой самой «связанной информации» (которое и обуславливает значимость информации для ее содержательного использования). И поэтому сторонникам атрибутивного подхода приходится вводить понятие «вторичной, или свободной» информации, которая «представляет собой некоторое «отражение» первичной информации и может быть отчуждена от своего первоисточника».

Понятие вот этой «вторичной информации» соответствует тому, что считается «информацией» сторонниками функционального подхода, которого придерживается и автор данной статьи[1].

Отметим, что в определении В. М. Глушкова рассматривается только аспект обозначения содержания, аспект использования не отражен вовсе. Поэтому эту дефиницию правильнее рассматривать не как определение понятия «информация», а как определение понятия «сведения». Отсюда следует следующее определение.

Определение 2**. Сведения о свойствах реальной или воображаемой сущности — это мера, описание, обозначение присущих данной сущности особенностей неоднородности реального или воображаемого распределения материи и энергии в пространстве и во времени и/или неравномерности протекания процессов в реальном или воображаемом мире.**

Итак, в соответствии с определением 2 сведения о какой-либо сущности — это описание её свойств. Но, в соответствии с определением 1, это описание может выступать в роли информации только в том случае, когда оно используется какой-нибудь системой для влияния на ход ее деятельности по манипулированию этой сущностью.

Отсюда: сведения есть описание, отражение свойств сущности, «объективное» с точки зрения отражающей системы. А информация — понятие субъективное, определяемое «потребностями» принимающей сведения системы.

4. Свойства сущностей

Свойство сущности — это присущее сущности некое качество, некий аспект неоднородности материи и энергии или неравномерности протекания процессов, взаимосвязей (отношений) с другими сущностями. Свойству соответствует множество возможных *вариантов его проявления*. Вариант проявления свойства — это вариант реализации соответствующего качества, степень выраженности некой характеристики неоднородности/неравномерности. Конкретной сущности присущ тот или иной вариант проявления свойства, называемый «значением свойства».

Как проявляются эти значения? Ответ на этот вопрос опирается на второе из вышеупомянутых свойств природы. А именно: «Между объектами реального мира постоянно осуществляются различные взаимодействия; только во взаимодействии с другими объектами могут проявиться и быть познанными свойства объекта» [11].

Таким образом, мера степени проявления свойств сущности определяется в результате (по результату) ее взаимодействия с другими сущностями. В ходе взаимодействия сущностей A и B в зависимости от значений свойств сущности A происходит изменение некоторых характеристик свойств сущности B. Изменения характеристик свойств сущности B отражают состояние сущности A. Как отмечается в работе [12], *отражение* «выражается в том, что из всего содержания взаимодействия выделяется лишь то, что в одной системе появляется в результате воздействия другой системы и соответствует (тождественно, изо- или гомоморфно) этой последней».

Пусть α — значение (вариант проявления) какого-то свойства a сущности A, и в результате его воздействия на сущность B изменилось значение какого-то свойства b сущности B. Вариант проявления этого свойства β — это *отражение* свойства сущности A в сущности B. Его можно рассматривать как обозначение проявления этого свойства, как знак, сопоставленный в свойствах сущности B соответствующему свойству сущности A, т. е. сам по себе *вариант проявления* β свойства b сущности B может пониматься как

[1] Нам представляется неудачным использование одного термина «информация» применительно к двум разным (хотя и связанным) понятиям. И для обозначения самих проявлений присущих реальным сущностям неоднородностей/неравномерностей следовало бы выбрать какой-нибудь другой термин.

обозначение варианта проявления α свойства *a* сущности *A*. (Например, глубина опускания чашки пружинных весов есть обозначение массы взвешиваемого предмета, несколько букв на бумаге — обозначение сказанного слова, совокупность пикселей на экране — обозначение формы предмета).

В свою очередь, в результате воздействия сущности *B* на третью сущность *D* варианту проявления β будет соответствовать некий вариант проявления δ, который может рассматриваться как обозначение для β и, опосредованно, для α. Заметим, что при воздействии сущности *A* на другую сущность *E* то же самое свойство *a* сущности *A* с проявлением α может воздействовать на некое свойство *e* сущности *E*, и тогда проявление ε свойства *e* может рассматриваться как другое обозначение проявления α свойства *a* сущности *A*. Так, одна и та же температура тела обозначается по-разному на термометрах Цельсия и Фаренгейта. Но подчеркнем сразу: осознать, что проявление β некого свойства сущности *B* есть обозначение значения какого-то свойства сущности *A*, может только некая система с целенаправленной деятельностью, на которую воздействует сущность *B*.

Проявление β свойства сущности *B,* отражающей свойства сущности *A*, т. е. собственно результат отражения, называется *данными* о сущности *A*.

Как правило, в реальности происходят многократные последовательные взаимодействия между несколькими сущностями. Проявление некоего свойства очередной сущности есть обозначение значения какого-то свойства предыдущей сущности и, опосредованно, обозначение проявления какого-то свойства начальной (исходной) сущности в цепочке. Например, для пружинных весов обозначениями веса взвешиваемого предмета (отражениями) могут выступать и глубина опускания чашки весов, и угол отклонения стрелки весов, и положение конца стрелки на шкале, и число, написанное у деления шкалы.

Факт проявления воздействия одной сущности на другую в виде данных присущ любым воздействиям как в мире косной материи, так и в мире живой материи и в ментальных процессах. Этот факт рассматривается сторонниками атрибутивного подхода к определению понятия информации как одно из оснований считать информацию присущей не только живой, но и косной материи.

Но собственно результат отражения (т. е. данные, непосредственно обозначающие значение свойства) не может полностью выполнять роль меры. Например, обозначение «37°» требует уточнений: во-первых, это температура или угол отклонения стрелки? во-вторых, какого пациента или какого прибора?

Таким образом, сведения как мера, характеризующая некую неоднородность/неравномерность, должны указывать, и о каком объекте неоднородности идет речь, и о каких его качествах. Т. е. сведения — это уже некоторое *структурное образование*, включающее не только результат отражения значения какого-то свойства, но и еще какие-то обозначения объекта, к которому относится это свойство. В качестве такого образования может выступать, например, триада обозначений (*обозначение сущности, обозначение ее свойства, обозначение варианта проявления этого свойства*) как характеристика этой самой неоднородности/неравномерности. Обозначение сущности и обозначение (название) свойства могут быть явными или неявными, следующими из контекста.

Подобную совокупность обозначений назовем *сведениями* о свойстве сущности, а всю совокупность данных, отражающих различные свойства одной сущности, назовем *сообщением*.

Сведения могут рассматриваться как законченное утверждение (предикативное отношение) некоего языка (или даже протоязыка). В. А. Курдюмов в [13], говоря о предикационной концепции языка, утверждает, что основой порождаемого текста

выступает первичная пара в сознании: «о чем? — что говорить?»[2], — которая оформляется в тексте в предикативное отношение (предикацию) между предицируемым (топик; то, о чём говорится) и предицирующим (то, что сообщается о топике) компонентом (комментарий). Далее: «Не будет ошибочным и простое описание: язык есть предикация. Мы полагаем, что в целом наиболее удобным и точным будет рабочее определение языка **как совокупности предикационных цепей**, поскольку все многомерное пространство как раз и покрывается этими цепями на разных уровнях».

Сведения (вышеупомянутая триада обозначений) представляются как предикация вида «У сущности *A* свойство *a* имеет значение α». Нам представляется, что повествовательные тексты могут сводиться к совокупности подобных утверждений.

Итак, для того, чтобы данные могли восприниматься как сведения о чем-то, они должны выражать не только обозначение варианта проявления свойства, но также и указания о том, проявлением какого свойства какой сущности является этот вариант. Такие указания могут отображаться как данными, входящими в состав рассматриваемого сообщения, так и данными из других сообщений. В полноценные сведения данные превращаются в процессе их осознания, т. е. восприятия системой с целенаправленной деятельностью, это она должна суметь «понять» какие свойства какой сущности представляются (отображаются) этими данными. Таким образом, отражение свойств одной сущности в свойствах другой происходит и в неживой природе, но осознать это отражение как информацию, могут только системы живого мира.

«Понимание» того, проявлением какого свойства являются воспринимаемые данные, в самых простых случаях определяется тем, что органы восприятия (рецепторы) реагируют только на конкретный вид раздражителя — электромагнитные колебания, давление, температуру, pH, pCO_2 и т.д. А понимание того, к какой сущности относятся эти данные, определяется неким «контекстом», их окружением: знанием структуры данных, или наличием «дополнительных» данных, о других свойствах, позволяющих определить «направление», откуда приходит сигнал (орган зрения уже определил объект, выступающий в роли источника шума, а затем орган слуха воспринимает исходящий от него звук). В развитых системах умение выявить структуру и суть сообщения заложено в модели самой воспринимающей системы.

Отметим, что речь идет не только о системах материального мира, манипулирующих материальными ресурсами в соответствии с законами физики и химии. Все изложенное в принципе можно отнести и к ментальным системам, манипулирующим сущностями субъективной реальности. В качестве ресурсов, которыми манипулируют такие системы, выступают ресурсы информационной природы (чувственные данные, представления, образы и другие содержания, наличествующие в уме), воплощенные в определенных мозговых нейродинамических системах, являющихся материальными носителями этой информации. Результаты преобразования таких ресурсов в материально представляемые тексты (устная и письменная речь, графика, скульптура и пр.) также являются ресурсами информационной природы.

Качественная специфика ресурсов информационной природы заключается в том, что при манипулировании ими существенны их информационные, а не вещественно-энергетические свойства; преобразуются же эти свойства в соответствии с законами логики, семиотики, лингвистики.

Когда речь идет о манипулировании ресурсами, имеющими информационную природу, разница между информацией, поступающей на ресурсный и на информационный входы системы, сохраняется. В этом случае будем называть (см. рис.): деятельность системы *S* «*преобразованием информации*»; обрабатываемую информационную сущность (ресурс *A*),

[2] Эта пара может выступать в роли источника информации информационной коммуникации (понятие, рассматриваемое нами далее в п. 4), обеспечивающей доставку текста принимающей системе.

