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»

Диссертация на соискание учёной степени
кандидата филологических наук

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

Москва – 2022

Оглавление

Введение

Автоматическое разрешение лексической неоднозначности является одной из ключевых задач обработки естественного языка, которая заключается в выборе того значения многозначного слова, в котором оно употреблено в конкретном контексте. «Неоднозначность, свойственная естественному языку и проявляющаяся на различных его уровнях, является серьёзным препятствием для компьютерного анализа текстов» [Митрофанова и др., 2008: 368], поэтому разрешение лексической многозначности широко используется в таких областях, как машинный перевод [Богуславский и др., 2005; Марчук, 2016; Gonzales et al., 2017; Liu et al., 2018; Pu et al., 2018; Raganato et al., 2019], автоматическое извлечение информации из текстов [Zhong, Ng, 2012; Delli Bovi et al., 2015; Hristea, Colhon, 2020], информационный поиск [Blloshmi et al., 2021], построение семантических графов [Alexeyevsky, 2018], разработка вопросно-ответных систем [Ramakrishnan et al., 2003], а также для улучшения качества классификации текстов [Епрев, 2010; Shimura et al., 2019]. Автоматическое разрешение неоднозначности также применяется в специализированных доменах, таких как биомедицина [Пашук и др., 2019; Martínez, Baldwin, 2011; Sabbir et al., 2017; Pesaranghader et al., 2019] и прогнозирование цен акций [Hogenboom et al., 2021].

Существуют три основных подхода к разрешению лексической неоднозначности: основанный на методах машинного обучения с учителем (*supervised machine learning*), на методах машинного обучения без учителя (*unsupervised machine learning)*, а также базирующийся на знаниях (*knowledge-based*). Ввиду того, что модели, обученные без учителя, не используют никаких заранее предопределенных меток значений, результаты их работы трудно оценивать и сравнивать с моделями других типов. Алгоритмы разрешения неоднозначности, базирующиеся на базах знаний, показывают точность предсказаний значений, сравнимую с методами обучения с учителем, но обычно они их не превосходят. Именно поэтому сейчас большинство передовых моделей

разрешения лексической неоднозначности основаны на методах обучения с учителем.

Необходимым компонентом любой системы машинного обучения с учителем является размеченный корпус, а если речь идет о подходах на основе нейронных сетей, то аннотированных данных требуется очень много. Ручная разметка больших текстовых коллекций требует много времени и трудозатрат, а иногда для разметки необходимо привлекать экспертов. Больших корпусов, аннотированных вручную, существует не так много, и в основном они для английского языка. Проблема отсутствия или недостатка размеченных данных в англоязычной терминологии обозначается как *knowledge acquisition bottleneck*, и для ее решения разрабатывается большое количество методов по автоматическому сбору и разметке обучающих коллекций. Исследователи используют различные подходы, в которых применяются электронные ресурсы (например, Википедия[1] и Викисловари[2]), лексико-семантические ресурсы, параллельные корпуса текстов, а также различные алгоритмы (например, алгоритм распространения меток).

Объектом исследования в данной работе являются корпуса с семантической разметкой.

Предметом настоящей 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.).

[1] https://www.wikipedia.org
[2] https://www.wiktionary.org

Целью данной диссертационной работы является разработка метода автоматического сбора и разметки корпуса русского языка для задачи разрешения лексической многозначности, а также его программная реализация. В рамках настоящего исследования на материале русского языка рассматривается подход, основанный на однозначных родственных словах.

Для достижения данной цели были поставлены следующие ***задачи***:

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 диссертационного 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 обеспечивает данному методу возможность найти обучающие примеры для подавляющего числа 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 и размеченных 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 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[3], тезаурус 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.

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.

[3] https://ru.wikinews.org/wiki/Заглавная_страница

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 и публикации.

