Московский государственный университет им. М.В. Ломоносова
Филологический факультет

На правах рукописи

Большина Ангелина Сергеевна

Методы разрешения лексической неоднозначности на основе автоматически размеченных семантических корпусов

Специальность 10.02.21 – «Прикладная и математическая лингвистика»

Автореферат

диссертации на соискание учёной степени
кандидата филологических наук

Научный руководитель:
д.т.н. Лукашевич Наталья Валентиновна

Москва – 2022

Работа выполнена на кафедре теоретической и прикладной лингвистики филологического факультета Московского государственного университета имени М. В. Ломоносова.

Научный руководитель: **Лукашевич Наталья Валентиновна**
доктор технических наук, профессор кафедры теоретической и прикладной лингвистики филологического факультета МГУ им. М. В. Ломоносова, ведущий научный сотрудник Научно-исследовательского вычислительного центра (НИВЦ) МГУ имени М. В. Ломоносова

Официальные оппоненты: **Зацман Игорь Моисеевич**
доктор технических наук, заведующий отделом Федерального исследовательского центра «Информатика и управление» Российской Академии Наук (ФИЦ ИУ РАН)

Ягунова Елена Викторовна
доктор филологических наук, профессор кафедры информационных систем в искусстве и гуманитарных науках Федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет»

Ляшевская Ольга Николаевна
кандидат филологических наук, профессор Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики»

Защита диссертации состоится «22» июня 2022 года в 16 часов 00 минут на заседании Диссертационного совета МГУ.10.04 Московского государственного университета имени М. В. Ломоносова по адресу: 119991, ГСП-1, г. Москва, Ленинские горы, МГУ, д.1, стр. 51, 1-й учебный корпус гуманитарных факультетов, филологический факультет.

E-mail: otipl@philol.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М. В. Ломоносова (Ломоносовский просп., д. 27).

С информацией о регистрации участия в защите и с диссертацией в электронном виде можно ознакомиться на сайте ИАС «ИСТИНА»: https://istina.msu.ru/dissertations/457125986/

Автореферат диссертации разослан «___» _____ 2022 года

Ученый секретарь
диссертационного совета,
кандидат филологических наук О. И. Беляев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Неоднозначность — это неотъемлемое свойство естественных языков. Автоматическое разрешение лексической неоднозначности (Word Sense Disambiguation, WSD) является одним из этапов семантического анализа текстов, который используется в машинном переводе, извлечении информации, классификации текстов. Задача разрешения неоднозначности состоит в выборе корректного значения многозначного слова в определенном контексте.

Наилучшие результаты снятия многозначности на различных наборах данных достигаются методом машинного обучения с учителем. Однако подобная парадигма обучения требует большого количества размеченных обучающих данных, которые доступны лишь для небольшого числа языков. Получение семантически аннотированных данных дорогостоящий процесс, требующий немало времени и трудозатрат. В связи с этим достижения в области автоматического разрешения неоднозначности не могут быть применены для решения этой задачи в языках с недостаточным количеством размеченных лингвистических ресурсов, к которым относится и русский язык.

Реферируемая диссертация посвящена задаче автоматической генерации семантически размеченных обучающих коллекций. В работе уделяется особое внимание методу формирования корпусов с помощью однозначных родственных слов, который рассматривается на материале русского языка.

Объектом исследования в данной работе являются корпуса с семантической разметкой.

Предметом реферируемой диссертации являются автоматически порожденные корпуса с семантический разметкой по значениям слов.

Актуальность темы исследования обусловлена тем, что существует необходимость разрешения лексической неоднозначности в условиях недостатка или отсутствия размеченных данных. Для языков, в которых наблюдается недостаток размеченных данных (к ним относится и русский язык), требуется разрабатывать методы автоматической генерации размеченных обучающих

коллекций с учетом имеющихся в языке источников лексической информации (например, тезаурусы, словари, параллельные корпуса и т.п.).

Степень разработанности проблемы. Ввиду того, что корпусов с разметкой значений слов существует не так много, создается большое количество методов по автоматическому сбору и разметке обучающих коллекций. Исследователи используют различные подходы, среди которых выделяются системы, базирующиеся на однозначных родственных словах, параллельных корпусах, базах знаний, алгоритме распространения меток и бутстрэппинге. Все ресурсы, используемые в данных методах, отличаются по сфере применимости и охвату значений многозначных слов. Подход, основанный на методе однозначных родственных слов, представляется наиболее доступным из всех представленных в литературе систем, т.к. основной ресурс, на который он опирается – это семантическая сеть. Данный источник знаний можно найти для большого числа языков. Например, один из наиболее распространённых семантических графов BabelNet[1] по состоянию на февраль 2021 года содержит в себе информацию для 500 языков. Помимо этого, данный подход позволяет набирать такое количество обучающих примеров для каждого целевого значения, какое требуется, размер обучающей выборки ограничен только размером корпуса, из которого извлекаются контексты.

Целью реферируемого диссертационного исследования является разработка метода автоматического сбора и разметки корпуса русского языка для автоматического разрешения многозначности, а также его программная реализация. В рамках данного исследования рассматривается метод, основанный на однозначных родственных словах, на материале русского языка.

Для достижения цели настоящей работы были поставлены следующие **задачи**:

[1] https://babelnet.org/

1) Проанализировать теоретические аспекты создания систем автоматической генерации размеченных обучающих коллекций для задачи разрешения лексической неоднозначности.
2) Реализовать алгоритм автоматической разметки текстовых коллекций, используя информацию об однозначных словах из лексико-семантического ресурса для русского языка RuWordNet [Loukachevitch et al., 2016].
3) Разработать метод фильтрации примеров, автоматически размеченных с помощью однозначных родственных слов, с целью обеспечения разнообразия контекстов и их семантической близости к целевым значениям.
4) Провести оценку корректности семантической разметки корпуса, аннотированного с помощью информации об однозначных родственных словах.
5) Обучить модель разрешения лексической многозначности для русского языка с применением полученного размеченного обучающего множества.

Научная новизна настоящего исследования заключается в следующем:

1) Предложен подход к автоматическому созданию и разметке корпуса для разрешения лексической многозначности на основе однозначных родственных слов, который учитывает далеко расположенных однозначных кандидатов. Это дает описанному методу возможность найти обучающие примеры для подавляющего большинства многозначных слов и их значений из тезауруса.
2) Реализован метод фильтрации однозначных родственных слов на основе близости их векторных представлений со словами семантически близкими целевому значению. Этот компонент повышает релевантность примеров, добавляемых в обучающую коллекцию, и, как следствие, уменьшает «шум» в данных.
3) Разработаны и обучены модели автоматического разрешения лексической многозначности для русского языка на материале автоматически собранных

корпусов. Данные модели хорошо обучаются на неточных метках значений и показывают качество, сравнимое с результатами моделей, обученных на вручную размеченных данных.

