На правах рукописи

Бручес Елена Павловна

МЕТОДЫ И АЛГОРИТМЫ РАСПОЗНАВАНИЯ И СВЯЗЫВАНИЯ СУЩНОСТЕЙ ДЛЯ ПОСТРОЕНИЯ СИСТЕМ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ НАУЧНЫХ ТЕКСТОВ

Специальность 05.13.17 – Теоретические основы информатики

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск – 2021

Работа выполнена в федеральном государственном бюджетном учреждении науки Институте систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН).

Научный руководитель:	**Батура Татьяна Викторовна** кандидат физико-математических наук, доцент, исполняющий обязанности заведующего лабораторией моделирования сложных систем, федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН)
Официальные оппоненты:	**Бобров Леонид Куприянович** доктор технических наук, профессор кафедры прикладной информатики, федеральное государственное бюджетное образовательное учреждение высшего образования «Новосибирский государственный университет экономики и управления «НИНХ» (НГУЭУ) **Саломатина Наталья Васильевна** кандидат физико-математических наук, старший научный сотрудник лаборатория анализа данных, федеральное государственное бюджетное учреждение науки Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук (ИМ СО РАН)
Ведущая организация:	Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ)

Защита состоится «17» марта 2022 года в 14 ч. 00 мин. на заседании диссертационного совета Д 219.005.02 в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет телекоммуникаций и информатики» (СибГУТИ) по адресу: 630102, г. Новосибирск, ул. Кирова, 86, ауд. 625.

С диссертацией можно ознакомиться в библиотеке СибГУТИ и на сайте: http://www.sibsutis.ru/science/postgraduate/dis_sovets/.

Автореферат разослан «__» г.

Ученый секретарь
диссертационного совета,
д.т.н., доцент Нечта И.В.

Общая характеристика работы

Актуальность темы исследования. В связи с бурным ростом количества данных, в том числе и текстовых, активно развивается область обработки естественных языков. Решение таких задач позволяет более эффективно анализировать информацию для своих целей, экономя силы и время.

В последнее время особый интерес представляет автоматический анализ научных публикаций. Согласно исследованиям, ежегодное количество публикаций с 2008 г. до 2018 г. выросло с 1.8 миллиона до 2.6 миллионов статей [White K., 2019]. Очень важно следить за трендами и исследованиями в научных статьях, сравнивать предлагаемые методы для тех или иных задач, находить нужную информацию и многое другое. Очевидно, что проделать всю эту работу вручную невозможно, именно поэтому разработка инструментов для текстов научной тематики сегодня является одной из самых актуальных задач.

Стоит отметить, что такие тексты отличаются от остальных особой морфологией и лексикой, а также определёнными синтаксическими и семантическими структурами. Кроме того, тексты научных статей состоят из блоков, которые располагаются в общепринятом порядке: так, например, сначала идёт название статьи, авторы и их аффилиации, затем аннотация статьи; основной текст состоит, как правило, из введения, обзора работ по данной теме, описания предложенного метода, результатов, заключения и списка литературы. Такое деление на блоки упрощает поиск нужной информации не только для человека, но и при автоматической обработке текстов.

Существует много работ, посвящённых извлечению различной информации из научных текстов: библиографических данных, условий экспериментов, наборов данных и полученных метрик, изображений и таблиц и др.

Современные подходы для решения таких задач подразумевают использование алгоритмов машинного обучения. Качество таких алгоритмов напрямую зависит от качества данных, которые используются для их обучения. Для подготовки и разметки данных необходимо наличие специалистов и времени. Поэтому сегодня особенно актуальными являются методы, не требующие большого количества размеченных данных.

Задача извлечения информации из текстов является не только важной задачей самой по себе, но также и основным этапом для других задач (например,

автоматического реферирования), поэтому требуется высокое качество её решения. Можно сказать, что эта задача хорошо решается для английского языка, что связано с наличием большого количества данных, исследователей, вовлечённых в работу, и пр. Но использовать такие системы для русского языка представляется невозможным, т.к. русский язык имеет свои морфологические и синтаксические особенности, которые должны учитываться при разработке подобных алгоритмов.