поступающую на ресурсный вход системы *S*, — «*входной информацией*»; информацию, полученную в результате обработки (ресурс *C*), — «*выходной информацией*», а информацию об информационном ресурсе *A*, поступающую на информационный вход системы *S*, — «*привлекаемой информацией*».

Так, в различных автоматизированных информационных системах преобразуются ресурсы информационной природы, называемые данными, а в роли привлекаемой информации выступают сведения, называемые метаданными (или метаинформацией).

5. Информационная коммуникация

На рисунке указано, что информация об используемом системой *S* ресурсе *A* поступает к системе опосредованно, через отражение ее в свойствах промежуточных сущностей. При этом ресурс рассматривается как источник информации (ИИ) и выполняются, как минимум, два преобразования сведений об этом ресурсе: порождение (генерация) сообщения, отображенного в свойствах промежуточной сущности *D*, и передача сообщения системе *S*. В реальности маршрут передачи сведений содержит значительно больше преобразований: *A* воздействует на некую сущность B_1, B_1 на B_2 и т.д. Совокупность этих преобразований называется информационной коммуникацией (ИК). Схема ИК выглядит как $A \rightarrow B_1 \rightarrow B_2 \rightarrow \ldots \rightarrow B_n \rightarrow S$. Здесь *A* — источник информации, B_k (k=1,2…n) — это промежуточные сущности (носители информации), участвующие в процессе передачи информации, стрелка обозначает преобразование информации на одном шаге (воздействие одной сущности на другую). *S* обозначает целевую систему, получающую информацию; назовем ее приемником информации.

Результат преобразования информации на каждом шаге — это сообщение M_k, совокупность некоторых данных. Данные могут выражаться либо статическими, либо динамическими (переменными во времени) характеристиками носителя информации *Bk*. В первом случае говорим о *статических данных*, во втором — о *динамических данных*, которые принято называть *сигналом*. Сигналы обеспечивают передачу сведений в пространстве, а статические данные — во времени.

Сам факт воздействия сущности B_k на B_{k+1} есть факт выполнения деятельности некой системы, преобразующей сообщение M_k в сообщение M_{k+1}. Назовем эту систему преобразователем информации, или *посредником*, T_k. Тогда схему ИК можно представить в виде $A \rightarrow M_1 \rightarrow M_2 \rightarrow \ldots \rightarrow M_n \rightarrow S$, первая стрелка обозначает операцию генерации сведений об источнике информации *A*, последняя — операцию передачи сведений системе — потребителю информации, остальные обозначают деятельность систем-посредников T_k (k=1,2…n-1). Отметим, что преобразование сообщения M_k в сообщение M_{k+1} может носить характер как функциональный (по законам физики, химии, логики), так и конвенциальный (по законам семиотики).

Термин «информационная коммуникация» используется нами как обозначение процесса направленного переноса сведений от источника информации к системе-потребителю этой информации. Описание такой ИК может рассматриваться как линейная (ее еще называют «трансмиссионная») модель коммуникации, которая рассматривает процесс передачи информации от активного участника (отправителя) к пассивному (с точки зрения процесса общения) получателю сообщения (см., например, [14]). При рассмотрении же более сложных коммуникативных актов, включающих наличие обратной связи, диалог между участниками коммуникации, процессы информационной коммуникации встречаются несколько раз на разных шагах реализации таких актов.

6. Потенциальная информация

Конкретная ИК может реализовываться в режиме либо непосредственного, либо отложенного общения, когда результаты некоторых шагов фиксируются на каком-либо

материальном носителе и могут быть использованы для последующих преобразований через неопределенный промежуток времени (или вообще никогда). Назовем такие результаты «промежуточными сообщениями». Примеры таких сообщений: следы на снегу, книга, база данных. С точки зрения конкретной системы эти сообщения можно назвать «информация на полпути».

Часто промежуточные сообщения рассматриваются как тексты в некоторой предметной области, формируемые в соответствии с правилами этой предметной области, например, как данные, организованные в соответствии с требованиями некой реляционной СУБД. Такие сообщения могут длительно храниться в БД, и какая-нибудь ИК может включать несколько шагов преобразования сообщения в рамках той же предметной области средствами соответствующей СУБД.

Строго говоря, с точки зрения данных нами определений, результатами всех шагов ИК, кроме последнего, является порождение и преобразование *сведений* о сущности, а не *информации*. Но в соответствии со сложившейся традицией будем называть все эти результаты информацией, уточняя для порядка — *потенциальной информацией*.

7. Восприятие информации

Последний шаг информационной коммуникации — восприятие информации, т. е. воздействие некоего сигнала, несущего информацию об источнике информации *A*, на получателя и реакция получателя на этот сигнал. Эта реакция заключается в том, что получатель (целевая система) выбирает тот или иной вариант действий в зависимости от содержащихся в сигнале сведений об источнике информации.

Выбор варианта система принимает на основе анализа внешних управляющих воздействий (привлекаемой информации о ресурсах, подвергаемых манипуляциям системы) и сопоставления их с имеющейся у системы внутренней моделью деятельности.

В соответствии с этим привлекаемая информация может рассматриваться и как описание свойств обрабатываемой сущности, и как задание на выполнение деятельности системой, как некая инструкция или внешняя модель этой деятельности. «Сигнал есть модель и в смысле отображения события, его вызвавшего, и в смысле плана события, которое он вызовет. Сигнал органически воплощает в себе отображение и предуготовленное им действие» [15, с. 252]. В семиотической терминологии «модель» может пониматься как «знак».

Итак, привлекаемая информация есть сообщение, воздействующее на деятельность некой системы с целенаправленной деятельностью, а восприятие этой информации (построение варианта реализации этой деятельности) представляет собой субъективную интерпретацию сообщения, выявление его *смысла*.

В Кратком психологическом словаре [16], указаны два толкования термина «смысл»:1) суть, главное, основное содержание (иногда скрытое) в явлении, сообщении, или поведении; 2) личностная значимость тех или иных явлений, сообщений или действий, их отношение к интересам, потребностям и жизненному контексту в целом конкретного субъекта.

Оба эти толкования могут быть привлечены для прояснения сути двух основных аспектов деятельности управляющей подсистемы системы с целенаправленной деятельностью.

Во-первых, подсистема должна выявить содержание сообщения, поступающего в систему с сигналом, понять смысл того, что обозначает сообщение (первое толкование термина «смысл»). Необходимо «опознать» информацию, установить, что сведения, содержащиеся во входном сигнале, каким-то образом характеризуют свойства ресурсов, которыми манипулирует система. При этом сначала подсистема опознает в поступающих с сигналом динамических данных сообщение «Свойство b_n сигнала B_n имеет значение β_n», а затем последовательно выясняет, что это сообщение должно быть «опознано» как

сообщение о свойствах результатов преобразований на предыдущих шагах ИК, и так вплоть до опознания свойств источника информации (ресурса *A)*. Однако, зачастую требуется значительно меньшая «глубина опознания», так как значение на промежуточном шаге ИК оказывается вполне репрезентативным для оценки ситуации. Так, самец колюшки начинает брачный танец, реагируя на раздутое брюшко самки (и не пытается осознать, что такое брюшко есть признак созревания икры и т.д., и т.п.). А вот следователю порой необходимо докопаться до истинного, глубинного, а не демонстрируемого, источника информации.

Во-вторых, управляющая подсистема должна «осознать» содержание опознанного сообщения, т. е. проанализировать это содержание, воспринять его как некоторую инструкцию и сформировать конкретный вариант реализации деятельности системы. Для выполнения этих процессов системе может потребоваться привлечение и обработка дополнительной информации как из внутренней информационной базы системы, так и из внешних информационных баз. Именно влияние сообщения на процесс формирования варианта деятельности и есть *интерпретация* сообщения, проявление его значимости для системы, т. е. проявление смысла сообщения (во втором толковании)[3].

Таким образом, «семиотически», процесс восприятия информации целевой системой включает реализацию двух этапов: выявление денотата, соответствующего входному сигналу, и определение концепта этого сигнала, в роли которого выступает выбираемый системой вариант ее деятельности.

Мы рассматриваем информацию как сведения, влияющие на результат деятельности некой целевой системы. Вынуждены отметить, что многие авторы избегают акцентировать внимание на наличии и роли целевой системы. Поэтому в их дефинициях действуют другие «действующие лица». Например: «Информация — определенная некоторым контекстом совокупность данных» [10] («контекстом», а не «системой»!).

8. Шаги информационной коммуникации

Анализ процесса реализации информационной коммуникации показывает, что информация может участвовать в информационных действиях следующих типов: генерация информации; использование информации о некой сущности для влияния на ход деятельности системы, манипулирующей этой сущностью; преобразование представления информации о сущностях (формальное или содержательное преобразование ресурсов информационной природы).

Шагами преобразования представления информации являются все шаги этапа отражения, кроме первого (генерации), и все шаги этапа восприятия, кроме последнего (собственно восприятия).

Важными характеристиками шагов ИК являются преднамеренность, целенаправленность.

---

[3] Нами представлена принципиальная схема процесса восприятия смысла привлекаемой информации. В реальности процессы восприятия информации целевой системой оказываются весьма сложными, но соответствуют предлагаемой нами схеме. Так, А. У. Хараш, говоря о восприятии текстов, пишет: «А. А. Леонтьев предлагает различать два значения слова «смысл». Одно из них, которое он называет «нетерминологическим», традиционно связывается с операцией «укрупнения значений», проводимой при восприятии текста, как такового, и ведущей к извлечению из него сложных семантических конфигураций. Извлечение из текста его собственного, «текстуального» смысла А. А. Леонтьев очень точно характеризует как «значенческое понимание». Другое значение слова «смысл», квалифицируемое А. А. Леонтьевым как собственно «терминологическое», предполагает усмотрение в сообщении некоего содержания, которого нет в самом тексте. Смысл в этом случае извлекается не из текста, а из предметного мира коммуникатора, из сферы действительных мотивов его деятельности, его целостного (а не только «коммуникативного») бытия… Реципиент обнаруживает, какими содержаниями целостной (внекоммуникативной) деятельности коммуникатора продиктован сам факт предъявления данного текста, данного «текстуального смысла», и тем самым так или иначе превращает эти содержания в содержание своей деятельности, познавательной или практической» [17].