Результаты исследования опубликованы в 9 статьях, 6 из которых 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. – 2022. – № 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. – 2020. – 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. – 2020. – 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, 2020. – 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. – 2020. – 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). – 2020.

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. – 2022. – Т.26(1). – С.185-189.
2) *Bolshina, A.* 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.

Основные положения докладывались на следующих ***конференциях***:

1) 26-я международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-2020»;
2) Международная конференция по компьютерной лингвистике в Болгарии (CLIB 2020);
3) 18-я национальная конференция по искусственному интеллекту (КИИ-2020);

4) Международная конференция «Лингвистический форум 2020: Язык и искусственный интеллект»;
5) Открытая конференция ИСП РАН им. В.П. Иванникова 2021;
6) XII Международная научная конференция «Интеллектуальные системы и компьютерные науки».

Структура и объем научно-квалификационной работы.

Работа 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.

Глава 1. Теоретические аспекты задачи автоматического разрешения неоднозначности

Неоднозначность – это неотъемлемое свойство естественных языков. Изучение лексической многозначности ведется не только в области прикладной лингвистики, но и в первую очередь в рамках теоретической семантики. Для определения полисемии Ю. Д. Апресян [Апресян, 1995: 186] использует понятие сходства значений: «Значения a_i и a_j слова А называются сходными, если существуют такие уровни семантического описания, на которых их толкования (семантические деревья) или коннотации имеют нетривиальную общую часть, и если она выполняет в толкованиях одну и ту же роль относительно других семантических компонентов». Таким образом, согласно [Апресян, 1995: 187] «Слово А называется многозначным, если для любых двух его значений a_i и a_j найдутся такие значения a_1, a_2, …, a_k, a_l, что a_i сходно с a_1, a_1 – с a_2 и т.д., a_k – с a_l и a_l – с a_j». Е. В. Падучева приводит следующее определение лексической многозначности [Падучева, 2004: 147]: «Многозначное слово полисемично (а не омонимично), если его значения связаны друг с другом системными, т. е. повторяющимися соотношениями».

В области лексической семантики строго различают полисемию и омонимию, которая подразумевает лишь внешнее совпадение слов, в значениях которых нет никакой общей части. Однако, как было отмечено в работе [Иомдин, 2014: 89], «в сфере компьютерной лингвистики различие между полисемией и омонимией, последовательно проводимое в теоретической семантике и (до последнего времени) в лексикографии <…> нерелевантно, поскольку для решения задачи определения значения слова в тексте наличие или отсутствие смысловой связи между возможными кандидатами несущественно». Также, стоит разделять языковую неоднозначность и речевую: «если языковая неоднозначность — это способность слова, выражения или конструкции иметь различные смыслы, т. е. это свойство языковых единиц, то речевая неоднозначность — это реализация данного свойства в конкретном высказывании» [Зализняк, 2006: 22].

В работах [Апресян, 1995; Падучева, 2004; Кустова, 2004] проводились аналогии между лексической многозначностью и словообразованием: связи между значениями многозначного слова напоминают отношения между словом и его словообразовательными дериватами, что позволяет говорить о «семантической деривации» как особом типе словообразовательных процессов» [Апресян, 1995: 187]. Ввиду того, что во внутренних характеристиках полисемии прослеживается близость к словообразованию, в литературе выделяют регулярную лексическую многозначность: «Полисемия слова А со значениями a_i и a_j называется регулярной, если в данном языке существует по крайней мере еще одно слово В со значениями b_i и b_j, семантически отличающимися друг от друга точно так же, как a_i и a_j, и если $a_i – b_i$, $a_j – b_j$ попарно несинонимичны» [Апресян, 1995: 189]. Модели семантической деривации можно считать продуктивными, если они «многократно повторяются в семантических парадигмах других слов» [Падучева, 2004: 249]. Примером продуктивной полисемии является следующее соотношение значений [Апресян, 1995: 192]: «всякое существительное со значением ‘сосуд’ может обозначать также ‘количество вещества, входящего в сосуд’».