Более того, русский язык считается малоресурсным – это означает, что количество данных (не только размеченных, но и неразмеченных) существенно ниже, чем для английского языка. Это тоже вызывает сложности при построении систем для решения любых задач обработки текстов для русского языка.

Эти факты обуславливают ***актуальность темы исследования***. В данной диссертационной работе рассмотрены методы и алгоритмы для решения нескольких задач извлечения информации, которые не требуют большого количества вручную размеченных данных. Полученные результаты показали, что при полном отсутствии вручную размеченных данных возможно разработать систему извлечения информации с достаточным качеством для применения на практике.

Степень разработанности темы исследования. В последнее время наблюдается рост публикаций, посвященных анализу именно научных текстов.

Исследования по извлечению информации различного рода представлены в работах В.Д. Гусева, Н.В. Саломатиной, Tkaczyk D., Epp S., Foppiano L. и др.

Извлечение научных терминов исследуется в трудах Н.В. Лукашевич, Е.И.Большаковой, Kucza M., Niehues J. и др.

Извлечение отношений в научных текстах является тесно связанной с извлечением терминов и решается такими исследователями, как Hearst M., Huang K., Wang G. и др.

Также в последнее время особое внимание уделяется задаче одновременного извлечения сущностей и отношений между ними, например, в работах Ryuichi T., Tianyang Z., Eberts M., Ulges A. и др.

Объектом исследования являются тексты научных статей на русском языке.

Предметом исследования являются методы автоматического извлечения информации из текстов на естественном языке.

Цель и задачи работы. Целью работы является исследование и разработка методов, применяемых для решения задач извлечения терминов и семантических отношений между ними, а также связывания их с внешней базой знаний, и реализация

основных компонентов системы извлечения информации из научных текстов на русском языке.

Требования к предлагаемым алгоритмам:

1. Реализация в условиях недостаточного количества размеченных данных;
2. Независимость от области знаний.

Для достижения поставленной цели были определены следующие задачи:

1. Предложить и реализовать метод извлечения научных терминов, слабо зависящий от области знаний;
2. Адаптировать метод извлечения отношений между терминами, основанный на переносе обучения моделей с английского языка на русский в постановке zero-shot learning;
3. Описать алгоритм и реализовать метод связывания терминов с сущностями в базе знаний;
4. Разработать методику разметки корпуса текстов на русском языке для обучения и оценки качества алгоритмов и методов;
5. Разработать программный комплекс для извлечения терминов и отношений из научных текстов и связывания терминов с внешней базой знаний.

Соответствие диссертации паспорту научной специальности. Диссертация соответствует области исследований специальности 05.13.17 – Теоретические основы информатики по п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»; п. 6 «Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке»; п. 12 «Разработка математических, логических, семиотических и лингвистических моделей и методов взаимодействия информационных процессов, в том числе на базе специализированных вычислительных систем».

Методы исследования. Методологической основой исследования являются методы компьютерной лингвистики, статистические методы и методы машинного обучения, успешно зарекомендовавшие себя в задачах анализа текстов. Для программной реализации системы использовались методы объектно-ориентированного программирования.

Научная новизна работы заключается в следующем:

1. Предложен новый метод извлечения терминов из научных текстов, основанный на частичном обучении, который может применяться к текстам разных областей знаний.
2. Разработан и реализован метод извлечения семантических отношений, позволяющий решать задачу в условиях ограниченного количества размеченных данных. Метод основан на технике "обучения без примеров" (zero-shot learning) путем переноса обучения моделей с английского языка на русский и потенциально применим для широкого круга малоресурсных языков.
3. Разработана методика подготовки и разметки данных. В ходе исследования подготовлен корпус текстов на русском языке, который содержит трехуровневую разметку и служит основой для обучения и оценки качества современных автоматических методов извлечения информации.