Процессы преобразований, происходящие в неживой природе (а зачастую и в живой), являются непреднамеренными и нецеленаправленными. Например, поваленные деревья вдоль трассы пролета Тунгусского метеорита; или образование годичных колец в результате прироста тканей дерева.

Однако, в результаты преобразований могут вноситься искажения за счет воздействия различных случайных факторов (случайный шум). Поэтому, например, естествоиспытатель, который хочет понять значимость какого-нибудь сигнала и его влияние на деятельность системы, должен определить, какие характеристики какого ресурса этот сигнал отражает и насколько велико возможное искажение этих характеристик в процессе реализации шагов соответствующей ИК.

Примером непреднамеренных целенаправленных преобразований могут служить следы петляющего зайца. Следы он оставляет непреднамеренно, а последовательность следов — целенаправленно запутывающая.

Для человеческой деятельности наиболее характерны преобразования информации преднамеренные и целенаправленные. Например, в медиасреде таковы практически все шаги в составе медиакоммуникации. При этом целенаправленность очередного преобразования заключается не только в том, что преобразуемые сведения рассчитаны на восприятие определенным потребителем, но и в том, что в преобразованном сообщении могут появиться искажения не только за счет случайного шума, но и в результате привнесения в сообщение добавлений, изменений лицом, выполняющим деятельность по преобразованию сообщения.

Некоторые операции преобразования информации могут представляться как самостоятельная деятельность по содержательному преобразованию информационных ресурсов (например, перепись населения, деятельность информационных фондов по накоплению, аккумуляции, анализу информации и пр.). Результаты преобразования далее могут использоваться в других преобразованиях этих и других ресурсов и т.д. Для людей, выполняющих такую деятельность, она подчас представляется самодостаточной. Однако, по сути, вся эта деятельность носит вспомогательный, подготовительный характер, она предназначена для обеспечения основной деятельности — удовлетворения информационных потребностей какой-либо системы (систем).

9. Выводы

И. В. Мелик-Гайказян утверждает, что информация есть не атрибут, объект или отношение, а «многостадийный необратимый во времени процесс» [18, с. 98]. По нашим же представлениям информация есть понятие, безусловно, связанное с процессом, однако, информация — это не сам процесс, а средство, обеспечивающее как связь между стадиями многостадийного необратимого во времени процесса, так и эффективность деятельности систем, реализующих выполнение этих стадий.

В качестве такого многостадийного процесса выступает информационная коммуникация, включающая многошаговый процесс доставки информации от сущности-источника информации к целевой системе, манипулирующей этой сущностью, и сам процесс интерпретации информации этой системой.

На разных шагах этого процесса информация выступает в разных ипостасях. На первом шаге происходит порождение (генерация) сведений о свойствах некоего ресурса целевой системы, выступающего в роли источника информации. На следующих шагах эти сведения подвергаются различным преобразованиям, т. е. сами эти сведения выступают как перерабатываемый ресурс информационной природы. На последнем шаге сведения используются целевой системой как информация, т. е. как сведения, влияющие на выбор варианта деятельности по манипулированию этим ресурсом.

В соответствии с этим информация всегда выступает как двойственный объект, свойства которого определяются связанностью как с точкой отправления, так и с точкой прибытия. Охарактеризуем эти двойственности.

А) Соотнесенность с двумя **разными процессами**: с одной стороны, это сведения, отражающие свойства некой сущности, с другой — сведения, предопределяющие выполнение определенных действий системой, принимающей информацию.

Б) **Семиотическая двойственность** сигнала: с одной стороны проявления свойств сигнала — это обозначение свойств сущности — источника информации, с другой — это обозначение варианта деятельности системы.

В) Два разных способа **«существования» информации**: с одной стороны это привлекаемая информация, сообщение, используемое управляющей подсистемой для влияния на ход деятельности системы, с другой стороны — потенциальная информация, хранимый результат реализации какого-либо шага информационной коммуникации, который, может быть, будет обработан и использован какой-то системой как привлекаемая информация.

Г) Два разных вида **информационной деятельности**: использование информации в процессе её восприятия системой, т. е. интерпретация смысла полученного сообщения, и использование информации при реализации шагов информационной коммуникации, т. е. формальное и содержательное преобразование сведений, их представления.

Д) Различие **содержательного смысла** разных этапов реализации информационной коммуникации: вначале смысл операций преобразования заключается в видоизменении описания свойств источника информации, в приведении его к форме, удобной для хранения, затем — в конструировании сигнала, содержание которого может выступать в роли инструкции, влияющей на ход деятельности целевой системы.

В заключение — об одном важнейшем свойстве информации (хотя, может быть, его следует приписать не информации, а системам, способным использовать информацию). Это свойство — обеспечение замещения физической причинности на информационную [18, с. 86].

Тем самым информация предоставляет системе возможность не «проверять на деле», каков результат (возможного) физического взаимодействия конкретных ресурсов, а проводить «проверку» на основании «сообщений» о свойствах этих ресурсов. И на основании результатов этой проверки выбирать соответствующий вариант деятельности. В действительности замещаться могут не только физические взаимодействия материальных ресурсов, но и взаимодействия ресурсов информационной природы.

Такое свойство замещения обеспечивает значительную экономию вещественно-энергетических и временных затрат как в области материальной, так и информационной деятельности.

Литература

[1] Винер Н. Кибернетика и общество. — М: Изд-во «Иностранная литература», 1958.
[2] ГОСТ 7.0-99. Информационно-библиотечная деятельность. Термины и определения (Электронный ресурс). — 1999. — URL: https://ivo.garant.ru/#/basesearch/ГОСТ 7.0-99 (дата обращения 07.09.2021).
[3] Ходоровский Л. А. К определению понятия «информация» // Научно-техническая информация. Сер. 2. 2021. № 10. С. 1–17.
[4] Соколов А. В. Философия информации. — СПб: СПбГУКИ, 2010. — 368 с.
[5] Моисеев Н. Н. Универсум. Информация. Общество. — М: Устойчивый мир, 2001. — 200 с.
[6] Дубровский Д. И. Проблема идеального. Субъективная реальность. Второе доп. издание. — М: Канон+, 2002. — 368 с.
[7] Сапрыкин М. Ю., Сапрыкина Н. А. Анализ понятия «информация» с позиции объектно-ориентированного подхода // Науковедение. 2016. Т. 8, №2. URL:

http://naukovedenie.ru/PDF/36TVN216.pdf. DOI: 10.15862/36TVN216. (дата обращения: 07.09.2021)
[8] Глушков В. М. Мышление и кибернетика // Вопросы философии. 1963. №1. С. 36–48.
[9] Колин К. К., Урсул А. Д. Информация и культура. Введение в информационную культурологию. — М: Изд-во «Стратегические приоритеты», 2015. — 288 с.
[10] Максимов Н. В., Лебедев А. А. К конструктивному определению свойств информации // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединенной научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2023.
[11] Философский энциклопедический словарь / гл. редакция: Л. Ф. Ильичёв, П. Н. Федосеев, С. М. Ковалёв, В. Г. Панов. — М: Советская энциклопедия, 1983.
[12] Урсул А.Д. Природа информации. Философский очерк. — М.: Политиздат, 1968.
[13] Курдюмов В. А. Идея и форма. Основы предикационной концепции языка. М.: Воен. ун-т, 1999. — 194 с.
[14] Долженков В. Н. Характеристика трансмиссионной модели коммуникации // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2017. № 3(69): в 3-х ч. Ч. 2. С. 88–90. ISSN 1997-2911.
[15] Дубровский Д. И. Психические явления и мозг. — Москва: «Наука», 1971. — 386 с.
[16] Краткий психологический словарь / Сост. Л. А. Карпенко, А. В. Петровский, М. Г. Ярошевский. — Ростов-на-Дону: «ФЕНИКС», 1998.
[17] Хараш А. У. Смысловая структура публичного выступления (об объекте смыслового восприятия)//Вопросы психологии. 1978. № 4. С. 84–95.
[18] Информационный подход в междисциплинарной перспективе (материалы «круглого стола» // Вопросы философии. 2010. № 2. С. 84–112.

Intelligence, Information and Information Communication

Leonard A. Khodorovskii

The relationship between the concepts "intelligence" and "information", which are elements of the "information communication" process, is considered. The definition of the concept of "intelligence about the properties of an entity" is given as a designation of the heterogeneity of the real or fictional world and the unevenness of processes in this world that characterize this entity. Information about an entity is defined as intelligence about it that affects the way how the activity in which the entity participates should be performed in the system. Information communication is considered as a multi-step process that ensures the transformation of intelligence about the entity and its transfer to the system that uses this intelligence. Intelligence is considered not as a simple reflection of the results of the transformation at the next step of information communication, but as a structured message about the properties of the entity.