Теоретическая значимость исследования состоит в дальнейшей разработке метода генерации обучающих коллекции на основе однозначных родственных слов для русского языка. Кроме того, в работе представлено выведение и обоснование компонента фильтрации однозначных кандидатов, который позволяет отбирать более репрезентативные контексты для обучающей выборки. Таким образом, теоретическая значимость исследования также состоит в развитии представлений об источниках семантически близких контекстов для заданного значения слова. Сформулированные в диссертационном исследовании выводы демонстрируют особенности и проблемные места систем автоматического разрешения лексической неоднозначности на русском языке.

Практическая значимость диссертационной работы определяется возможностью применения разработанного подхода, основанного на методе однозначных родственных слов, к автоматической генерации и разметке обучающих коллекций для задачи разрешения лексической неоднозначности, а также для других задач, где требуется семантическая разметка текстов. Помимо этого, предложенный метод можно использовать для дополнения уже имеющихся аннотированных текстовых данных, и, как следствие, повышения точности моделей обработки естественного языка. Полученные экспериментальные данные могут способствовать развитию подходов для автоматической генерации и аннотации текстовых коллекций.

Экспериментальным материалом реферируемого диссертационного исследования послужили новостной корпус, сегменты корпуса «Тайга», относящиеся к новостным ресурсам и художественной литературе, наборы данных с технологического соревнования RUSSE-2018, новости с ресурса Wikinews, лексико-семантический ресурс для русского языка RuWordNet. Анализ текстовых данных, программная реализация моделей разрешения неоднозначности и

алгоритма сбора и разметки текстов, базирующегося на однозначных родственных словах, были осуществлены с помощью языка программирования Python.

Теоретико-методологическую основу исследования составили работы по теоретической семантике Апресяна Ю. Д., Кобозевой И. М., Падучевой Е. В., Зализняк А. А., Кустовой Г. И., основным компонентам систем разрешения неоднозначности Navigli R., Bevilacqua M., Pasini T., Raganato A. и др., работы, описывающие процедуры создания, обучения и оценки моделей разрешения неоднозначности, Ляшевской О. Н., Митрофановой О. А., Loureiro D., Pilehvar M. T., Camacho-Collados J., Scarlini B., Berend G. и др., работы, посвященные методам автоматического сбора и разметки обучающих коллекций, Mihalcea R., Martinez D., Agirre E., Delli Bovi C., Taghipour K., Ng H. T., Otegi A. и др.

На защиту выносятся следующие **положения**:

1) **Метод автоматической генерации и разметки** семантически аннотированных коллекций, базирующийся на однозначных родственных словах, извлекаемых из тезауруса для русского языка RuWordNet.
2) **Компонент фильтрации** однозначных родственных слов и контекстов, в которых они употребляются.
3) **Готовые к применению модели** разрешения неоднозначности, обученные на текстовых коллекциях, собранных с помощью метода однозначных родственных слов.
4) **Список наиболее успешных стратегий** обработки текстовых данных и извлечения контекстуализированных векторных представлений слов, а также эффективных архитектур моделей, с помощью которых достигаются максимальные показатели качества предсказания значений слов в русском языке.

Достоверность результатов настоящей диссертационной работы обеспечивается методологической базой исследования, успешным практическим применением разработанного подхода на основе однозначных родственных слов

для генерации семантически размеченных обучающих коллекций, а также открытым кодом реализованных методов и моделей.

Личный вклад соискателя заключается в проведении основного объема теоретических и экспериментальных исследований, а также в разработке и программной реализации метода автоматического сбора и разметки обучающих коллекций и моделей разрешения неоднозначности. Подготовка части материалов к публикации проводилась совместно с научным руководителем, причем вклад диссертанта был определяющим.

Апробация работы. Основные положения и результаты работы докладывались на научных конференциях: 26-я международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-2020», международная конференция по компьютерной лингвистике в Болгарии (CLIB 2020), 18-я национальная конференция по искусственному интеллекту (КИИ-2020), международная конференция «Лингвистический форум 2020: Язык и искусственный интеллект», открытая конференция ИСП РАН им. В.П. Иванникова 2021, XII Международная научная конференция «Интеллектуальные системы и компьютерные науки».

Результаты исследования опубликованы в 9 статьях, 6 из которых в изданиях, рекомендованных для защиты в диссертационном совете МГУ им. М.В. Ломоносова.

Структура диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы и двух приложений. Общий объем диссертационной работы составляет 163 страницы. Список литературы содержит 257 наименований. Код, реализующий разработанный в рамках диссертационного исследования метод однозначных родственных слов и модели разрешения неоднозначности, а также данные, необходимые для их запуска, доступны по ссылке https://github.com/loenmac/russian_wsd_data.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во **введении** характеризуется актуальность, научная и практическая значимость, цель работы, формулируются задачи исследования, а также приводятся основные положения, выносимые на защиту.

В **главе 1** рассматриваются теоретические аспекты полисемии и задачи разрешения лексической неоднозначности, представлен обзорный анализ подходов к автоматизации построения семантически размеченных коллекций и история развития систем автоматического разрешения лексической неоднозначности. В **первом разделе первой главы** описываются принципы представления значений слов в системах снятия лексической неоднозначности. Во **втором разделе первой главы** рассказывается об основных ресурсах, которые используются в системах разрешения неоднозначности, например, наборы данных с разметкой значений, семантические сети, текстовые корпуса и т.п. В **третьем разделе первой главы** рассматриваются способы представления контекста употребления многозначного слова для систем разрешения неоднозначности. В **четвертом разделе первой главы** описаны метрики, с помощью которых автоматически оцениваются и сравниваются различные модели разрешения неоднозначности.

В пятом разделе первой главы описаны основные направления в области автоматических методов генерации семантически размеченных обучающих коллекций.

Подраздел 1.5.1. посвящен подходу, базирующемся на однозначных родственных словах, который позволяет собирать и размечать текстовые коллекции на основании слов, семантически связанных с ключевым значением многозначного слова в каком-либо семантическом графе и при этом имеющих только одно значение. Среди исследователей, разрабатывающих методы данной группы, можно выделить Martinez D., Mihalcea R.

Методы генерации обучающих данных с помощью параллельных корпусов обсуждаются в **подразделе 1.5.2.** Здесь стоит отметить исследования Taghipour K., Ng H. T., Otegi A., Hauer B.

Подраздел 1.5.3. описывает методы, использующие различные базы знаний, прежде всего семантические сети (например, WordNet), Википедию и Викисловари. Такие исследователи, как Pasini T., Navigli R., Scarlini B., посвящали свои работы подходам этой группы.