В работах по лексической семантике различают два основных типа связи между значениями многозначного слова – это метафора и метонимия [Зализняк, 2006: 57]. Исследователи также выделяют связь значений на основе синекдохи и по функции [Кобозева, 2000: 170]. Механизм семантической деривации, основанный на метафоре, базируется на сходстве описываемых явлений или объектов: «*язык* пламени, *язык* колокола и *язык* во рту человека похожи по форме» [Кобозева, 2000: 170]. Метонимическая связь между двумя значениями характеризуется отношением смежности обозначаемых концептов: например, существительное *зал* в значениях ‘зал для собраний, занятий’ и ‘зрители в зале’. Частным случаем метонимии является синекдоха, представляющая собой перенос свойств с части на целое или наоборот: существительное *копейка* в значениях ‘денежная единица, равная одной сотой рубля’ и ‘денежные средства’. Примером

значений, схожих по функциям определяемых объектов, являются значения слова *язык* ‘орган’ и ‘пленный’ («по функции участия в передаче информации» [Кобозева, 2000: 170]).

В работе [Апресян, 1995: 182] выделяется три топологических типа полисемии: радиальная, цепочечная и радиально-цепочечная. Если одно центральное значение объединяет все остальные значения многозначного слова, то такая многозначность обозначается как радиальная. Если каждое значение связано только с другим ближайшим к нему значением, то речь идет о цепочечной полисемии. Радиально-цепочечная полисемия является наиболее частым случаем.

Стоит также отметить такое свойство семантики многозначных слов как диффузность, когда в их значениях есть «дискретные» области (дискретность которых обязана их противопоставленности другим дискретным областям) и «диффузные» области, границы между которыми носят градуальный характер» [Зализняк, 2006: 57]. Эта характеристика связана с природой многозначности, «которая устроена не дискретно в языке как системе» [Зализняк, 2004: 41]. Границы между некоторыми значениями слов являются нечеткими, и для того, чтобы их выявить, необходимо ответить на вопрос, «представляют ли два употребления слова одно и то же его узуальное значение[4] или два разных значения» [Кобозева, 2000: 162]. Например, в работе [Кобозева, 2000: 166] предлагается учитывать следующие факторы при выделении значений слов: «морфосинтаксическую, лексическую и семантическую сочетаемость лексемы по ее валентностям; парадигматические связи (корреляции) лексемы; грамматические ограничения».

Характеризуя общие теоретические аспекты полисемии, можно отметить, что имеются две различные задачи в области представления лексической многозначности: «Одна состоит в том, чтобы оптимальным образом организовать информацию, нужную для пользователя. Другая — в том, чтобы понять, как

[4] «Узуальное значение слова, или узема, есть абстракция от в принципе бесконечного ряда актуальных значений слова в речи, инвариант актуальных значений, все различия между которыми могут быть объяснены действием экстралингвистических факторов» [Кобозева, 2000: 158].

устроена многозначность в языке, а это означает провести, тем или иным образом, границу между воспроизводимым и порождаемым и выявить механизмы семантической деривации, которыми пользуется говорящий» [Зализняк, 2006: 45].

Человеку обычно не составляет труда понять, в каком значении было употреблено слово в том или ином контексте, однако разработка автоматической системы, которая бы выполняла подобную задачу, является нетривиальной задачей. Автоматическое разрешение лексической неоднозначности описывается, как ИИ-полная задача [Mallery, 1988], то есть проблема, для решения которой необходимо создать «сильный искусственный интеллект» [Searle, 1980], способный мыслить, как люди. Научные исследования в этой области имеют многолетнюю историю: задача разрешения лексической неоднозначности была сформулирована в конце 40-х гг. 20-го века в рамках работы, посвященной машинному переводу [Weaver, 1955]. Сложность решения этой задачи считалась одним из значимых препятствий на пути разработки систем машинного перевода [Bar-Hillel, 1960].