Keywords: information, information communication, intelligence, signal, information transformation, potential information

Reference for citation: Khodorovskii L. A. Intelligence, Information and Information Communication // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 67–80.DOI: 10.17586/2541-9781-2024-7-67–80

Reference

[1] Wiener N. Kibernetika i obshchestvo. — M.: Izd-vo «Inostrannaya literatura», 1958. (in Russian)

[2] GOST 7.0-99. Informatsionno-bibliotechnaya deyatel'nost'. Terminy i opredeleniya (Elektron. resurs). — 1999. — URL: https://ivo.garant.ru/#/basesearch/GOST 7.0–99 (access date: 07.09.2021). (in Russian)

[3] Khodorovskij L. A. K opredeleniyu ponyatiya «informaciya» // Nauchno-tekhnicheskaya informaciya. Ser. 2. 2021. № 1 Pp. 1–17. (in Russian)

[4] Sokolov A. V. Filosofiya informatsii. — SPb: SPbGUKI, 2010. — 368 p. (in Russian)

[5] Moiseyev N. N. Universum. Informatsiya. Obshchestvo. — M.: Ustoychivyy mir, 2001. — 200 s. (in.Russian)

[6] Dubrovskii D. I. Problema ideal'nogo. Sub"yektivnaya real'nost'. Vtoroye dop. izdanie. — M.: Kanon+, 2002. — 368 p. (in Russian)

[7] Saprykin M. Yu., Saprykina N. A. Analiz ponyatiya «informaciya» s pozicii ob"ektno-orientirovannogo podhoda // Naukovedenie. 2016. T. 8, №2. URL: http://naukovedenie.ru/PDF/36TVN216.pdf. DOI: 10.15862/36TVN216 (access date: 07.09.2021). (in Russian)

[8] Glushkov V. M. Myshlenie i kibernetika // Voprosy filosofii. 1963. №1. S. 36-48 (in Russian).

[9] Kolin K.K., Ursul A.D. Informatsiya i kul'tura. Vvedeniye v informatsionnuyu kul'turologiyu. — M: Izd-vo «Strategicheskiye prioritety», 2015. — 288 p. (in Russian)

[10] Maksimov N. V, A. A. Lebedev A. A. K konstruktivnomu opredeleniyu svoystv informatsii // Komp'yuternaya lingvistika i vychislitel'nyye ontologii. Vypusk 7 (Trudy XXVI Mezhdunarodnoy ob"yedinennoy nauchnoy konferentsii «Internet i sovremennoye obshchestvo», IMS-2023, Sankt-Peterburg, 26–28 iyunya 2023 g. Sbornik nauchnykh statey).— SPb.: Universitet ITMO, 2023. (in Russian)

[11] Filosofskii enciklopedicheskij slovar' / gl. red.: L.F. Il'ichyov, P. N. Fedoseev, S. M. Kovalyov, V. G. Panov. — M.: Sovetskaya enciklopediya, 1983. (in Russian)

[12] Ursul A. D. Priroda informacii. Filosofskij ocherk. — M.: Politizdat, 1968.

[13] Kurdyumov V. A. Ideya i forma. Osnovy predikatsionnoy kontseptsii yazyka. M.: Voyen. un-t, 1999. — 194 p. (in Russian)

[14] Dolzhenkov V. N. Kharakteristika transmissionnoy modeli kommunikatsii // Filologicheskiye nauki. Voprosy teorii i praktiki. Tambov: Gramota, 2017. № 3(69): v 3 -kh ch. Ch. 2. Pp. 88–90. (in Russian)

[15] Dubrovskii. Psihicheskie yavleniya i mozg. — M.: «Nauka», 1971. — 386 s. (in Russian)

[16] Kratkij psihologicheskij slovar' / Sost. L. A. Karpenko, A. V. Petrovskij, M. G. Yaroshevskij. — Rostov-na-Donu: FENIKS, 1998. (in Russian)

[17] Kharash A. U. Smyslovaya struktura publichnogo vystupleniya (ob ob"yekte smyslovogo vospriyatiya) // Voprosy psikhologii. 1978. № 4. Pp. 84-95. (in Russian)

[18] Informacionnyj podhod v mezhdisciplinarnoj perspektive (materialy «kruglogo stola») // Voprosy filosofii. 2010. № 2. Pp. 84–112. (in Russian)

Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей

А. В. Чижик

Университет ИТМО

chizhik@itmo.ru

Аннотация

Анализ тональности текстов является одной из актуальных задач, которая способна выявлять важные факторы, влияющие на вектор социального настроения общества. При использовании для решения этой задачи методов машинного обучения требуется преобразовать текст в его векторное представление. Существует ряд методов векторизации текстов. В данной статье сравниваются три актуальных на данный момент подхода к созданию векторного представления: учет веса слова в документе (TF-IDF), использование дистрибутивной семантики при создании векторов слов (Word2Vec) и векторизация целых предложений (Laser). Сравнивая эти три модели векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей, можно сказать, что каждая из них имеет свои преимущества и недостатки. В статье описан дизайн исследования, приведены метрики качества, описаны данные, на которых проводились опыты.

Ключевые слова: векторизация текстов, анализ тональности, социальные медиа

Библиографическая ссылка: Чижик В. А. Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 81–89. DOI: 10.17586/2541-9781-2024-7-81–89

1. Введение

Информационные технологии, проникая во все сферы жизни, изменяют способы взаимодействия между людьми, обеспечивая доступ к новым источникам информации и создавая новые формы коммуникации. Благодаря доступности интернета, развитию социальных сетей, мессенджеров и других цифровых инструментов коммуникации в текущий момент наблюдается резкое увеличение количества цифровых взаимодействий между людьми. Для индивидов присутствие виртуальной коммуникативной среды обеспечивает расширение круга общения и позволяет находить единомышленников и достигать общих целей вне пространственных, временных и социальных ограничений. С точки зрения социокультурной динамики это означает увеличение интерактивности общественного пространства через более активное участие людей в различных социальных и политических процессах. Таким образом, цифровые технологии и цифровые взаимодействия являются мощным инструментом создания новых связей между социальными группами и отдельными индивидами. При этом онлайн-среда представляет собой многочисленные горизонтальные связи, образующие масштабный неориентированный граф, именно эта структура и обеспечивает преодоление социальных барьеров и приводит к пересечению индивидов из различных слоев общества.

Коммуникация в виртуальной реальности имеет ряд важных характеристик, две из которых стратегически важны в рамках актуализации потенциала этой коммуникативной среды как объекта исследований. Во-первых, большинство цифровых взаимодействий происходит в форме обмена текстовыми сообщениями, что означает наличие зафиксированного в удобном формате цифрового следа; во-вторых, часто коммуникация происходит в публичной зоне (посты и комментарии к ним, чаты) и имеет ряд метаданных (например, дата поста, социально-демографические характеристики автора), а, значит, доступна для структурированного сбора данных и дальнейшего изучения компьютерными методами.

Таким образом, возрастающее количество публичных текстовых данных превратили анализ текстов (natural language processing, nlp) в актуальный метод для анализа социальной динамики общества. Основным объектом исследований стал контент из социальных и новых медиа. Стоит отметить, что nlp-методы применяются в различных областях, в том числе активно используются в сфере маркетинга для оптимизации пути клиента, в медицине для обеспечения дистанционного взаимодействия «клиника-пациент», но, главное, они позволяют анализировать поведение социальных групп на микро- и макроуровнях, привязывая его к временной шкале и событийным фактам. Так выявляются и объясняются скрытые закономерности, приводящие общество в движение, а также предсказываются вероятности наступления явления (например, общественных волнений или, наоборот, апатии социальной группы или общества в целом).

Одним из важных nlp-методов, который актуален для исследования медиапространства, является анализ тональности текста (sentiment analysis). Он дает возможность понять отношения, мнения и эмоции, лежащие в основе онлайн-текста. Формализуя понятие, можно дать следующее определение: анализ тональности — это класс методов контент-анализа в компьютерной лингвистике, основная задача которого заключается в классификации текста по его настроению. Обобщая тональность текстов, можно вычислять индекс субъективного благополучия, прогнозировать результаты выборов или экономических показателей, оценивать реакцию на события или новости.

По сути, тон текста помогает понять эмоциональное состояние автора и определить его отношение к поднятой теме. Так как любой публичный пост подразумевает наличие серии комментариев на него, то становится доступным целый ряд научных рефлексий: исследование общей реакции социальной группы на тему, анализ реакции активных акторов на проблему, детекция реакции пассивных акторов коммуникации на лидеров мнений.

2. Постановка проблемы

В простых случаях задача анализа тональности сводится к бинарной классификации текстов на две категории: позитивные и негативные (в ряде случаев также включают категорию «нейтральный текст»). Однако подобное разделение на 2-3 класса не всегда репрезентативно для выявления глобальных социальных закономерностей, и задача переформатируется в мультиклассовую классификацию, когда необходимо более четко определить эмоциональные состояния индивидов. В таком случае дополнительная фаза исследования отводится под разработку актуальной шкалы, способной связать в единую логику используемые для анализа данные и выявляемые закономерности. В таких целях может использоваться численная шкала или категории типа «страх», «злость», «печаль», «счастье». В результате могут быть вынесены суждения об индексе социального благополучия или векторе социального настроения. Такие классы легко связываются с количественными данными (например, в задачи социального картирования, где необходимо визуально показать взаимосвязь эмоций жителей страны, города или района и ряда количественных данных, отражающих различные характеристики жизни в этой

локации). Глобально анализ тональности текстов можно разделить на три направления методов:

1. Подходы на основе правил (rule-based). В них используются размеченные словари эмоций, для русского языка крупнейшими являются RuSentiLex и LINIS Crowd [1, 2], которые имеют информацию о привязке слов к категориям «позитивно» и «негативно», то есть не дают четких характеристик эмоций в отличие от англоязычных SenticNet, SentiWordNet и SentiWords [3, 4, 5]. Так же эта группа методов предполагает вручную созданные наборы правил классификации. Очевидным минусом подхода является низкая способность к обобщению (невозможно масштабировать для анализа текстов, не имеющих предсказуемой конкретной тематики).
2. Подходы на основе машинного обучения, которые подразумевают автоматическое извлечение признаков из текста, что позволяет анализировать тексты, относящиеся к разным тематикам, в едином конвейере. Часто используемыми в рамках анализа тональности моделями машинного обучения являются логистическая регрессия, дерево решений и метод опорных векторов. Последние несколько лет помимо классических алгоритмов машинного обучения для решения этой задачи применяются свёрточные (CNN) и рекуррентные (RNN) нейросети [6, 7]. Группа этих методов показывает хорошие результаты с точки зрения метрик качества (точность от 70% в зависимости от конкретной задачи) и масштабируемости (применимость для текстов разных типов и дискурсов).
3. Гибридные подходы, которые объединяют в себе первые два (примером может служить ALDONAr [8]).