В отдельный **подраздел 1.5.4.** вынесены подходы, которые для автоматической (и полуавтоматической) семантической разметки используют алгоритм распространения меток, бутстрэппинг и активное обучение. Среди исследователей, изучающих подходы такого типа, можно выделить Khapra M. M., Barba E.

Ввиду того, что имеющиеся размеченные ресурсы включают в себя не все многозначные слова из инвентаря значений и покрывают не все возможные значения многозначных слов, разрабатываются различные методы для преодоления этих недостатков. Подобные системы описаны в **подразделе 1.5.5.**

Подраздел 1.5.6. реферируемой диссертации представляет собой сравнительный анализ качества предсказаний моделей, обученных на автоматически сгенерированных коллекциях и наборах данных, размеченных вручную. Полученные в ходе различных экспериментов метрики свидетельствуют о том, что системы, в которых использовались синтетические данные, не уступают системам, в которых для обучения применялись данные, размеченные экспертами.

Во **шестом разделе первой главы** приведен обзор систем разрешения лексической неоднозначности, сгруппированных по основному методу классификации, который в них используется.

В **подразделе 1.6.1.** описываются подходы к снятию лексической неоднозначности, основанные на знаниях (*knowledge-based methods*). Для работы такие методы требуют только базу знаний (например, семантический граф типа WordNet или толковый словарь), однако такие системы уступают моделям, которые базируются на обучении с учителем, в плане качества предсказаний. В данной области можно особо выделить работы Lesk M., Moro A., Agirre E., Basile P.

Подраздел 1.6.2. посвящен алгоритмам разрешения лексической неоднозначности, использующим методы машинного обучения с учителем

(*supervised word sense disambiguation*). Здесь описываются как классические подходы (например, наивный байесовский классификатор и метод опорных векторов), так и различные архитектуры моделей на основе нейронных сетей (например, LSTM и трансформеры). Для подобной группы методов необходимы обучающие коллекции со снятой многозначностью, а в случае, если в качестве классификатора используется нейронная сеть, то размер обучающего набора данных должен быть очень большим. При этом результаты последних технологических соревнований показывают, что наилучшее качество разрешения неоднозначности на различных наборах данных достигается именно методами машинного обучения, которые основаны на обучении с учителем. Среди исследователей, занимающихся изучением методов данной группы, можно отметить Zhong Z., Ng H. T., Bevilacqua M., Navigli R., Iacobacci I., Raganato A.

В **подраздел 1.6.3.** вынесены алгоритмы снятия неоднозначности, которые применяют в работе методы машинного обучения без учителя (*word sense induction*). Несомненным преимуществом таких моделей является то, что для своей работы они не требуют размеченный корпус, однако качество их работы трудно количественно оценивать. В данной группе методов не используется никакой инвентарь значений, что осложняет соотнесение смысловых кластеров, которые были выведены моделью, с тем или иным набором значений слов. В этой области можно выделить работы исследователей Amrami A., Goldberg Y., Amplayo R. K.

Седьмой раздел первой главы реферируемой диссертационной работы посвящен исследованиям систем разрешения лексической неоднозначности и автоматического извлечения значений для русского языка. Среди исследований в области снятия неоднозначности стоит отметить следующих авторов: Ляшевская О. Н., Митрофанова О. А., Кобрицов Б. П., Толдова С. Ю. Среди исследователей, занимающихся задачей автоматического извлечения значений, можно выделить следующих: Лопухин К. А., Лопухина А. А., Арефьев Н. В., Кутузов А. Б.

Вторая глава посвящена методике сбора и разметки обучающих коллекций с помощью однозначных родственных слов, разработанной в рамках диссертационного исследования.

В **первом разделе второй главы** описывается система отбора и ранжирования однозначных кандидатов. В качестве базы знаний и источника таких «родственников» чаще всего используется лексико-семантические сети типа WordNet. Базовым понятием таких сетей является синсет, представляющий собой синонимический ряд, в который входят слова со схожими значениями. Синсеты формируют узлы семантического графа и соединяются друг с другом ребрами, представляющими такие отношения, как гипонимия, гиперонимия, меронимия и т.п.

Однозначные родственные слова – это слова или словосочетания, связанные с целевым многозначным словом каким-либо отношением в графе WordNet и имеющие только одно значение, т.е. принадлежащие одному синсету. Подходы, использующие информацию об однозначных родственных словах, основаны на заменах. Сначала однозначные «родственники» отбираются с помощью того или иного метода, затем из корпуса извлекаются контексты, в которых они встречаются. В этих текстах они заменяются на целевые многозначные слова, а тексты добавляются в обучающую коллекцию. Целевые многозначные слова – это слова, с которых требуется снять неоднозначность.

В разработанной в рамках диссертационного исследования системе учитываются не только ближайшие к значению целевого слова однозначные родственные слова (в семантическом графе), как это делалось во многих предыдущих работах, но и более удаленные. Расстояние между значениями подсчитывалось как количество ребер в семантической сети, соединяющих два узла, т.е. значения слов. Для определения списка тех однозначных слов, которые рассматриваются алгоритмом при построении обучающей коллекции, используется термин ***однозначные кандидаты («родственники-кандидаты»)***, обозначающий однозначные слова или словосочетания, которые могут быть расположены в пределах четырех шагов от многозначного слова. Принимаются во внимание только те слова или словосочетания, которые встречаются в корпусе хотя бы 50 раз.

Стоит отметить, что не все контексты, в которых употребляются однозначные родственные слова, хорошо отражают значение ключевого многозначного слова. Помимо этого, некоторые слова, описанные в тезаурусе как однозначные, в корпусе могут иметь несколько значений. Ввиду этих факторов в рамках реферируемого исследования был разработан компонент фильтрации на основе коэффициента схожести однозначного «родственника-кандидата» и синсетов, близких к целевому значению многозначного слова, которые обозначаются понятием ***гнездо синсета***. Группа синонимов для ключевого значения, а также все слова из непосредственно связанных синсетов в пределах двух шагов от целевого слова составляют ***гнездо синсета*** для целевого значения.

Ниже приведен фрагмент гнезда синсета для слова *такса* ‘порода собак’:

(1) *охотничий пёс, охотничья собака, пёсик, четвероногий друг, псина, собака, терьер, собачонка, борзая собака*… и т.д.

Модели векторных представлений слов используются в компоненте фильтрации для выбора наиболее подходящих однозначных «родственников», чей контекст может служить хорошей репрезентацией значения целевого слова. В данной работе для анализа семантической близости контекстов с однозначными кандидатами и целевого значения применяются модели векторных представлений слов word2vec, основанные на архитектуре нейронных сетей CBOW.