В этой главе будут подробно рассмотрены все неотъемлемые составляющие задачи разрешения многозначности: представление значений слов, подготовка обучающей коллекции, формирование признакового пространства и выбор конкретного алгоритма разрешения неоднозначности.

1.1. Представление значений слов

«Традиционно к лексическому значению относят наиболее существенную часть связанной с лексемой информации – ее денотат, сигнификат и некоторую часть прагматической информации» [Кобозева, 2000: 80]. Обычно толкования слов в словарях состоят именно из сигнификативного компонента, который включает в себя существенные признаки и свойства обозначаемых словом объектов.

Инвентарь значений – это система, формализующая представление значений слов. Она определяет сам словарь, какие значения есть у слов, и какими

метками они обозначаются. Обычно в исследованиях в качестве инвентаря значений применяются толковые словари или семантические графы.

Для описания значений слов обычно используется перечислительный подход, при котором значения представлены в виде нумерованного списка. Однако в разных источниках представлены разные способы деления слов на значения. Рассмотрим, к примеру, толковые словари русского языка Ушакова[5] и Ожегова[6]. В словарной статье к слову *ключ* в словаре Ожегова описано 6 значений этого слова, в то время как в словаре Ушакова – 4. В словаре Ожегова есть следующее отдельное значение: «Приспособление для отвинчивания или завинчивания, откупоривания, приведения в действие механизма». А в словаре Ушакова оно входит в один ряд с другими: «Металлическое приспособление для отпирания и запирания замка. *Запереть дверь на ключ. Подобрать ключ к замку*. || То же для отвинчивания гаек и болтов. *Подвинтить гайку французским ключом*. || То же для вскрытия консервных банок. || То же для электрических выключателей особого вида. || То же для завода часов и всяких иных механизмов. || То же для натягивания струн в струнных инструментах типа фортепьяно, арфы.»

Данный пример хорошо иллюстрирует, что в разных источниках деление слов на значения устроено по-разному, и поэтому перед исследователем в области автоматического разрешения неоднозначности встает задача определения степени **гранулярности** (детализации) значений многозначного слова. Многое в этом вопросе зависит от предполагаемой области применения приложения. Бывают случаи параллельной многозначности переводных эквивалентов слов в разных языках, поэтому для машинного перевода можно не учитывать совпадающие в нескольких языках значения. Однако для других задач обработки языка это может быть неприменимо: «<…> например, слово *interest* является многозначным в английском, итальянском и французском. Вследствие этого для задачи машинного перевода будет излишним выделять все его значения (например, «доля» и «увлечение»), а в других приложениях, таких как извлечение

[5] https://ushakovdictionary.ru/
[6] https://slovarozhegova.ru/

информации, это необходимо, так как это позволит отделить тексты про финансы и хобби» [Navigli, 2009: 5]. Как отмечалось в [Иомдин, 2014: 90], «основные трудности, на которые наталкиваются разработчики, – отсутствие единообразного описания значений и недостаточная системность существующих лексикографических источников».

Помимо этого, исследователи должны делать выбор относительно количества выделяемых значений не только исходя из приложения, где будет применяться система, но и из вычислительных возможностей, так как большое число значений увеличивает количество параметров в моделях, и, как следствие, замедляет их обучение. В исследовании [Vial et al., 2019] описывались методы, сокращающие количество значений, которые использовались в системе разрешения многозначности: были оставлены только те значения, которые необходимы для различения смысла имеющихся в лексической базе знаний слов. Было показано, что данная процедура позволяет улучшить качество снятия неоднозначности, а также сократить результирующий размер моделей.