Из перечисленных групп методов с точки зрения применимости для анализа процессов социальной динамики выделяются подходы на основе машинного обучения. Возможность их применения первично строится на необходимости переформатирования текста в числовые векторы, так как алгоритмы машинного обучения подразумевают манипуляции в математическом пространстве. К тому же идея заключается в том, что векторы (embedding), представленные в геометрическом пространстве, могут быть описаны через расстояние до соседей, что дает информацию об их взаимосвязях. Эмбеддинги слов могут быть созданы различными методами векторизации: самая простая из них — «мешок слов» (bag of words), также часто используется tf-idf векторизация, которая учитывает важность слова в документе, а не только частоту его появления. В более сложных системах для генерирования эмбеддингов слов применяются модели дистрибутивной семантики, например, Word2Vec, GloVe и FastText [9, 10, 11]. Существуют подходы к векторизации, позволяющие создать эмбеддинги предложений или параграфов (а не слов), к этой логике векторизации относятся, например, модели ELMo, BERT и LASER [12, 13, 14].

Стоит отметить, что, несмотря на частое появление в методологиях исследований компонента анализа тональности, метод не имеет четких рамок и устоявшихся правил использования: тональность, содержащуюся в тексте, можно анализировать на уровне бинарной классификации, или детализировать на несколько классов (часто используют пятибалльную шкалу). В зависимости от того, какая модель векторизации будет использована, примененный в дальнейшем алгоритм машинного обучения для задачи классификации тональности текста сработает точнее или наоборот более ординарно. Таким образом, исследование применимости моделей векторизации к конкретным типам текстов является актуальной исследовательской проблемой.

В рамках данного исследования была поставлена задача анализа успешности моделей векторизации применительно к коротким текстам из социальных сетей. Было решено сфокусироваться на бинарной классификации, так как основной вопрос: какая техника создания векторного представления точнее фиксирует особенности коротких текстов на русском языке (разговорного формата) с точки зрения возможности далее ml-моделью уловить негативные и позитивные тональности.

3. Данные

В настоящее время интерес представляют два социальных медиа: ВКонтакте (консервативная по формату социальная сеть, состоящая из пабликов и групп с наличием публичных постов и комментариев к ним) и Телеграмм (мессенджер с большим количеством публичных чатов, где обсуждение тем может развиваться параллельно, без наличия побуждающего нулевого поста). С точки зрения возможностей привязывать анализируемые текстовые данные к реальности (например, к геоданным) полезнее оказывается информация, полученная из социальной сети ВКонтакте. Поэтому для тестирования моделей векторизации было собрано два набора данных именно из этой сети: 1) датасет постов и комментариев к ним из публичных районных сообществ города Санкт-Петербурга (18 групп, выкачивались данные за 2019-2020 гг.); 2) датасет, составленный из контента пабликов «Подслушано» (4 группы, выкачивались данные за 2022 год). Средняя длина комментариев в первом наборе данных — 21 слово, а постов — 41 слово; во втором датасете средняя длина постов — 11 слов, комментариев к ним — 15 слов. Полярность тематик собранных датасетов — намеренная стратегия, так как дискурсы текстов и длина сообщений — важные характеристики, влияющие на подбор метода векторизации. Идея заключалась в том, чтобы проверить, будет ли какое-то заметное различие в метриках качества для двух датасетов. Общий объем данных — 319 335 записей. Собранные текстовые данные были поэтапно предобработаны по следующей схеме: 1) разбиение текстов реплик на токены; 2) удаление спецсимволов, эмодзи, ссылок и знаков пунктуации; 3) удаление стоп-слов; 4) нормализация токенов. На выходе из такого пайплайна препроцессинга был получен предобработанный текст, готовый к вероятностной векторизации. Также в рамках очистки датасетов от данных, не вносящих концептуальный вклад в эксперимент, посты (начало дискуссии по теме), не содержащие более пяти комментариев, были удалены. Это было сделано исходя из того, что для проводимого эксперимента была важна оценка тональности поста и серии комментариев к нему с точки зрения направленности социального настроения, таким образом, нейтрально окрашенные темы (например, обсуждение потерянных ключей или графика работы какого-то учреждения) не представляли для данного исследования интереса. После этого этапа предобработки данных были получены обновленные датасеты, общим объемом 204 107 строк.

4. Описание эксперимента

В качестве базовой модели векторизации была выбрана TF-IDF (учитывались биграммы). Также были обучены: модель Word2Vec (size = 100, sg = 1, min_count = 1, window = 5, учитывались биграммы) и базирующаяся на библиотеке глубокого обучения PyTorch модель LASER (использовалась предобученная модель для русского языка из библиотеки laserembeddings). Таким образом, эксперимент заключался в сравнении успешности трех основных подходов к созданию векторных представлений текстов.

Прежде чем обратиться к обучению классификатора, было решено исследовать данные методом кластеризации, чтобы удостовериться, что в текстах действительно есть закономерности, которые с математической точки зрения заметны. Кластеризация данных является классической задачей восстановления распределения данных: это дает понимание того, как объекты распределены в пространстве признаков, какие наиболее характерные значения у них есть, где объектов мало, а где они лежат плотным облаком. Таким образом, на первом этапе анализа данных мы использовали эмбеддинги текстов, полученные тремя способами, и было принято решение посмотреть способность анализируемых моделей эмбеддингов к разделению на кластеры. Кластеризация была проведена с использованием алгоритма K-средних (k=2), так как он позволяет задать количество искомых кластеров. Результаты разбиения на два кластера представлены на рис. 1. Для визуализации кластеризации использовался алгоритм понижения размерности PCA.

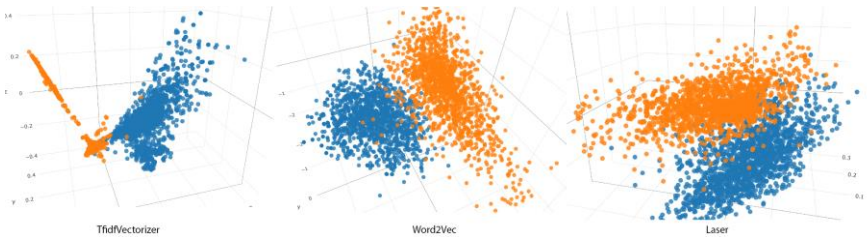

Рис. 1. Анализ разделимости классов с использованием 3D-сжатия векторного пространства с использованием алгоритма PCA

Графики показывают, что векторизация методом TF-IDF дает неплохие результаты: кластеры визуально выглядят сепарированными друг от друга. Векторизация с помощью Word2Vec и Laser гораздо хуже фиксирует особенности текстов, позволяющие их сепарировать как легко вчитывающиеся кластеры, по крайней мере при k=2. Стоит отметить, что кластерный анализ не дает точного представления о конкретных особенностях текстов: признаки, по которым алгоритм делит данные на группы, остаются не интерпретируемыми. Однако, этот опыт показывает, насколько векторное представление в принципе фиксирует полярность кластеров (по какому-либо признаку). Заметим, что неожиданным стала низкая репрезентативность Word2Vec-векторизации, так как этот метод обычно хорошо улавливает синонимичность слов, что, как следствие, помогает близкие по значению тексты отнести к одному кластеру.

На втором этапе эксперимента часть собранных данных была размечена вручную на два класса (негативный и позитивный).

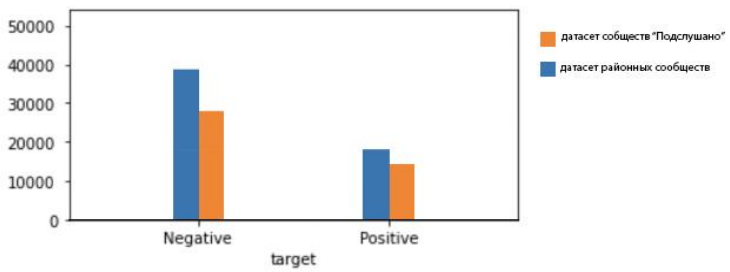

Рис. 2. Распределение размеченных классов в двух датасетах

После этого были выделены слова (и словосочетания), вносящие наибольший вклад в каждый из классов, и построены облака слов для обоих классов (рис. 3).

Рис. 3. Датасет районных сообществ: облако слов, вносящих наибольший вклад в «негативный» класс (слева); облако слов, вносящих наибольший вклад в «позитивный» класс (справа)

Такое визуальное представление классов дает возможность выдвинуть важную гипотезу: «негативные» тексты гораздо важнее правильно детектировать нежели «позитивные», так как они явно более содержательны с точки зрения возможностей

дальнейшего анализа контекстов. То есть, выявив «негативный» класс далее отдельно с ним можно проводить дополнительные исследования: тематическое моделирование, выделение именованных сущностей, анализ тональности уже с мультиклассовым разделением на эмоции. Соответственно, для оценки качества работы модели классификатора можно использовать матрицу ошибок. Она дает информацию о процентном содержании истинно-положительного, истинно-отрицательного, ложно-положительного и ложно-отрицательного решений классификатора. Таким образом, отдельно от общей производительности модели, становится возможным проверить, в скольких случаях был спрогнозирован «негативный» класс, и это оказалось правдой.

Далее размеченный набор данных был разделен на обучающую и тестовую выборку (пропорция 70% и 30%). В качестве алгоритма классификации была выбрана логистическая регрессия. На рис. 4 представлены результаты работы логистической регрессии при анализе датасета районных сообществ.