Предложенная в диссертационном исследовании система ранжирования однозначных родственных слов позволяет приписывать каждому однозначному кандидату вес, который помогает оценить потенциал контекстов употребления этого кандидата для отображения значения целевого многозначного слова. Предполагается, что контексты с однозначными родственными словами, которые расположены выше в рейтинге, являются более подходящими примерами для репрезентации значения целевого слова. В реферируемом диссертационном исследовании система отбора и ранжирования однозначных родственных слов состоит из следующих шагов:

1. Извлекаются все однозначные кандидаты в пределах 4 шагов от целевого многозначного значения s_j.

2. Формируется гнездо синсета ns_j, которое состоит из слов близких к целевому значению s_j, например, синонимов, гипонимов, гиперонимов и когипонимов. Гнездо синсета ns_j состоит из N_k синсетов.

3. Для каждого однозначного кандидата r_j с помощью word2vec модели, обученной на том или ином корпусе, извлекается 100 наиболее похожих слов.

4. Необходимо найти пересечение этого списка слов со словами, которые входят в гнездо синсета ns_j для целевого значения s_j.

5. Для каждого слова в пересечении берется его косинусная мера близости, вычисленная с помощью модели word2vec. Этот вес приписывается синсету, которому принадлежит слово. Результирующий вес синсета в гнезде ns_j определяется максимальным весом среди всех слов $w^j_{k_1}, \dots, w^j_{k_i}$, которые представляют этот синсет в пересечении.

6. Финальный вес однозначного кандидата r_j вычисляется как сумма весов всех синсетов из гнезда синсета ns_j. Благодаря такому подсчету больший вес получат те кандидаты, которые схожи с большим числом синсетов, входящих в гнездо синсета для целевого значения многозначного слова. Таким образом, общий вес кандидата определяется следующей формулой:

$$Weight_{r_j} = \sum_{k=1}^{N_k} max\left[cos\left(r_j, w^j_{k_1}\right), \dots, cos\left(r_j, w^j_{k_i}\right)\right] \quad (1)$$

В примерах 2 и 3 приведены фрагменты списка однозначных родственных слов с соответствующими значениями близости (приведены в скобках), которые были получены для существительного *гвоздика*:

(2) *гвоздика* ‘приправа’: *чёрный перец* (7.5), *кардамон* (6.8), *корица* (6.5), *имбирь* (6.4), *мускатный орех* (6) … и т.д.

(3) *гвоздика* ‘растение’: *фиалка* (14.0), *орхидея* (13.8), *тюльпан* (13.6), *астра* (12.8), *ландыш* (12.6) ….

Во **втором разделе второй главы** содержится обзор данных, использованных в реферируемом диссертационном исследовании. В

разработанном методе генерации и разметки обучающих коллекций для извлечения однозначных кандидатов и для вычисления их характеристик, применялся тезаурус для русского языка RuWordNet. В этом разделе приведены количественные характеристики семантической сети RuWordNet.

Помимо этого, в данной главе приведена информация о составе и объеме следующих корпусов, являющихся экспериментальным материалом настоящего исследования: новостной корпус, сегменты корпуса «Тайга», относящиеся к новостным ресурсам и художественной литературе. Новостной корпус и обозначенные выше сегменты корпуса «Тайга» использовались для извлечения контекстов с однозначными родственными словами, а также для обучения моделей векторных представлений слов (*эмбеддингов*) word2vec, основанных на архитектуре нейронных сетей CBOW.

В **третьем разделе второй главы** на конкретных примерах из исходных данных демонстрируются основные характеристики метода однозначных родственных слов, и описывается процедура подготовки обучающей коллекции.

Результаты применения описанного метода к материалу RuWordNet показывают, что с помощью него можно найти однозначных «родственников» почти для всех многозначных слов в тезаурусе и, таким образом, создать обучающую коллекцию для моделей автоматического разрешения неоднозначности. Анализ полученных данных также показывает, что большинство однозначных родственников расположены на более удаленных расстояниях от ключевого слова. Эти результаты свидетельствуют о том, что стратегия выбора однозначных родственных слов в рамках широкого диапазона расстояний от целевого значения многозначного слова позволяет создавать обучающие коллекции с большим покрытием значений и неоднозначных слов. Процентное соотношение однозначных «родственников», соединенных различными отношениями с целевым значением многозначного слова и расположенных на различных расстояниях от него, показывает, что больше всего однозначных родственных слов находится на расстоянии 2-3 шагов от целевого синсета и являются его когипонимами.

Для сравнения в рамках реферируемого диссертационного исследования были созданы две отдельные обучающие коллекции из текстов новостного корпуса и сегмента корпуса «Тайга», относящегося к художественной литературе (Проза.ру), соответственно. Все однозначные кандидаты сортируются в зависимости от веса, который был приписан им в ходе работы системы отбора. На основе результатов ранжирования однозначных «родственников» в реферируемом исследовании были предложены два различных способа сбора обучающих коллекций. В соответствии с первым методом сбора обучающей коллекции была собрана коллекция только с тем однозначным родственным словом, которое получило наибольший вес во время процедуры ранжирования. Для удобства эта коллекция обозначалась как Корпус-1000, потому что в ней для каждого значения многозначного слова было получено по 1000 примеров. Во втором подходе примеры для обучающей коллекции собирались с помощью всех имеющихся однозначных родственных слов с ненулевым весом. Количество примеров для каждого однозначного «родственника» определялось его весом, полученным на этапе ранжирования. Эта коллекция обозначалась как сбалансированная, так как выбор обучающих примеров не был ограничен только одним однозначным родственным словом.

В **третьей главе** реферируемого диссертационного исследования описываются методики и процедуры проведения различных экспериментов, произведенных на текстовых коллекциях, размеченных и сгенерированных предложенным методом с помощью однозначных родственных слов.

В **разделе 3.1.** описывается серия из трех экспериментов, которая была направлена на оценку моделей разрешения неоднозначности, обученных на автоматически собранных и аннотированных коллекциях.

Помимо обучающих коллекций, описанных в предыдущей главе, в данном исследовании использовались наборы данных с технологического соревнования RUSSE-2018. Инвентарь значений слов из этих текстовых коллекций был сопоставлен с инвентарем из тезауруса RuWordNet, что привело к исключению некоторых многозначных слов и их значений из выборки. Полученный тестовый

набор данных обозначался в реферируемом диссертационном исследовании как RUSSE-RuWordNet, потому что он является пересечением инвентаря значений RUSSE-2018 и значений слов, описанных в RuWordNet. Помимо этого, был собран еще один тестовый набор данных, состоящий из новостных статей из газеты «Комсомольская правда», которая является одним из сегментов корпуса «Тайга». Вручную в рамках диссертационного исследования в этих текстах было размечено 385 предложений с 27 целевыми неоднозначными словами, входящими в выборку RUSSE-RuWordNet. Набор данных со словарными дефинициями и примерами употребления слов использовался в данной серии экспериментов для обучения базовой модели (бэйслайн, *baseline*) задачи разрешения неоднозначности, а также расширения обучающих выборок.