1.2. Лингвистические ресурсы

Существует два основных типа источников знаний: структурированные и неструктурированные. К первой категории относятся, например, машиночитаемые (электронные) словари, которые представляют собой базу данных со словарными статьями, по которым можно быстро осуществлять поиск, а также удобно использовать для различных вычислительных задач. Электронные словари были особенно популярны в период с 80-х гг. 20-го века до возникновения и широкого распространения семантической сети WordNet [Miller, 1995; Fellbaum, 1998]. Среди электронных словарей можно отметить такие, как Oxford Dictionary of English[7], Longman Dictionary of Contemporary English (LDOCE)[8], многоязычный Викисловарь (Wiktionary)[9].

[7] https://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/acref-9780199571123
[8] https://www.ldoceonline.com/
[9] https://www.wiktionary.org/

Еще одним структурированным ресурсом является тезаурус. **Тезаурус** – это «словарь, в котором слова и словосочетания с близкими значениями сгруппированы в единицы, называемые понятиями, концептами или дескрипторами, и в котором явно (в виде отношений, иерархии) указываются семантические отношения между этими понятиями (концептами, дескрипторами)» [Лукашевич, 2011: 20]. Примерами тезаурусов являются следующие ресурсы: информационно-поисковый тезаурус Европейского союза EUROVOC[10]; тезаурус исследовательской службы Конгресса США (Legislative Indexing Vocabulary)[11]; РуТез (тезаурус для русского языка) [Лукашевич, 2011].

Самым известным лексико-семантическим ресурсом, используемым в области автоматической обработки текстов, является тезаурус WordNet для английского языка. Он состоит из семантических сетей для глаголов, существительных, прилагательных и наречий. Базовым понятием семантических графов такого типа является **синсет**, представляющий собой синонимический ряд, в который входят слова со схожими значениями. Синсеты формируют узлы семантического графа и соединяются друг с другом такими отношениями, как гипонимия, гиперонимия, меронимия, антонимия и т.д. Тезаурусы типа WordNet также имеются и в других языках, например, RuWordNet для русского языка [Loukachevitch et al., 2016], GermaNet[12] для немецкого языка, DanNet[13] для датского языка.

С понятием тезауруса непосредственно связано понятие онтологии. В области автоматической обработки естественного языка под понятием «онтология» понимается «некоторый компьютерный ресурс, представляющий собой некоторое описание взгляда на мир применительно к конкретной области интересов» [Лукашевич, 2011: 101]. Среди наиболее известных онтологий можно выделить такие, как FrameNet[14] и BabelNet[15] [Navigli, Ponzetto, 2012]. FrameNet –

[10] https://eur-lex.europa.eu/browse/eurovoc.html
[11] https://www.congress.gov/browse/legislative-indexing-vocabulary/106th-congress
[12] https://uni-tuebingen.de/en/142806
[13] https://cst.ku.dk/english/projects/dannet/
[14] https://framenet.icsi.berkeley.edu/fndrupal/
[15] https://babelnet.org/about

это лексический ресурс, базирующийся на фреймовой семантике [Fillmore, Baker 2001]. Как гласит определение с официального сайта проекта, BabelNet – это инновационный многоязычный энциклопедический словарь, содержащий обширную лексикографическую и энциклопедическую информацию о входящих в него терминах, а также семантическую сеть/онтологию, которая соединяет понятия и именованные сущности в очень большой граф семантических отношений, состоящий примерно из 20 миллионов узлов. Эта онтология объединяет в себе различные источники: сеть WordNet для английского языка, коллекции семантических сетей для различных языков (Open Multilingual WordNet[16]), Википедию, Викисловари, OmegaWiki[17] (коллаборативный мультиязычный словарь, созданный интернет-пользователями), WikiData[18] [Vrandečic´, 2012] и другие ресурсы. BabelNet версии 5.0 содержит данные для 500 языков.