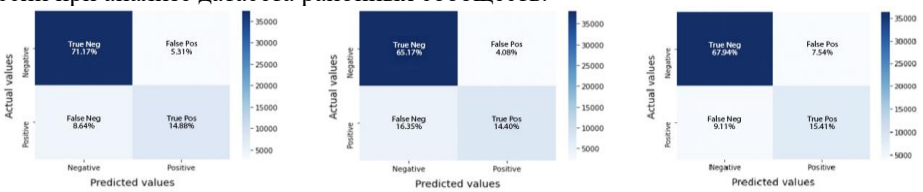

Рис. 4. Результаты работы модели логистической регрессии при отправленных в нее векторных представлениях, полученных тремя способами (слева направо: tf-idf, w2v, Laser)

Как видно из результатов, все три метода векторизации сработали достаточно хорошо. Однако различия касаются степени ошибок первого (ложно-положительное решение) и второго (ложно-отрицательное решение) рода. По приведенным матрицам видно, что лучше всего «негативный» класс детектируется логистической регрессией, работающей на векторном представлении tf-idf. Удивительным фактом является то, что модель векторизации Laser сработала достаточно хорошо, это значит, что для анализа тональности коротких текстов актуальным подходом может быть векторизация целых предложений.

Эксперименты над вторым датасетом дали схожие результаты, при этом стоит отметить, что Laser показал лучший результат относительно tf-idf, а модель Word2Vec осталась на третьем месте (56,17% истинно-отрицательных решений, что немного хуже, чем это было на данных из районных сообществ). Таким образом, появляется гипотеза, что чем менее текст насыщен контекстов (и присутствует только эмоция), тем хуже Word2Vec способствует улавливанию нюансов, важных для классификации по тону сообщения. И в то же время Laser, вероятно, показывает наиболее успешные результаты в рамках векторизации коротких текстов, чем меньше в них присутствует категория «содержание».

5. Заключение

При сравнении моделей векторизации TF-IDF показала лучшую способность улавливать необходимые особенности в коротких текстах. Модель логистической регрессии с использованием данного векторного представления показала хорошую итоговую производительность (F1_score=0,81), к тому же именно эта векторизация позволяет при анализе тональности точнее детектировать «негативный» класс, который, как было показано выше, является более интересным с точки зрения дальнейших поисков закономерностей при анализе социального настроения. Однако стоит отметить, что Word2vec предоставляет дополнительные инструменты анализа текстов (благодаря учету квази-синонимичности) и, соответственно, при определенной постановке задачи может быть полезным. Касательно целесообразности использования Laser, стоит дополнительно отметить, что модель требует больших ресурсов оперативной памяти (и сама векторизация

занимает достаточно длительное время), однако само векторное представление показало неплохие результаты при использовании в классификаторе.

В дальнейшем планируется доразметить собранные наборы данных, составить из них датасет, содержащий уравновешенное количество примеров из обоих классов, и затем повторить опыт с теми же настройками, что описаны в данном эксперименте (так как некоторая туманность в оценке «позитивного» класса на данный момент остается).

Исследование выполнено при поддержке Российского научного фонда и Санкт-Петербургского научного фонда, грант № 23-28-10069 «Прогнозирование социального самочувствия с целью оптимизации функционирования экосистемы городских цифровых сервисов Санкт-Петербурга» (https://rscf.ru/project/23-28-10069/).

Литература

[1] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 1171–1176.

[2] Koltsova O. Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. — 2016. — Vol. 2016. — P. 277–287.

[3] Cambria E., Poria S., Bajpai R., Schuller B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. — 2016. — P. 2666–2677.

[4] Baccianella S. et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining // Lrec. — 2010. — Vol. 10. — № 2010. — P. 2200–2204.

[5] Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Transactions on Affective Computing. — 2015. — Vol. 7. — №. 4. — P. 409–421.

[6] Baziotis C. et al. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns // arXiv preprint arXiv:1804.06659. — 2018.

[7] Baziotis C., Pelekis N., Doulkeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis // Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). — 2017. — P. 747–754.

[8] Meškelė D., Frasincar F. ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model // Information Processing & Management. — 2020. — Vol. 57. — №. 3. — Art. 102211.

[9] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — Vol. 26. — P. 3111–3119.

[10] Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.

[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — Vol. 2. — P. 427–431.

[12] Lee K., Filannino M., Uzuner Ö. An Empirical Test of GRUs and Deep Contextualized Word Representations on De-Identification // MedInfo. — 2019. — P. 218–222.

[13] Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2018. — Vol. 1. — P. 4171–4186.

[14] Chizhik A., Zherebtsova Y. Challenges of Building an Intelligent Chatbot // International Conference «Internet and Modern Society» (IMS-2020). CEUR Proceedings. — 2021. Vol. 2813. — P. 277–287.

Comparison of Text Vectorization Models for the Sentiment Analysis of Short Messages from Social Media

Anna V. Chizhik

ITMO University

Sentiment analysis is one of the urgent tasks that can identify important factors which affect the vector of the social mood. When using machine learning methods to solve this problem, it is required to convert the text into its vector representation. There are a number of text vectorization methods. This paper compares three currently relevant approaches to creating a vector representation: taking into account the weight of a word in a document (TF-IDF), using distributive semantics when creating word embeddings (Word2Vec), and sentence embedding models (Laser). Comparing these three text vectorization models for the task of analyzing the sentiment of short messages from social networks, it is obvious that each of them has its own advantages and disadvantages. The paper describes the design of the study, provides quality metrics, describes the data on which the experiments were conducted.

Keywords: text vectorization, embeddings, sentiment analysis, social media

Reference for citation: Chizhik A.V. Comparison of Text Vectorization Models for the Sentiment Analysis of Short Messages from Social Media // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 81–89. DOI: 10.17586/2541-9781-2024-7-81–89

Reference

[1] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 1171–1176.

[2] Koltsova O. Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. — 2016. — Vol. 2016. — P. 277–287.

[3] Cambria E., Poria S., Bajpai R., Schuller B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. — 2016. — P. 2666–2677.

[4] Baccianella S. et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining // Lrec. — 2010. — Vol. 10. — № 2010. — P. 2200–2204.

[5] Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Transactions on Affective Computing. — 2015. — Vol. 7. — №. 4. — P. 409–421.

[6] Baziotis C. et al. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns // arXiv preprint arXiv:1804.06659. — 2018.

[7] Baziotis C., Pelekis N., Doulkeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis // Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). — 2017. — P. 747–754.
[8] Meškelė D., Frasincar F. ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model // Information Processing & Management. — 2020. — Vol. 57. — №. 3. — Art. 102211.
[9] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — Vol. 26. — P. 3111–3119.
[10] Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.
[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — Vol. 2. — P. 427–431.
[12] Lee K., Filannino M., Uzuner Ö. An Empirical Test of GRUs and Deep Contextualized Word Representations on De-Identification // MedInfo. — 2019. — P. 218–222.
[13] Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2018. — Vol. 1. — P. 4171–4186.
[14] Chizhik A., Zherebtsova Y. Challenges of Building an Intelligent Chatbot // International Conference «Internet and Modern Society» (IMS-2020). CEUR Proceedings. — 2021. — Vol. 2813. — P. 277–287.

Поддержка модели превентивной медицины: модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент»

А. В. Чижик[1,2], М. П. Егоров [1],
М. Ю. Якубова[1], Д. А. Погребной[1], А. С. Кривошапкина[1]

[1] Университет ИТМО, [2] Санкт-Петербургский государственный университет

chizhik@itmo.ru, egorovm@niuitmo.ru, shentorin@gmail.com,
pogrebnoy.inc@gmail.com, aitalina.kr@gmail.com

Аннотация

Модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент» является важным инструментом в поддержке модели превентивной медицины, так как позволяет улучшить качество обслуживания пациентов и повысить степень их участия в предупреждении заболеваний. В статье описывается разработанный модуль, с помощью которого можно детектировать симптомы и их отрицания, затем на этом основании выносить предварительный диагноз и маркер срочности приема. Авторами описывается общий алгоритм, особенности моделей машинного обучения, которые вкачены в общий конвейер работы модуля, приводятся метрики качества.

Ключевые слова: здравоохранение, разговорный ИИ, языковые модели, человекомашинный диалог

Библиографическая ссылка: Чижик А. В., Егоров М. П., Якубова М. Ю., Погребной Д. А., Кривошапкина А. С. Поддержка модели превентивной медицины: модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент» // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединенной научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 90–96. DOI: 10.17586/2541-9781-2024-7-90–96

1. Введение

Здоровье населения является одним из факторов, влияющих на полюс, к которому стремится социальное настроение, и с этой позиции оно является достаточно сложным феноменом, определяющимся через взаимодействие социальных, психологических, экономических и только затем биологических, генетических и физиологических факторов. Иными словами, во многом уровень здоровья населения зависит от мировоззренческого контекста, присутствующего в обществе и способствующего выстраиванию постоянной и доверительной коммуникации с медицинскими учреждениями. Предпосылкой для проведения этого исследования стал опыт создания мультимодального интеллектуального помощника для автоматизации процесса приема пациентов и оказания первичной медицинской помощи. Системы здравоохранения во всем мире используют автоматизацию для решения проблемы нехватки персонала, а также для эффективного управления и сортировки пациентов в больших масштабах в больницах и клиниках. Одним из видов автоматизации являются чат-боты и сопутствующие технологии, позволяющие автоматизировать процесс первичного общения с потенциальным

пациентом (включая фиксацию симптомов на стороне клиники и предоставление человеку необходимой справочной информации). Это позволяет преодолеть географические и временные барьеры между службами здравоохранения и их пользователями. Таким образом, конечной целью этих усилий является переход от неотложной помощи к профилактической, что возможно только на основе управления взаимодействием на основе данных. Важно отметить, что есть новые социологические исследования, которые показывают тенденцию: каждый пятый врач покидает профессию в течение двух лет по причине профессионального выгорания, что объясняется большим количеством рутинной деятельности (преимущественно офисного характера) [1, 2]. Чат-боты могут облегчить нагрузку на врачей, медсестер и других медицинских работников, автоматизируя задачи, которые лучше подходят для компьютера, и в результате освобождая медицинские бригады для более продуктивного выполнения своей основной работы.

Проведя серию экспериментов по созданию чат-бота с открытым доменом для общения пациента с клиникой, мы поняли, что невозможно создать полностью универсальный диалоговый агент, который любая клиника могла бы запустить в работу без дополнительных усилий. В то же время стало понятно, что клиникам нужны готовые модули, из которых можно было бы собрать нужную функциональную конфигурацию диалогового агента непосредственно на стороне клиники.