Тестирование моделей осуществлялось на наборе данных RUSSE-RuWordNet и наборе данных из «Комсомольской правды». Анализ проводился на легко интерпретируемой модели – методе ближайших соседей (kNN), объектами классификации которого выступали контекстуализированные представления слов, извлеченные из языковых моделей ELMo и BERT.

В первом эксперименте сравнивалось качество предсказаний моделей, обученных на текстовых коллекциях разного жанра (новостные статьи и художественная литература). Помимо этого, при извлечении обучающих данных из этих ресурсов применялись две различные стратегии формирования выборки на основе рейтинга однозначных родственных слов, которые были описаны выше. Кроме того, сравнивалось качество снятия неоднозначности, выполненное с помощью моделей, обученных на лемматизированных и нелемматизированных текстах.

Ещё одним конфигурируемым параметром при тестировании выступал тип языковой модели, которая применялась для представления текстов с помощью сжатых контекстуализированных векторов. В экспериментах использовались две различные модели ELMo – одна обученная коллективом DeepPavlov на русской части корпуса WMT News, а другая модель RusVectōrēs, обученная на лемматизированном корпусе «Тайга». Также применялись две модели BERT:

BERT-base-multilingual-cased от Google Research и RuBERT, который был обучен DeepPavlov на русскоязычной части Википедии и новостных материалах. В реферируемом диссертационном исследовании сравнивались два способа извлечения контекстуализированных представлений из моделей ELMo – вычисление вектора для всего предложения, содержащего ключевое слово, или отдельного вектора для целевого неоднозначного слова. Для извлечения контекстуализированных представлений из предварительно обученной модели BERT использовался следующий метод: конкатенировались векторные представления слов из четырех последних слоев трансформера.

Таблицы 1, 2 и 3 демонстрируют результаты, полученные с помощью разных обучающих коллекций, контекстуализированных представлений слов и обучающих параметров. В качестве минимально работающего прототипа модели разрешения неоднозначности использовалась модель, обученная на словарных дефинициях и примерах употреблений слов. Данные, приведенные ниже, демонстрируют, что все системы смогли показать результаты лучше, чем базовое решение. Алгоритм, основанный на контекстуализированных представлениях слов ELMo от RusVectōrēs, превзошел по качеству все остальные модели, его F1-мера составила 0.857. Модель RuBERT от DeepPavlov показала второй результат, следом за ней идет модель ELMo от DeepPavlov. Самая низкая F1-мера была получена на представлениях слов из языковой модели Multilingual BERT.

Модели, обученные на Проза.ру, показывают более высокие результаты, чем модели, использовавшие материал новостного корпуса. Это можно объяснить тем, что жанр коллекции Проза.ру совпадает с жанром тестового набора данных. Чтобы проверить гипотезу о том, что качество разрешения неоднозначности выше для тех моделей, чей обучающий корпус по жанру совпадает с тестовым, был проведен эксперимент с помощью тестовой выборки «Комсомольская правда». Результаты показали, что на этом наборе данных лучше работает модель, обученная на новостном корпусе (**0.78 F1**), чем та, что была обучена на Проза.ру (**0.74 F1**). Эти метрики подтверждают высказанное ранее предположение.

Таблица 1. Значения метрики F1 для моделей, основанных на представлениях слов из языковой модели BERT.

Модель	**RuBERT DeepPavlov (коллекция Корпус-1000)**		**Multilingual BERT (коллекция Корпус-1000)**		**RuBERT DeepPavlov (сбалансир. коллекция)**		**Multilingual BERT (сбалансир. коллекция)**	
k (кол-во соседей)	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус
5	0.793	0.771	0.694	0.667	0.792	0.769	0.717	0.682
7	**0.804**	**0.774**	0.699	0.673	0.802	0.768	0.723	0.683
9	0.802	0.769	**0.7**	**0.677**	**0.812**	**0.774**	**0.729**	**0.688**
Словарн. дефиниц.	0.667		0.672		0.667		0.672	

Таблица 2. Значения метрики F1 для моделей, основанных на представлениях слов из языковой модели ELMo.

Модель	**ELMo RusVectōrēs (ключевое слово, корпус-1000)**		**ELMo DeepPavlov (всё предложение, корпус-1000)**		**ELMo RusVectōrēs (ключевое слово, сбалансир. коллекция)**		**ELMo DeepPavlov (всё предложение, сбалансир. коллекция)**	
k (кол-во соседей)	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус	Проза.ру	Новостной корпус
1	0.809	0.794	0.765	**0.752**	0.812	0.797	0.745	0.758
3	0.826	0.811	**0.773**	0.749	0.833	0.81	0.775	0.753
5	0.834	**0.819**	0.77	0.748	0.845	0.81	0.776	0.756
7	**0.841**	**0.819**	0.767	0.746	**0.857**	0.815	**0.793**	**0.759**
9	0.84	0.816	0.762	0.747	0.856	**0.821**	0.791	0.753
Словарн. дефиниц.	0.772		0.716		0.772		0.716	

Что касается разницы в показателях качества между сбалансированной коллекцией и Корпус-1000, то здесь можно отметить небольшое снижение F1-меры для всех моделей, обученных на коллекции Корпус-1000. В Корпус-1000 включались только контексты с одним из «родственников», что повлияло на контекстное разнообразие данной коллекции. Сбалансированная коллекция, напротив, более репрезентативна в плане множества примеров.

Таблица 3. Значения метрики F1 для моделей, основанных на представлениях слов ELMo: Проза.ру, сбалансированная.

Модель / k	**ELMo RusVectōrēs (всё предложение)**	**ELMo DeepPavlov (целевое слово)**	**ELMo-ruwikiruscorpora (нелемматизированная, целевое слово)**
1	0.807	0.723	0.776
3	0.824	0.73	**0.794**
5	**0.827**	0.738	0.792
7	0.824	0.736	0.792
9	0.821	**0.742**	**0.794**
Словарн.дефиниц.	0.772	0.716	-

Контекстуализированные представления слов из языковой модели ELMo можно извлекать двумя разными способами, поэтому был проведен эксперимент с целью выявить, какой из них лучше подходит для задачи разрешения неоднозначности, а также конкретной модели. В первых двух столбцах таблицы 3 показаны результаты классификации на векторах слов ELMo от RusVectōrēs (извлекался вектор всего предложения) и DeepPavlov (вычислялся только вектор ключевого слова). Полученные результаты демонстрируют, что эти способы извлечения данных из языковых моделей ухудшили качество классификации для обеих моделей (по сравнению с данными из таблицы 2).