Что касается неструктурированных ресурсов, то к ним обычно относят корпуса и текстовые коллекции, использующиеся в качестве обучающих и тестовых данных на различных технологических соревнованиях. **Корпуса** – это «представительный массив текстов, собранный по определённому принципу (по жанру, авторской принадлежности и т.п.)» [Большакова и др., 2017]. Различают аннотированные и неаннотированные корпуса. Неразмеченные корпуса можно использовать для обучения языковых моделей или моделей разрешения неоднозначности методом обучения без учителя [Navigli, 2009: 7]. К ресурсам такого типа относят следующие: Британский национальный корпус[19]; Американский национальный корпус[20]; собрание текстов Wall Street Journal [Charniak et al., 2000]. Для решения задач автоматической обработки текстов на русском языке сейчас широко используются Национальный корпус русского языка (НКРЯ)[21], корпус Araneum Russicum [Benko, 2014], Генеральный Интернет-

[16] http://compling.hss.ntu.edu.sg/omw/
[17] http://www.omegawiki.org/
[18] https://www.wikidata.org/wiki/Wikidata:Main_Page
[19] http://www.natcorp.ox.ac.uk/
[20] https://www.anc.org/
[21] https://ruscorpora.ru/new/

корпус русского языка [Belikov et al., 2013; Piperski et al., 2013], корпус «Тайга» [Shavrina, Shapovalova, 2017]. Также в последнее время в качестве корпусов всё чаще стали использовать веб-ресурсы, например, тексты статей Википедии.

Семантическая разметка – это «разметка, сопоставляющая лексеме толкование (номер значения в авторитетном толковом словаре) или указывающей место в лексической классификации (тезаурусе)» [Рахилина и др., 2006: 445]. Семантически размеченные корпуса – это текстовые коллекции, в которых каждому многозначному слову приписана метка, отражающая его значение в конкретном контексте. Тексты, которые формируют корпус, могут быть взяты из различных источников: художественная литература, новостные и научные статьи, социальные сети и т.п. Для аннотирования корпусов используются заранее выбранные инвентари значений, и сейчас чаще всего в качестве них применяются инвентари семантических сетей типа WordNet для соответствующих языков или значения из семантической сети BabelNet. Рассмотрим пример аннотации предложения с помощью инвентаря значений из семантической сети для русского языка RuWordNet.

(1)	Дом	Бхутто	в	Лахоре	оцеплен	полицией.
	2526-N 130946-N	-	-	-	115194-V	9828-N

Все примеры в корпусе автоматически разбиваются на отдельные слова, и те слова, что представлены в семантическом графе RuWordNet, затем размечаются экспертами. Здесь и далее в работе, под **целевым (ключевым) многозначным словом** будет подразумеваться многозначное слово, с которого требуется снять неоднозначность. В представленном предложении есть многозначное слово *дом*, которое входит в два синсета в RuWordNet: 130946-N ‘домашний очаг’ и 2526-N ‘здание’. При его разметке учитывался контекст, в соответствии с которым была проставлена нужная метка значения (обозначена подчеркиванием). Все остальные аннотированные слова в данном примере имеют только одно значение, а те слова, что отсутствуют в словаре тезауруса, помечены

прочерками. При выполнении семантической разметки нередко возникает следующая проблема, упомянутая в [Иомдин, 2014: 90]: «несогласие экспертов (носителей языка) при выделении значений слов и интерпретации конкретных употреблений, затрудняющее работу над созданием аннотированных корпусов текстов (inter-judge variance)».

Выделяются два типа задач, на которых тестируются системы разрешения лексической неоднозначности: снятие неоднозначности с ограниченного, заранее определенного набора многозначных слов (*lexical sample* или *targeted WSD*), и предсказание значений всех слов открытых классов единиц языка[22] в тексте (*all-words WSD*). Создание моделей для решения задач второго типа является более трудоёмкой задачей, требующей обучающие корпуса с широким охватом неоднозначных слов. Ввиду наличия двух формулировок задачи разрешения лексической неоднозначности, существуют корпуса, в которых в каждом предложении размечено только одно ключевое многозначное слово, и корпуса, где аннотировано каждое слово (иногда даже однозначные). Первый тип данных обычно применяется для обучения и тестирования моделей, ориентированных на lexical sample WSD. Вторая разновидность корпусов используется как для lexical sample, так и для all-words WSD.