Теоретическая ценность и практическая значимость состоит в том, что в работе даны формальные описания предлагаемых алгоритмов и методов. На базе разработанных методов создан программный комплекс для извлечения информации из научных текстов на русском языке. Разработанные методы, алгоритмы и программное обеспечение могут применяться для построения систем машинного понимания текста, систем автоматической обработки текста, информационно-поисковых систем и других информационных систем, основанных на знаниях. Предложенные методы могут быть легко адаптированы к текстам других областей знаний.

Полученная система использовалась в работе, которая ведётся в рамках проекта РФФИ № 19-07-01134 «Создание моделей, методов и программных средств анализа текстов на естественном языке для использования в интеллектуальных информационных системах», а также поддержана стипендией Правительства Российской Федерации для студентов высшего профессионального образования и аспирантов, обучающихся по имеющим государственную аккредитацию образовательным программам, соответствующим приоритетным направлениям модернизации и технологического развития экономики России.

Получено свидетельство о государственной регистрации программы для ЭВМ № 20216111340 от 26.01.2021.

Основные положения, выносимые на защиту:

1. Разработана методика подготовки и разметки данных для задач извлечения терминов, отношений и связывания сущностей с элементами Wikidata. С помощью этой методики подготовлен корпус. Показана значимость данного корпуса для исследовательских целей. В частности, он может служить основой для обучения и оценки качества современных автоматических методов извлечения информации.
2. Предложен новый метод извлечения терминов из научных статей. Метод основан на частичном обучении и не зависит от области знаний и жанра текстов.
3. Адаптирован метод извлечения семантических отношений, основанный на технике "обучения без примеров" (zero-shot learning). Показано, что метод переноса обучения моделей с английского языка на русский хорошо работает для задачи классификации отношений.
4. Реализован алгоритм автоматического связывания научных терминов с сущностями в базе знаний Wikidata. Предложен ряд метрик для оценки качества метода, учитывающих различные аспекты. Описанные метрики показали сильные и слабые стороны реализованного алгоритма.

Достоверность результатов. Все полученные результаты подтверждаются экспериментами, проведенными в соответствии с общепринятыми стандартами.

Апробация результатов исследования. Основные результаты работы докладывались на следующих конференциях:

1. XXIII "Data analytics and management in data intensive domains" conference (DAMDID), Россия, Москва, 2021;
2. XXII Всероссийская конференция молодых учёных по математическому моделированию и информационным технологиям, Россия, Новосибирск, 2021;
3. Science and Artificial Intelligence conference (SAIC-2020), Россия, Новосибирск, 2020;
4. Международная научно-техническая конференция "Автоматизация" (RusAutoCon), Россия, Сочи, 2018;
5. 12-ая международная научно-практическая конференция «Виртуальные и интеллектуальные системы – ВИС-2017», Россия, Барнаул, 2017;
6. International Conference on Analysis of Images, Social Networks and Texts 2016 (AIST 2016), Россия, Екатеринбург, 2016.

Кроме того, результаты исследования обсуждались на ряде регулярных семинаров в Институте систем информатики им. А.П. Ершова СО РАН, Федеральном исследовательском центре информационных и вычислительных технологий, Новосибирском государственном университете.

Публикации. Основные результаты диссертации опубликованы в 10 научных статьях, из них: 3 в журналах из перечня ВАК РФ, 3 в изданиях, индексируемых Scopus; докладывались автором на 6 международных научных конференциях (Москва, Екатеринбург, Барнаул, Сочи, Новосибирск).

Получено 1 свидетельство о государственной регистрации программ для ЭВМ.

Основные результаты диссертации содержатся в работах [1 – 11].

Личный вклад соискателя. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Все представленные в диссертации результаты получены лично автором.

Объём и структура диссертационной работы. Диссертация состоит из введения, пяти глав, заключения и 8 приложений. Полный объем диссертации составляет 112 страниц, включая 7 рисунков и 22 таблицы. Список литературы содержит 105 наименований.

Основное содержание работы

Во ***введении*** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

В ***первой главе*** формулируются задачи извлечения сущностей, отношений между ними, а также связывания сущностей с внешней базой знаний. Приводится обзор существующих работ для каждой из этих задач. Целью данной главы является анализ достоинств и недостатков подходов к каждой из обозначенных задач.