Для изучения влияния лемматизации (приведения слов к нормальной форме) на результаты снятия лексической неоднозначности применялись две модели ELMo от RusVectōrēs: модель, обученная на лемматизированном корпусе «Тайга», и модель, обученная на нелемматизированном Национальном корпусе русского языка и русском сегменте Википедии. В качестве обучающей коллекции использовалась сбалансированная коллекция Проза.ру в двух вариантах – с приведением слов к нормальной форме и без. Результаты для нелемматизированной обучающей коллекции представлены в последнем столбце таблицы 3, а для лемматизированной – в таблице 2. Эксперимент показывает, что для сгенерированных обучающих коллекций ELMo-модель, обученная на леммах, дает более высокое качество, чем модель, обученная на ненормализованных словоформах. Таким образом, для русского языка для разрешения лексической

неоднозначности предпочтительнее обучать и тестировать модели на лемматизированных данных, так как они не содержат дополнительной морфологической информации, являющейся излишней для лексико-семантической задачи.

Ввиду того, что словарные дефиниции являлись ценным источником информации для моделей разрешения неоднозначности, начиная с самых ранних работ в этой области, был проведен эксперимент с целью оценки модели, обученной на автоматически сгенерированной обучающей коллекции, расширенной словарными дефинициями и примерами употребления ключевых слов. Для изучения влияния дополнительных данных на качество предсказаний словарные дефиниции и примеры употребления слов, уже использованные ранее в базовом решении, были добавлены в сбалансированные коллекции новостей и Проза.ру. Для снятия неоднозначности применялась наиболее успешная конфигурация модели – контекстуализированные представления ELMo RusVectōrēs для ключевых слов, к которым затем был применен kNN-классификатор. Результаты разрешения неоднозначности представлены в таблице 4. Полученные экспериментальные данные демонстрируют совсем небольшие улучшения в качестве предсказаний для модели, обученной на Проза.ру, и увеличение F1 на 2% для новостной модели.

Таблица 4. Значения метрики F1 для моделей, основанных на представлениях слов ELMo: Проза.ру и Новостной корпус, сбалансированные коллекции, дополненные словарными дефинициями.

Модель / k	**ELMo RusVectōrēs (ключевое слово) Проза.ру**	**ELMo RusVectōrēs (ключевое слово) Новостная коллекция**
1	0.819	0.824
3	0.835	0.832
5	0.847	0.828
7	**0.859**	0.834
9	0.858	**0.842**

В последнем эксперименте этой серии на тестовом наборе данных RUSSE-RuWordNet сравнивалось качество предсказаний моделей, обученных на

автоматически и вручную размеченных данных. Здесь, как и в предыдущих случаях, использовались контекстуализированные вектора слов ELMo от RusVectōrēs. Для каждого целевого значения из набора данных RUSSE-RuWordNet создавались 5 случайных разбиений примеров на обучающие и тестовые данные в соотношении 2:1. Затем эти данные применялись для обучения и тестирования пяти различных моделей разрешения неоднозначности. Среди результатов, полученных для каждого классификатора, брался максимальный, а затем финальное значение F1 представлялось как среднее этих пяти значений. F1-мера на вручную размеченных данных при такой процедуре обучения составила **0.917**. Качество предсказаний модели, обученной на новостном корпусе, на этих 5 тестовых выборках составило **0.84 F1**. Результат работы системы, обученной на комбинации новостной коллекции с каждым из обучающих подмножеств, описанных выше, составил **0.94 F1**. Таким образом, обучающая выборка, представленная комбинацией автоматически и вручную размеченных данных, показывает наилучший результат предсказаний (0.94 F1) среди всех рассмотренных в экспериментах конфигураций.

Во **втором разделе третьей главы** описывается эксперимент, в котором с помощью разработанного метода генерации обучающих данных была создана обучающая коллекция для многозначных существительных, глаголов и прилагательных. На основе этой коллекции была обучена модель, которая применялась для предсказания значений всех многозначных слов в тексте. В качестве тестовых данных, которые размечала обученная модель, использовались статьи из категории экономика портала Wikinews.

В **подразделе 3.2.1** приведены количественные характеристики многозначных слов трех частей речи и их однозначных родственных слов. Большинство многозначных слов в тезаурусе имеют 2 или 3 значения, а предложенный метод способен обеспечить 80-90% многозначных слов каждой части речи однозначными родственниками минимум для двух значений. Более 75% неоднозначных слов каждой из трех частей речи имеют однозначных родственников для всех своих значений. Таким образом, алгоритм может

обеспечить размеченными обучающими примерами большинство многозначных слов в тезаурусе.

Также в данном подразделе продемонстрированы характеристики самих однозначных родственных слов относительного того, к многозначному слову какой части речи они относятся. Большинство однозначных родственников расположены на более удаленных расстояниях от ключевого слова. Можно отметить, что пропорция близких родственников, таких как гипонимы и гиперонимы, выше для существительных и глаголов, тогда как для прилагательных эти пропорции относительно низкие. Помимо этого, когипонимы и когипонимы, расположенные на расстоянии 3 или 4 шагов, вносят большой вклад в широкий охват многозначных слов и их значений. Эти факты справедливы для всех рассматриваемых частей речи.

Подраздел 3.2.2. посвящен самому эксперименту по полуавтоматической разметке всех многозначных слов в тексте. Он состоял из следующих этапов: сначала создавалась обучающая коллекция для всех многозначных слов, которые были обнаружены в выборке новостных статей из секции экономики с ресурса Wikinews, затем на полученной коллекции обучалась модель, приписывающая значения многозначным словам из этого набора данных. Наконец, эти предсказания вручную были проверены и скорректированы.

Конфигурация модели и параметры обучающих данных были выбраны на основании ранее полученных выводов: в описываемом эксперименте использовалась лемматизированная сбалансированная новостная коллекция, kNN-классификатор, базирующийся на контекстуализированных представлениях слов ELMo-модели от RusVectōrēs, обученной на лемматизированном корпусе «Тайга».

В ходе ручной верификации сделанных моделью предсказаний были получены следующие метрики качества: **0.8 F1** для существительных; **0.72 F1** для глаголов; **0.8 F1** для прилагательных. Помимо этого, результаты экспертного анализа размеченных моделью текстов позволяют сделать вывод, что предварительная аннотация, сделанная классификатором, облегчает ручную работу по разметке данных.

В **подразделе 3.2.3** описаны типы ошибок предсказаний значений, выявленные в процессе ручной обработки аннотированных моделью текстов.

1) Ошибки, вызванные отсутствием нужного значения многозначного слова в тезаурусе RuWordNet.
2) Ошибки, связанные с тем, что многозначное слово является частью устойчивого словосочетания.
3) Ошибки, связанные с тем, что многозначное слово является частью имени собственного.
4) Ошибки из-за пересечения морфологической неоднозначности с лексической.
5) Ошибки, сделанные собственно моделью разрешения неоднозначности.