Поэтому мы решили создать библиотеку для языка Python, которая могла бы:

- обнаружить симптомы в реплике пользователя (на вход языковая модель получает короткий текст на ЕЯ, содержащий ответы со стороны пользователя на вопросы бота порядка «опишите свое самочувствие»);
- использовать эту информацию для вынесения интерпретируемого суждения о возможном диагнозе (т. е. диагноз и сопроводительную информацию о вероятностном распределении — концепция «второго мнения»);
- присвоить пациенту метку срочности (мультиклассификация пациентов для распределения потоков пациентов в клинике).

На рис. 1 показана логика разработанного модуля.

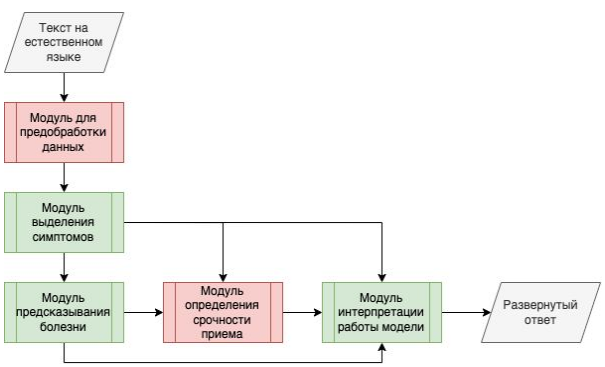

Рис. 1. Блок-схема разрабатываемого модуля

В рамках поднимаемой проблемы актуальными являются исследования, посвященные методам обработки и понимания естественного языка (NLP, natural language processing и NLU, natural language understanding). При разработке диалогового агента обычно решаются следующие классические задачи обработки естественного языка: сегментация, токенизация и лемматизация, NER и нахождение семантических связей [2]. Существуют и специфические задачи [3]: обработка последовательности из нескольких фраз, дополняющих друг друга; поиск ссылок с одной фразы на другую; обработка чередования разных типов интентов подряд; генерация уточняющих вопросов и их обработка. В нашем исследовании мы в первую очередь сосредоточились на проблеме выявления симптомов

(и их отрицания) и последующем использовании этой информации для определения диагноза [4, 5]. Следует отметить, что задача создания специализированных языковых моделей является достаточно динамично развивающейся, в частности, можно упомянуть следующие два современных исследования (относящихся к области медицины) [6, 7, 8], подход которых заключается в использовании условно закрытых данных (электронные медицинские карты пациентов, ЭМК).

2. Данные

Ролевая модель дистанционного взаимодействия между клиникой и пациентом подразумевает, что диалог строится в формате разговорного русского языка, свойственного социальным сетям (так как сам интерфейс любого диалогового агента напоминает мессенджер). Таким образом, при формировании набора текстовых данных необходимо стремиться к близости собираемых реплик к языку пациентов, а не к служебному языку медиков. Поэтому мы решили отойти от общей тенденции использования при разработке подобных модулей текстовых данных, взятых из ЭМК (анамнез и диагнозы), и собрали данные из открытых веб-источников:

- основа набора данных — 5 193 описания пациентов своих заболеваний с маркером категории болезни (источник: https://meduniver.com);
- датасет дополнен 292 заболеваниями с их описанием (источник: https://health.mail.ru/disease/adneksit/);
- также были собраны данные о симптомах из Википедии (272 симптома).

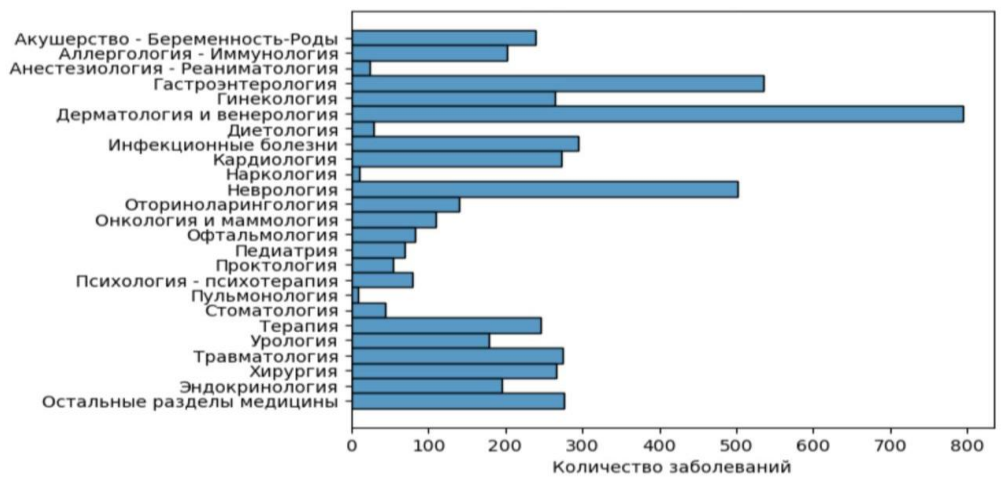

Рис. 2. Распределение категорий болезней

На рис. 2 показано распределение заболеваний по медицинским категориям. В ходе анализа данных, было выяснено, что в среднем пациент упоминает 2-3 симптома, присутствующих в его самочувствии, и тратит на описание около 66 слов.

3. Метод

На рис. 3 показана логика взаимодействия с текстовыми данными, которой мы придерживались при разработке данного модуля.

Было решено разработать систему подмодулей, что должно обеспечить возможность использовать библиотеку не только в полном функционале, но и частично, например, только для выделения симптомов.

Рис. 3. Пайплайн обработки и классификации медицинских текстов: текущая реализация

Процесс предобработки данных в рамках нашего проекта практически ничем не отличался от стандартного набора процедур, однако в отличие от классических подходов к этому этапу, мы решили сохранять некоторые стоп-слова, чтобы не потерять отрицание симптома (исходя из того, что отрицание симптома тоже является симптомом).

Далее нам потребовался список симптомов для их последующего извлечения из текстов. Хотя существуют методологии детекции ключевых слов, которые можно применить, в нашем случае они оказались не очень полезными. Поэтому был использован готовый список симптомов из открытой базы знаний «Википедия». Далее для формирования необходимой информации мы использовали фреймворк Scapy. Он может извлекать необходимые объекты из текста, используя предварительно обученную модель машинного обучения. ML-модель, доступная во фреймворке по умолчанию, не смогла справиться с большинством симптомов. Поэтому был создан некоторый набор правил-подсказок, чтобы помочь модели. Каждая такая подсказка — паттерн, написанный отдельно для каждого симптома. Как уже было отмечено выше, перед нами стояла задача детектирования отрицания симптомов. Для этой цели мы использовали пакет Python negex, который работает со всеми найденными сущностями и пытается найти отрицание для каждой из них (это реализуется за счет определения границ частей предложения и поиска в этих границах специальных слов и других признаков, полезных для задачи поиска отрицания).

После этих двух шагов у нас появился общий анамнез, сформированный на основании данных основного датасета, который включил все возможные симптомы со статусами (yes, no, no_info, confused). Диаграмма этого процесса представлена на рис. 4.

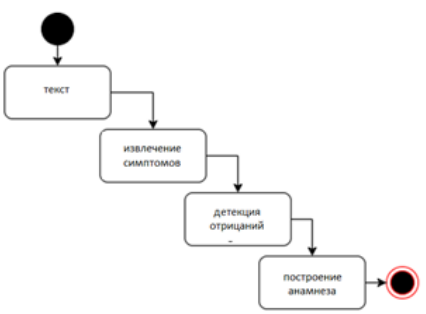

Рис. 4. Логика формирования набора симптомов со статусами

В моменте взаимодействия с репликой пользователя появляется задача мультиклассовой классификации, которая в нашем модуле решается с помощью модели логистической регрессии.

Отметим, что логистическая регрессия — это статистическая модель, которая используется для прогнозирования вероятности возникновения некоторого события, в данном случае для диагностики конкретного заболевания. Модель применяется к подготовленным данным и подразумевает создание матрицы признаков, в которой каждый столбец представляет определенный симптом, а каждая строка представляет конкретный случай пациента. Значения в ячейках матрицы указывают на наличие (1), отсутствие (0) или отрицание (-1) симптома. Используя подготовленную матрицу признаков и соответствующие метки классов (заболеваний), модель логистической регрессии обучается. В процессе обучения модель определяет оптимальные веса для каждого признака (симптома), которые позволяют наиболее точно классифицировать заболевания.

Преимуществом логистической регрессии является интерпретируемость результатов. Веса, присвоенные каждому симптому, отражают важность этого симптома в определении заболевания. Таким образом, врачи и другие медицинские работники могут анализировать эти веса и понимать логику результата модели. Более того, логистическая регрессия учитывает наличие, отсутствие и отрицание симптомов, что делает результаты еще более точными и надежными. Из вышеизложенного ясно, что нам становится легко получить удобочитаемую интерпретацию диагноза пациента. Текущая точность модели составляет 86%.

Можно обозначить конечной целью нашего модуля определение срочности приема/госпитализации пациента. Система здравоохранения в России предполагает 3 формы помощи: экстренную, срочную и плановую. Поэтому, чтобы разметить данные по срочности, мы собрали симптомы из различных открытых источников по категориям экстренных, срочных и плановых приемов. Для тестирования точности наших моделей машинного обучения мы сосредоточились на заболеваниях из категории «кардиология». Размеченные на три класса срочности симптомы были верифицированы на предмет применимости к задаче кардиологами Национального медицинского исследовательского центра им. В. А. Алмазова. Отметим, при отнесении случая к первым двум категориям (экстренный и срочный прием) пациенту требуется госпитализация, поэтому было принято решение маркировать данные бинарно: госпитализация требуется или не требуется.