Во ***второй главе*** проводится анализ существующих размеченных наборов данных для задачи извлечения сущностей, отношений между ними, а также связывания сущностей с внешней базой знаний. Описывается процедура разметки корпуса для поставленных задач: приводится подробная инструкция разметки, процесс, а также анализ полученного корпуса.

В рамках данной работы, вручную были размечены 80 аннотаций научных статей по теме “Информационные технологии”.

Всего в 80 размеченных текстах содержатся 11 157 токенов и 2 047 терминов. Средняя длина термина – 2.43 слова. Самый длинный термин состоит из 11 токенов. Процент согласия аннотаторов в задаче выделения сущностей составил 51.77%, что показывает высокую степень субъективности при нахождении слов и фраз, являющихся терминами, а также при определении точных границ сущностей.

Отношения между сущностями выделялись только в границах одного предложения, ограничения на количество отношений, в которые может вступать одна сущность, не накладывались. Всего в размеченной части корпуса было выделено 604 отношений между сущностями, из них CAUSE – 19, COMPARE – 9, ISA – 95, PARTOF – 90, SYNONYMS – 22, TOOL – 38, USAGE – 331. Больше половины составляют отношения использования (54.8%), на втором месте таксономические отношения (15.7%).

Для разметки данных для задачи связывания сущностей с внешней базой знаний использовалась информация также и о вложенных терминах. В качестве внешней базы знаний была выбрана база Викиданные. Всего в корпусе выделено 3386 терминов (с учётом вложенных сущностей), 1337 из которых удалось связать с сущностями в Викиданных. Средняя длина связанной сущности – 1,55 токен, минимальная длина – 1 токен, максимальная – 8 токенов.

В ***третьей главе*** дано формальное описание задачи извлечения научных терминов. Описаны алгоритмы, которые были реализованы в рамках данной работы: словарный подход, статистический подход, а также подходы, основанные на использовании алгоритмов глубокого обучения. Предложены метрики для оценки качества реализованных подходов, а также проведён анализ полученных результатов.

Назовем токеном x_i – слово или знак препинания в тексте. Рассмотрим последовательность всех токенов $X = \{x_0, x_1, ..., x_n\}$ и множество меток $Y = \{B\text{–}TERM,\ I\text{–}TERM,\ O\}$, где $B\text{–}TERM$ – метка для токена, который занимает первую позицию в термине, $I\text{–}TERM$ – метка для токена, который занимает вторую и последующие позиции в термине, O – метка для токена, который не входит в состав термина.

Требуется построить классификатор, который произвольной входной последовательности токенов ставит в соответствие последовательность меток, т.е. $\varphi: X \to Y$.

В качестве базового алгоритма был реализован метод на основе словаря. Его идея состоит в том, чтобы собрать конечный словарь фраз, которые являются терминами, а затем искать их во входном тексте. Как правило, метод такого типа обладает высокой точностью, но низкой полнотой, т.к. учесть разнообразие всех форм терминов, а также появление новых, невозможно. В рамках работы был собран словарь из 17252 терминов.

Также были проведены эксперименты с инструментом RAKE, основанном на статистическом подходе. Rapid automatic keyword extraction (RAKE) – алгоритм, предназначенный для автоматического извлечения ключевых слов [Rose S. et al., 2010]. Сначала применяется список стоп-слов и разделителей для выделения многословных терминов. После чего используется статистическая информация: для каждого слова из ключевых фраз-кандидатов оценивается частота, с которой оно встречается, и количество связей между этим словом и остальными. На основании этих двух величин вычисляется вес ключевой фразы, и все фразы сортируются по весам, наиболее вероятные ключевые фразы получают максимальный вес. Этот алгоритм хорошо применим к динамическим корпусам документов и к абсолютно новым доменам, при этом не зависит от языка и его особенностей.