Подраздел 3.2.4. содержит итоги эксперимента по предсказанию значений всех слов в тексте. Его результаты демонстрируют, что разработанный в рамках реферируемого диссертационного исследования алгоритм подходит для генерации обучающих данных вне зависимости от их части речи.

Третий раздел третьей главы посвящен исследованию разрешения неоднозначности на основе псевдоаннотированной коллекции. В этой части работы изучается метод для автоматической разметки текстов, который использует ансамбль моделей, предварительно обученных на данных, размеченных методом однозначных родственных слов.

В **подразделе 3.3.1.** описываются применяемые в данном эксперименте текстовые данные и модели. Контексты для обучающей коллекции извлекались из новостного корпуса и новостного сегмента корпуса «Тайга». Оценка моделей разрешения неоднозначности проводилась на вручную размеченных статьях, содержащих ключевые многозначные слова, из ресурса Wikinews.

Коллекция, собранная с помощью метода однозначных «родственников», выступала в качестве обучающих данных для трех моделей: две из них используют предварительно обученную языковую модель ruBERT от DeepPavlov, а другая базируется на языковой модели ELMo от RusVectōrēs, обученной на

лемматизированном корпусе «Тайга». Первая модель – это тонко настроенный (*fine-tuned*) ruBERT с выходным слоем, предназначенным для классификации последовательностей: линейный слой, получающий на вход конкатенированные представления ключевого слова с четырех последних слоев предварительно обученного трансформера. Вторая модель (*context-gloss pair BERT*) – это классификатор пар предложений, которые были представлены контекстом с ключевым многозначным словом и словарной дефиницией одного из его значений. Также для предсказания значения слова применялась логистическая регрессия, использующая векторные представления слов из ELMo в качестве признаков.

В **подразделе 3.3.2.** описывается методология разметки текстов на основе ансамбля моделей. Для предсказания значений ключевых слов в корпусе предварительно обученные модели использовались в ансамбле, так как синтетические данные, полученные с помощью метода однозначных родственных слов, могут содержать ошибки, и, соответственно, вносить «шум» в модели. В данном эксперименте также учитывалась степень «уверенности» каждой из моделей – вероятностные предсказания на тестовом наборе данных анализировались для выявления областей, где те или иные модели совершают наибольшее число ошибок. Чтобы получить окончательную метку класса из вероятностных оценок моделей, к ним применялась весовая функция. В предлагаемой системе каждое предсказание базового классификатора умножалось на значение точности того или иного класса, полученное в ходе оценки модели на тестовом наборе данных. Затем все взвешенные результаты суммировались, и индекс максимальной вероятности возвращался в качестве конечной метки смысла для примера с целевым словом. Эта схема взвешивания позволяла учитывать только предсказания классификаторов с высокой степенью «уверенности».

Стоит также отметить, что в экспериментах применялись различные варианты разметки текстов: учитывая (или не учитывая) принцип “One sense per discourse” (подразумевающем, что все употребления неоднозначного слова в рамках одного текста имеют одно и то же значение) и с дополнением (или без дополнения) данных словарными дефинициями и примерами употребления слов.

В **разделе 3.3.3.** приводятся выводы по результатам вышеописанного эксперимента по разметке текстовых коллекций с помощью ансамбля предварительно обученных моделей.

Результаты, полученные системами, обученными на различных наборах данных, приведены в таблице 5: без использования принципа “One sense per discourse” (вариант (а) – без добавления словарных дефиниций и примеров употреблений слов в обучающие данные, (б) – с добавлением) и с применением данного принципа ((в) – без добавления, (г) – с добавлением). Результаты моделей, повторно обученных на новых текстах, которые были размечены ансамблями, показывают, что эта процедура улучшает качество систем разрешения неоднозначности.

Таблица 5. Усредненные значения F1-меры для всех ключевых многозначных слов на тестовом наборе данных.

Модель / **Набор данных**	**ELMo LogReg**	**Fine-tuned BERT**	**Context-gloss pair BERT**
Набор, размеченный с помощью метода однозначных родственных слов	0.85	0.81	0.79
(а)	0.86	0.84	**0.87**
(б)	**0.86**	0.85	**0.86**
(в)	**0.87**	0.84	0.86
(г)	0.87	**0.88**	0.87

Раздел 3.4. реферируемого диссертационного исследования содержит визуализации контекстуализированных представлений примеров из обучающей коллекции. Вектора также, как и ранее, извлекались из языковой модели ELMo от RusVectōrēs. Все представления слов, взятые для контекстов из обучающей и тестовой коллекции, а также коллекции, использованной в базовом решении, были визуализированы с помощью алгоритма t-SNE, чтобы изучить, как они расположены относительно друг друга в векторном пространстве. Изображение многозначных слов из обучающей коллекции на координатной плоскости продемонстрировали, что большинство значений формируют легко разделимые смысловые кластеры.

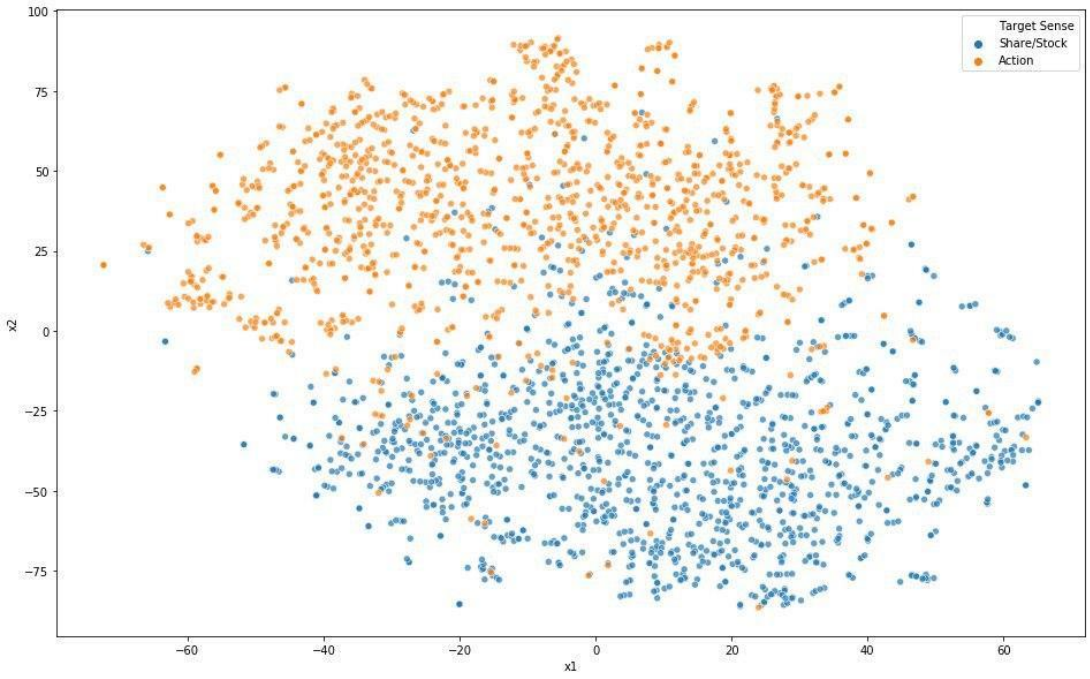

Рисунок 1. Представления для слова *акция*, извлеченные из RusVectōrēs ELMo модели, контексты были взяты из автоматически сгенерированный обучающей коллекции; визуализировано с помощью t-SNE.