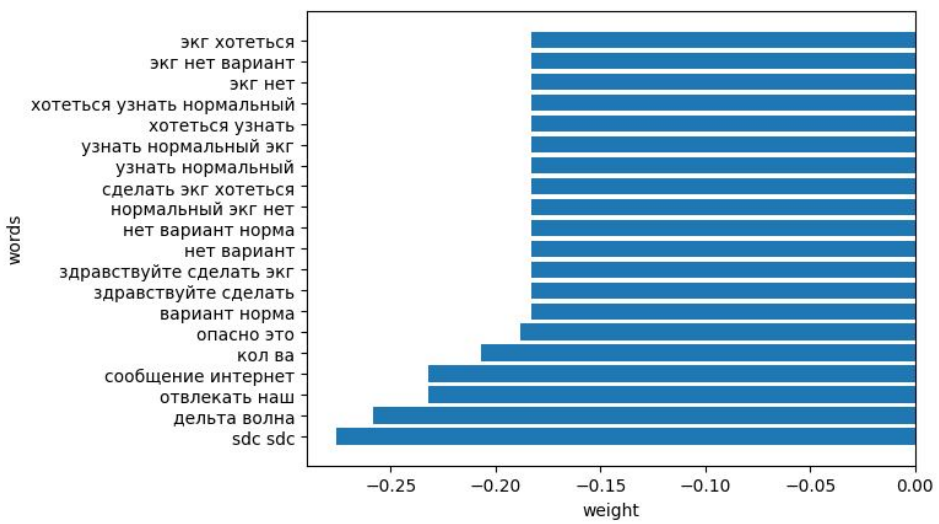

Рис. 5. Содержания класса «плановый прием»

Из предобработанных симптомов были составлены биграммы и триграммы. Дальнейшая логика была такова: если в предобработанном тексте пациента присутствовал хотя бы один из этих 2-3-граммов, то кейс маркировался как срочный. В результате мы получили датасет, содержащий 41 срочный случай и 231 плановый. Базовой идеей было использование модели логистической регрессии и tf-idf векторизатора. Однако модель переобучилась из-за несбалансированности классов и присутствию в «несрочном» классе большого количества шума (класс содержал просьбы пациентов интерпретировать результаты анализов) — рис. 5.

Поэтому следующим шагом стало обучение нейросети на Self Attention. Для каждого слова были взяты word2vec-эмбеддинги, затем Self Attention был использован для анализа контекста каждого слова. В итоге получены следующие метрики качества модели маркировки срочности: Accuracy = 0.93 и F1 = 0.96.

4. Заключение

В настоящее время коллектив авторов работает над улучшением значений метрик качества и планирует измерять качество модуля за счет привлечения медицинских экспертов для тестирования. Кроме того, наборы данных планируется дополнить новыми случаями. На наш взгляд, текущие тесты показывают, что модуль применим на практике. Наборы данных и библиотека находятся в свободном доступе на github (https://github.com/NIRMA-PATIENT-INTAKE).

Исследование проведено в рамках НИР Университета ИТМО № 622275 «Разработка модуля для предсказания предварительного диагноза: поддержание логистики потоков пациентов и концепции второго мнения при взаимодействии с пациентом через диалоговые системы».

Литература

[1] Sinsky C. A., Brown R. L., Stillman M. J., Linzer M. COVID-Related Stress and Work Intentions in a Sample of US Health Care Workers // Mayo Clinic Proceedings: Innovations, Quality & Outcomes. 2021. Vol. 5 (6). P. 1165–1173.

[2] Sinsky C., Colligan L., Li L., Prgomet M., Reynolds S., Goeders L., Westbrook J., Tutty M., Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties // Ann Intern Med. 2021. Vol. 165 (11). P. 753–760. DOI: 10.7326/M16-0961.

[3] Lalwani T. et al. Implementation of a Chatbot System using AI and NLP // International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 2018. Vol. 6 (3). P. 26–30. DOI: 10.2139/ssrn.3531782.

[4] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall, 2008.

[5] Freedman M. S., Gray T. A. Vascular headache: a presenting symptom of multiple sclerosis // Canadian journal of neurological sciences. 1989. Vol. 16 (1). P. 63–66.

[6] Chizhik A., Egorov M. Multimodal Intelligent Assistants for Automating the Patient Intake Process and Primary Care // 15th International Conference on Theory and Practice of Electronic Governance. — 2022. — P. 573–575.

[7] Legnar M. et al. Natural Language Processing in Diagnostic Texts from Nephropathology // Diagnostics. — 2022. — Vol. 12 (7). — Art. 1726. DOI: 10.3390/diagnostics12071726.

[8] Zhou B. et al. Natural language processing for smart healthcare // IEEE Reviews in Biomedical Engineering. — 2022. DOI: 10.48550/arXiv.2110.15803.

Support for the Preventive Medicine Model: Natural Language Processing Module for Remote Clinic-Patient Interaction

Anna V. Chizhik [1,2], Michil P. Egorov [1], Maria Yu. Yakubova [1], Dmitrii A. Pogrebnoi [1], Aitalina S. Krivoshapkina [1]

[1] ITMO University, [2] St.Petersburg State University

The natural language processing module for remote clinic-patient interaction is an important tool in supporting the model of preventive medicine, as it allows you to improve the quality of patient care and increase their degree of participation in disease prevention. The article describes the developed module, which can be used to detect symptoms and their denials, then, on this basis, make a preliminary diagnosis and a marker of the urgency of admission. The authors describe the general algorithm, the features of machine learning models that are injected into the general pipeline of the module, and provide quality metrics.

Keywords: health service, conversational AI, language models, human machine dialogue

Reference for citation: Chizhik A. V., Egorov M. P., Yakubova M. Yu., Pogrebnoi D. A., Krivoshapkina A. S. Support for the Preventive Medicine Model: Natural Language Processing Module for Remote Clinic-Patient Interaction // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). - St. Petersburg: ITMO University, 2024. P. 90–96. DOI: 10.17586/2541-9781-2024-7-90–96

Reference

[1] Sinsky C. A., Brown R. L., Stillman M. J., Linzer M. COVID-Related Stress and Work Intentions in a Sample of US Health Care Workers // Mayo Clinic Proceedings: Innovations, Quality & Outcomes. 2021. Vol. 5 (6). P. 1165–1173.

[2] Sinsky C., Colligan L., Li L., Prgomet M., Reynolds S., Goeders L., Westbrook J., Tutty M., Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties // Ann Intern Med. 2021. Vol. 165 (11). P. 753–760. DOI: 10.7326/M16-0961.

[3] Lalwani T. et al. Implementation of a Chatbot System using AI and NLP // International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 2018. Vol. 6 (3). P. 26–30. DOI: 10.2139/ssrn.3531782.

[4] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall, 2008.

[5] Freedman M. S., Gray T. A. Vascular headache: a presenting symptom of multiple sclerosis // Canadian journal of neurological sciences. 1989. Vol. 16 (1). P. 63–66.

[6] Chizhik A., Egorov M. Multimodal Intelligent Assistants for Automating the Patient Intake Process and Primary Care // 15th International Conference on Theory and Practice of Electronic Governance. — 2022. — P. 573–575.

[7] Legnar M. et al. Natural Language Processing in Diagnostic Texts from Nephropathology // Diagnostics. — 2022. — Vol. 12 (7). — Art. 1726. DOI: 10.3390/diagnostics12071726.

[8] Zhou B. et al. Natural language processing for smart healthcare // IEEE Reviews in Biomedical Engineering. — 2022. DOI: 10.48550/arXiv.2110.15803.

Сведения об авторах

Быкова Анна Павловна, Санкт-Петербургский государственный университет, студент, ORCID 0009-0002-8887-2876.

Гаврилкина Анастасия Сергеевна, Национальный исследовательский ядерный университет «МИФИ», инженер, ORCID 0000-0003-2167-1287.

Голицына Ольга Леонидовна, кандидат технических наук, Национальный исследовательский ядерный университет «МИФИ», доцент, ORCID 0000-0002-3848-4755.

Егоров Мичил Прокопьевич, Университет ИТМО, студент, ORCID 0000-0002-0125-7540.

Клименко Екатерина Владиславовна, Санкт-Петербургский государственный университет, студент.

Кривошапкина Айталина Сергеевна, Университет ИТМО, студент, ORCID 0000-0001-6504-8038.

Лебедев Александр Анатольевич, Национальный исследовательский ядерный университет «МИФИ», ведущий математик, ORCID 0000-0002-3780-8092.

Локалина Юлия Сергеевна, Санкт-Петербургский государственный университет, студент.

Максимов Николой Вениаминович, доктор технических наук, профессор, Национальный исследовательский ядерный университет «МИФИ», профессор, ORCID 0000-0002-8191-1521.

Мельничук Дмитрий Вадимович, кандидат физико-математических наук, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, доцент, ORCID 0000-0002-6689-8904.

Носкина Анастасия Викторовна, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, студент, ORCID 0009-0005-2105-9292.

Погребной Дмитрий Андреевич, Университет ИТМО, студент, ORCID 0000-0001-6392-8445.

Татур Екатерина Михайловна, Санкт-Петербургский государственный университет, студент.

Ходоровский Леонард Абрамович, кандидат технических наук, доцент, Санкт-Петербург, ORCID 0009-0000-1593-9027.

Чижик Анна Владимировна, кандидат культурологии, Санкт-Петербургский государственный университет, старший преподаватель, ORCID 0000-0002-4523-5167.

Якубова Мария Юрьевна, Университет ИТМО, студент, ORCID 0009-0005-6427-0485.

Авторский указатель

Содержание

Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26 — 28 июня 2023 г. Сборник научных трудов). — СПб.: Университет ИТМО, 2024. — 100 с.

Компьютерная лингвистика и вычислительные онтологии

Выпуск 7

Сборник научных трудов

Под редакцией В. П. Захарова и А. В. Чижик
Дизайн обложки С. Н. Ушаков
Оригинал-макет П. В. Мякишева, А. С. Метелева
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н. Ф. Гусарова
Подписано к печати 05.06.24
Заказ 4772 от 05.06.24
Тираж 100 экз.

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