Затем была проведена серия экспериментов с использованием методов машинного обучения. Сложность проведения экспериментов с использованием различных алгоритмов машинного обучения заключается в отсутствии размеченных данных. Эта проблема была решена следующим образом. Были взяты 1118 полных текстов научных статей (включая, аннотацию и основную часть), которые предварительно были очищены от формул, таблиц, схем и пр., и автоматически разметили тексты терминами из словаря, описанного в предыдущем разделе. Таким образом, у нас получился размеченный набор данных, общим объёмом 1992498 токенов и содержащий 177050 терминов.

Была поставлена гипотеза, что обобщающая способность модели позволит находить термины в аннотациях текстов, где, предположительно, концентрация терминов выше, в то время как, модель была обучена на полных текстах статей, в которых концентрация терминов ниже. Также, таким способом, будут находиться термины в текстах, которые отсутствовали в исходном словаре.

Были проведены эксперименты с посимвольной нейронной сетью, а также предложили итеративный метод на основе слабоконтролируемого обучения к извлечению терминов (Bert-LSTM и BertForTokenClassification). Идея предложенного подхода заключается в том, чтобы обучить модель на небольшом количестве размеченных данных, а затем разметить полученной моделью некоторое количество новых текстов, добавить их к обучающему множеству и обучить вторую модель.

Для более точного определения границ терминов, были реализованы несколько эвристик, которые учитывали части речи слов, входящих в состав термина, и ближайших к термину, а также некоторые другие грамматические характеристики.

Все подходы сравнивались друг с другом по основным метрикам информационного поиска – точность, полнота, F-мера. Для большей информативности учитывалось также, была ли найдена сущность полностью или только частично – из-за того, что определение границ термина является субъективной задачей, это разделение видится важным. Полученные значения для всех подходов представлены в Таблице 1.

Таблица 1. Полученные результаты для задачи извлечения терминов

Метод	Полное совпадение			Частичное совпадение		
	Точность	Полнота	F1	Точность	Полнота	F1
Словарный подход	0.25	0.17	0.20	**0.82**	0.34	0.48
RAKE	0.36	0.28	0.32	0.62	0.63	0.63
RAKE оптимизированный	**0.44**	**0.35**	**0.39**	0.65	0.57	0.61
Нейронная сеть	0.19	0.13	0.15	**0.82**	0.28	0.42
Bert-LSTM + эвристики + словарный подход	0.39	0.31	0.35	0.78	**0.78**	**0.77**
BertForTokenClassification + эвристики + словарный подход	0.40	0.31	0.35	0.77	0.77	**0.77**

Полученные результаты показали, что статистический подход с определёнными улучшениями даёт лучшие значения метрик при определении чётких границ терминов, в то время как модели, полученные на основе слабо контролируемого обучения, показывают значительно более высокие результаты, чем остальные методы, и являются достаточными для применения подхода для решения практических задач.

Также стоит отметить, что все эксперименты проводились на текстах из области информационных технологий, но реализованные алгоритмы могут быть применимы и расширены для других областей при наличии только неразмеченных текстов и начального словаря терминов.

В ***четвёртой главе*** дано формальное описание задачи извлечения отношений между научными терминами. Решена задача классификации отношений в постановке zero-shot learning. Реализованы алгоритмы для задачи извлечения отношений: с использованием лексико-синтаксических шаблонов, с использованием модели для классификации отношений, алгоритмы zero-shot learning с различными подходами к сэмплированию данных, а также ансамбль указанных решений.

Пусть дано предложение $S = \{x_0, x_1, \ldots, x_m\}\,(m \leq n)$, где x_i – токены. Для его элементов определена операция сцепления (конкатенации): $e_i = x_k x_{k+1} \ldots x_{k+l}\,(0 \leq i,\ k \leq m;\ l \geq 0),)$

Такое e_i назовём сущностью.

Рассмотрим пару сущностей (e_i, e_j) для $i \neq j$, и множество меток $Rel = \{CAUSE,\ COMPARE,\ ISA,\ PART\_OF,\ SYNONYMS,\ TOOL,\ USAGE,\ NONE\}$. Требуется построить классификатор, который паре (e_i, e_j) сопоставляет метку из Rel, т.е. $\gamma\colon (e_i, e_j) \to Rel$.