В **пятом разделе третьей главы** реферируемого диссертационного исследования описан эксперимент по использованию автоматически сгенерированных коллекций для решения задачи Word-in-Context (WiC). Каждый рассматриваемый пример в этой задаче представлен двумя контекстами с ключевым словом и отражает какое-либо из его значений. Цель состоит в том, чтобы определить, относятся ли эти два контекста к одному значению многозначного слова или к разным. Для оценки полученной модели использовался набор данных для задачи WiC из проекта Russian SuperGLUE[2]. Для классификации данных в этой задаче применялся многослойный перцептрон. В качестве входных данных классификатору подавались контекстуализированные представления ELMo, извлеченные для ключевого слова из каждого контекста, а также вектор ключевого слова, полученный без какого-либо контекста. Accuracy классификатора, обученного на автоматически размеченной выборке, составила **0.91**. На данных, размеченных вручную, была проведена 5-кратная кросс-

[2] https://russiansuperglue.com/

валидация, и усредненное значение метрики accuracy, полученное с помощью этих классификаторов, оно составило **0.8**. Результаты этого эксперимента свидетельствуют о том, что метод генерации и разметки обучающих коллекций подходит не только для задачи разрешения неоднозначности, но и для задачи WiC.

В **заключении** реферируемого диссертационного исследования сформулированы основные результаты работы, дана итоговая оценка и формулируются перспективы дальнейшей разработки темы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

При выполнении диссертационной работы были получены следующие основные **результаты**:

1) Произведен анализ теоретических аспектов создания систем автоматической генерации размеченных обучающих коллекций для задачи разрешения лексической неоднозначности, а также анализ принципов построения моделей разрешения неоднозначности.
2) Разработан метод автоматического сбора и разметки обучающих текстовых коллекций для решения задачи разрешения лексической неоднозначности, базирующийся на концепте однозначных родственных слов. Он учитывает однозначных кандидатов, далеко расположенных от целевых многозначных слов в семантическом графе RuWordNet. Эта особенность позволяет находить обучающие примеры для подавляющего большинства многозначных слов и их значений из тезауруса.
3) Для оценки извлекаемых контекстов был разработан механизм, использующий коэффициент схожести однозначного «родственника-кандидата» и синсетов, близких к целевому значению многозначного слова. В данном методе используется модель word2vec, обученная на корпусе, из которого извлекаются контексты, что позволяет учитывать в оценке дистрибутивные признаки однозначных родственных слов. Этот компонент

фильтрации повышает релевантность примеров, добавляемых в обучающую коллекцию, и, как следствие, уменьшает «шум» в данных.

4) Для оценки метода отбора и ранжирования однозначных родственных слов, разработанного в рамках диссертационного исследования, был проведен сравнительный анализ моделей, обученных на сгенерированных с помощью него обучающих коллекций. Были выделены наиболее успешные стратегии обработки текстовых данных и использования контекстуализированных векторных представлений слов в моделях, с помощью которых достигаются максимальные показатели качества предсказания значений слов в русском языке. Полученные в ходе экспериментов результаты могут быть полезны для дальнейших исследований в области автоматического разрешения лексической неоднозначности на материале русского языка.
5) Разработаны и обучены модели разрешения неоднозначности для русского языка различных архитектур: kNN, логистическая регрессия и многослойный перцептрон на контекстуализированных представлениях слов из предварительно обученных языковых моделей; тонко настроенный ruBERT (*fine-tuned ruBERT*); ruBERT, тонко настроенный на классификацию пар предложений (*context-gloss pair BERT*). Данные модели, обученные на автоматически сгенерированных коллекциях, показывают качество разрешения неоднозначности на уровне с моделями, обученными на вручную размеченных корпусах.

Публикации по теме диссертации

Основные результаты работы изложены в 9 научных статьях:

В изданиях, рекомендованных для защиты в диссертационном совете МГУ им. М.В. Ломоносова:

1) Большина, А. С. Методы автоматического формирования семантически размеченных корпусов [Текст] / А. С. Большина // Вестник Московского университета. Сер. 9. Филология. – 2022а. – № 2. – С. 173–183.
2) Bolshina, A. Generating training data for word sense disambiguation in Russian [Текст] / A. Bolshina, N. Loukachevitch // Proceedings of Conference on

Computational linguistics and Intellectual technologies Dialog-2020. – 2020a. – С.119-132.

3) Bolshina, A. All-words Word Sense Disambiguation for Russian Using Automatically Generated Text Collection. [Текст] / A. Bolshina, N. Loukachevitch // Cybernetics and Information Technologies. – 2020b. – Т. 20., №. 4. – С. 90-107.
4) Bolshina, A. Automatic Labelling of Genre-Specific Collections for Word Sense Disambiguation in Russian [Текст] / A. Bolshina, N. Loukachevitch // Russian Conference on Artificial Intelligence. – Springer, Cham, 2020c. – С. 215-227.
5) Bolshina A. Comparison of Genres in Word Sense Disambiguation using Automatically Generated Text Collections. [Текст] / A. Bolshina, N. Loukachevitch // Fourth International Conference Computational Linguistics in Bulgaria. – 2020d. – С. 156-165.
6) Bolshina A. Exploring the Limits of Word Sense Disambiguation for Russian using Automatically Labelled Collections [Текст] / A. Bolshina, N. Loukachevitch // Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence (LFLAI 2020). – 2020e.

Публикации в журналах, включенных в перечень ВАК:

1) Большина А. С. Создание псевдоаннотированного обучающего корпуса для задачи разрешения лексической неоднозначности с помощью ансамбля моделей [Текст] / А. С. Большина // Интеллектуальные Системы. Теория и приложения. – 2022б. – Т.26, №1. – С.185-189.
2) Bolshina A.S. Weakly Supervised Word Sense Disambiguation Using Automatically Labelled Collections [Текст] / A. Bolshina, N. Loukachevitch // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). – 2021. – Т.33, №6. – С.193-204.

Прочие публикации:

1) Большина А. Оценка лексических замен в задаче автоматической разметки значений слов [Текст] / А. Большина // «Труды молодых учёных», Москва. – 2020. – С.14-26.