Все отношения, кроме SYNONYMS и NONE, являются *асимметричными*, т.е. для остальных отношений выполняется условие: $\forall\, R \in Rel \setminus \{SYNONYMS,\ NONE\}\ (e_i R e_j \Rightarrow \neg e_j R e_i)$.

Для извлечения информации о семантических отношениях была применена техника zero-shot learning, идея которой заключается в следующем. Сначала межъязыковую модель дообучают на данных того языка, которые представлены в достаточном количестве, а затем применяют эту модель к данным малоресурсного языка без дообучения. В качестве данных для обучения был использован размеченный корпус на английском языке SciERC.

Анализ результатов показал, что данный метод хорошо работает для задачи классификации отношений – для заданной пары сущности известно, что они связаны отношением, и нужно определить тип этого отношения. Модель, которая не видела примеров на русском языке, тем не менее, показывает хорошее качество классификации.

Задача извлечения отношений подразумевает ещё определение того, связана ли пара сущностей отношением или нет. В качестве базовых алгоритмов был реализован подход с вручную созданными лексико-синтаксическими шаблонами, а также использование модели, обученной для задачи классификации отношений, но отсутствие отношения между сущностями определялось по порогу.

Затем были проведены эксперименты не только с различными моделями, но также исследовали влияние сэмплирования на качество алгоритма. Очевидно, что примеров пар сущностей, которые не связаны отношениями, гораздо больше, чем тех, которые связаны – отсюда возникает дисбаланс классов, который влияет на работу моделей. Были поставлены эксперименты с двумя способами сэмплирования обучающих данных для сглаживания этого дисбаланса .

Объединение двух подходов: применение лексико-синтаксических шаблонов и использование модели в постановке zero-shot learning – позволило повысить качество данной задачи.

В целом, задача извлечения отношений видится сложной и нуждается в дальнейшем исследовании, что подтверждают полученные метрики (Таблица 2), а также анализ результатов работ других исследователей.

Таблица 2. Полученные результаты для задачи извлечения отношений

Подход	F1-micro	F1-macro
Для оценки качества используются все типы отношений		
Лексико-синтаксические шаблоны	0.88	0.23
Для оценки качества используются отношения из SciERC		
Использование модели классификации (порог=0.8)	0.65	0.21
Zero-shot learning	0.85	0.23
Ансамбль: лексико-синтаксические шаблоны и zero-shot learning	**0.86**	**0.27**

В ***пятой главе*** дано формальное описание задачи автоматического связывания сущностей с внешней базой знаний, а также приведен алгоритм, основанный на эвристическом и статистическом подходе.

Назовём $Entities$ множество сущностей и $Properties$ множество свойств. База знаний состоит из множества троек вида $< Subject\ Predicate\ Object >$, где $Subject$ и $Object$ являются элементами множества $Entities$, а $Predicate$ – элементом множества $Properties$.

Назовём токеном x_i – слово или знак препинания в тексте. Рассмотрим последовательность токенов $X = \{x_1, x_2, ..., x_n\}$. Сущностью Ent будет называться подпоследовательность таких токенов, которая представляет собой термин. Тогда мощность множества E, которое содержит в себе сущности Ent, всегда меньше либо равна мощности множества X, включая значение 0.

Задача автоматического связывания сущности состоит в построении такой функции F, которая бы для каждой сущности из множества E ставила бы в соответствие элемент из множества $Entities$ либо ε, где ε – отсутствие сущности в заданной базе знаний:

$$F\colon E \ -> \ Entities \cup \varepsilon.$$

В данной работе в качестве базы знаний используется база знаний Викиданные.

В качестве входных данных алгоритму подается последовательность или единичный токен, соответствующий термину. Далее выполняются два основных шага: создание массива кандидатов для связывания и нахождение наиболее подходящей сущности в полученном множестве кандидатов.

Все сущности – входные и в базе знаний – проходят лемматизацию с помощью MyStem. Это нужно для более точного поиска совпадений, т.к. русский язык отличается богатой морфологией и большим количеством словоформ.

На этапе создания массива кандидатов ищется построчное совпадение входной сущности с сущностями в базе знаний. Кроме того, для более полного формирования этого списка ищутся не только полное название входной сущности, а также униграммы, биграммы и триграммы, полученные из её названия.

Этап нахождения релевантной сущности из базы знаний рассматривался как задачу ранжирования. Чтобы учитывать не только название сущности, но и её контекст, использовалась дополнительная информация:

1. Для входного упоминания – название сущности, 5 токенов до неё и 5 токенов после неё (без учёта границ предложений);
2. Для сущности из Викиданных – название сущности, её синонимы и описание.

Каждая сущность (входную и из полученного множества кандидатов) была представлена в виде вектора V, который был получен по формуле:

$$V = \frac{\sum_{i=0}^{n} vector_i}{n}, \text{ где}$$

$vector_i$ – векторное представление для i-ого токена сущности,

n – количество токенов в сущности.

Векторные представления были получены с использованием предобученной модели Fasttext.

Затем полученные для каждого вектора сущности из базы знаний было рассчитано косинусное расстояние между ним и вектором входного упоминания. Кандидаты были отранжированы по этому расстоянию, далее кандидат, вектор которого наиболее близок к вектору входной сущности, считается связанной сущностью.

Для оценки качества алгоритма использовался ряд метрик.

1. ***Accuracy*** – определяется как отношение количества верно связанных терминов ко всем терминам. Так как нам удалось связать не все термины в корпусе, информативнее будет разделить эту метрику на две: ***Accuracy*** – принимает во внимание все сущности, и ***LinkedAccuracy*** – считается только на том наборе терминов, для которых нашлась сущность в графе знаний в корпусе. Таким образом, $Accuracy$ вычисляется по формуле:

$$Accuracy = \frac{CorrectEntities}{AllEntitties}, \text{ где}$$

$CorrectEntities$ – количество верно связанных терминов,

$AllEntities$ – количество всех терминов в корпусе.

Обозначим $AllLinkedEntities$ количество всех терминов в корпусе, которые имеют связь с сущностью в Викиданных. Тогда $LinkedAccuracy$ вычисляется по формуле:

$$LinkedAccuracy = \frac{CorrectLinkedEntities}{AllLinkedEntities}, \text{ где}$$

$CorrectLinkedEntities$ – количество верно связанных алгоритмом терминов среди всех связанных терминов.

2. ***Среднее количество кандидатов***. Эта метрика показывает, насколько хорошо работает этап генерации кандидатов: если значение относительно мало, то это означает, что можно улучшить алгоритм, например, также рассматривать синонимы, переводы, альтернативные написания сущностей и др. Если значение, наоборот, велико, то это может вызвать сложности при ранжировании кандидатов. Эта метрика также была разбита на две: $AveragedCandidates$ – среднее количество кандидатов для всех сущностей и $LinkedAveragedCandidates$ – среднее количество кандидатов для набора терминов, которые удалось связать.

$$AveragedCandidates = \frac{\sum_{1}^{n} |Candidates_i|}{AllEntities}, \text{ где}$$

$Candidates_i$ – множество полученных кандидатов для сущности.

Обозначим $LinkedCandidates_i$ множество сгенерированных кандидатов для всех терминов, связанных с Викиданными. Тогда формула для метрики $LinkedAveragedCandidates$ имеет вид:

$$LinkedAveragedCandidates = \frac{\sum_{1}^{n} |Linked\_candidates_i|}{AllLinkedEntities}.$$

3. ***Наличие подходящего кандидата в списке, найденном алгоритмом***. Данная метрика считалась только для множества терминов в корпусе, которые имеют связь с сущностью из графа знаний, и вычислялась по формуле:

$$TopCandidates = \frac{CorrectSets}{AllLinkedEntities}, \text{ где}$$

$CorrectSets$ – это количество сгенерированных списков кандидатов, содержащих верную сущность.

Полученные значения метрик представлены в Таблице 3.

В ***заключении*** сделаны выводы, подведены итоги проведенного исследования, а также изложены рекомендации и перспективы дальнейшей разработки темы.

Таблица 3. Полученные результаты для задачи связывания сущности

Accuracy	LinkedAccuracy	Averaged Candidates	LinkedAveraged Candidates	TopCandidates
0.38	0.23	10.29	7.38	0.76

Основные результаты

1. Собран и размечен корпус научных текстов для задач извлечения научных терминов, извлечения отношений и связывания сущностей с внешней базой знаний.
2. Исследованы различные методы извлечения терминов из научных текстов: словарный метод, статистический метод, с использованием машинного обучения.
3. Предложен подход для извлечения терминов на основе слабоконтролируемого обучения, идея которого заключается в обучении модели на большом количестве данных с автоматической разметкой.
4. Адаптирован метод извлечения отношений между терминами, основанный на переносе обучения моделей с английского языка на русский в постановке zero-shot learning.
5. Описан алгоритм и реализован метод связывания терминов с сущностями в базе знаний. Предложен ряд метрик для оценки качества метода, учитывающий различные аспекты.
6. Разработан программный комплекс для извлечения информации из научных текстов.

Публикации автора по теме диссертации

Статьи в журналах из перечня ВАК:

1. Бручес Е. П., Батура Т. В. Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения. Вестник НГУ. Серия: Информационные технологии. 2021 Т.19, №2. С. 5–16. DOI: 10.25205/1818-7900-2021-19-2-5-16
2. Батура Т.В., Бручес Е.П., Паульс А.Е., Исаченко В.В., Щербатов Д.Р. Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей. Программные продукты и системы. 2021. Т. 34. № 1. С. 132–144. DOI: 10.15827/0236-235X.133.132-144
3. Мезенцева А. А., Бручес Е. П., Батура Т. В. Автоматическое связывание терминов из научных текстов с сущностями базы знаний. Вестник НГУ. Серия: Информационные технологии. Т.19, №2. с. 65–75. 2021. DOI: 10.25205/1818-7900-2021-19-2-65-75

Статьи в изданиях, индексируемых в Scopus/Web of Science:

4. Bruches E.P., Pauls A.E., Batura T.V., Isachenko V.V. Study of Methods for Entity Recognition and Relation Extraction in Scientific Texts. Science and Artificial Intelligence conference (SAIC-2020). 2020. p. 41–45. DOI: 10.1109/S.A.I.ence50533.2020.9303196
5. Batura T.V., Bruches E.P. A combined approach to the problem of part-of-speech homonymy resolution in Russian texts. Proceedings of the International Russian Automation Conference (RusAutoCon 2018). September 9-16, 2018. pp. 4–9. DOI 10.1109/RUSAUTOCON.2018.8501718
6. Bruches E., Karpenko D., Krayvanova V. The Hybrid Approach to Part-of-Speech Disambiguation. International Conference on Analysis of Images, Social Networks and Texts (AIST 2016). 2016. pp. 21–26.

Статьи в изданиях, индексируемых в РИНЦ:

7. Крайванова В.А., Бручес Е.П., Минаков А.М., Анкудинов К.Л., Пчельников Д.В. Архитектура категоризатора событий в гетерогенном пространстве параметров // Ползуновский альманах, № 4, 2018, с.134–138.
8. Бручес Е.П., Крайванова В.А. О способе векторизации морфологической информации словоформы. Сборник научных трудов «Нечеткие системы и

мягкие вычисления. Промышленные применения», г. Ульяновск, 2017. с. 232–239.

9. Бручес Е. П., Крайванова В. А. Снятие омонимии геолокаций на основе частоты встречаемости контекстов. Ползуновский альманах № 4, 2017, т. 3, с. 103–105.
10. Batura T.V., Bruches E.P., Strekalova S.E. A combined approach to part-of-speech homonymy resolution. Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2017. Is. 41. pp. 13–25.

Свидетельства о регистрации программ для ЭВМ:

11. Бручес Е.П., Батура Т.В. Свидетельство о регистрации программ для ЭВМ №2021611340 «Система автоматического извлечения терминов из научных текстов «Term Extractor». Дата регистрации: 26.01.2021.