Как и во многих других областях обработки естественного языка, наибольший объем размеченных данных имеется для английского языка. Самым популярным корпусом с разметкой значений слов считается SemCor [Miller et al., 1994], который является подкорпусом Брауновского корпуса[23], выпущенного в 1967 году. Он состоит из 352 документов, 226 036 токенов, размеченных вручную значениями из инвентаря значений тезауруса WordNet, и покрывает 38 022 уникальных неоднозначных слова и 25 915 синсетов [Pasini, 2020: 4938]. По сравнению с другими ресурсами с ручной разметкой этот корпус включает в себя наибольшее число неоднозначных слов и синсетов, поэтому он чаще всего

[22] «Класс считается открытым, если он очень большой и легко может быть увеличен. <…> Открытые классы образуют корневые морфемы существительных, глаголов, прилагательных, которые и считаются носителями лексического значения» [Кобозева, 2000: 76].
[23] http://korpus.uib.no/icame/brown/bcm.html

используется для обучения моделей разрешения неоднозначности методом обучения с учителем. Однако исследователи отмечают, что частотность некоторых слов в корпусе отличается от статистики их употребления в настоящее время ввиду того, что Брауновский корпус был создан в 60-е года. Например, наибольшая встречаемость у слова *pipe* в корпусе в значении ‘курительная трубка’, в то время как сейчас наиболее частотное значение этого слова – это ‘металлическая труба’ [Pasini, 2020: 4938]. Помимо этого, для некоторых значений слов в данном корпусе очень мало примеров [Agirre, De Lacalle, 2004: 1123].

Для преодоления проблемы высокой степени детализации значений сети WordNet, и, как следствие, корпуса SemCor, который использует его инвентарь значений, исследователи [Hovy et al., 2006] создали корпус OntoNotes. В этом корпусе значения слов организованы иерархически: более специализированные значения слов объединяются в группы более глобальных и общих значений. Таким образом, создатели корпуса OntoNotes пытались решить проблему чрезмерной гранулярности значений, присущую WordNet. При этом корпус включает не очень много уникальных неоднозначных слов (3 380 слова). Более того, в нем нет примеров для прилагательных и наречий.

Корпус WNGT (Princeton WordNet Gloss Corpus)[24] был создан полуавтоматически. Он состоит из всех словарных определений и примеров употреблений слов, содержащихся в сети WordNet, в которых слова размечались либо экспертами, либо с помощью вручную созданных эвристик. Данный корпус содержит 441 656 размеченных примеров для 31 396 уникальных многозначных слов [Pasini, 2020: 4938].

В исследовании [Navigli, 2009: 7] упоминаются также такие корпуса, как DSO [Ng, Lee, 1996], созданный Организацией оборонных исследований (Defense Science Organisation), и набор данных Open Mind Word Expert [Mihalcea, Chklovski, 2003], аннотированный совместными усилиями пользователей сети

[24] https://wordnetcode.princeton.edu/glosstag.shtml

Интернет. Сводная статистика по некоторым из размеченных корпусов наглядно представлена в работе [Pasini, 2020].

Регулярно проводятся технологические соревнования для сравнения качества работы различных моделей разрешения неоднозначности. Оценка алгоритмов осуществляется на одних и тех же размеченных данных, и на данный момент большинство существующих аннотированных наборов было создано в рамках соревнований SemEval (ранее Senseval), которые проходят каждые три года, начиная с 1998. Все текстовые коллекции, выпущенные в рамках данных соревнований, различаются по объему и сложности. Они содержат разметку для различных частей речи и доменов, и бывают доступны для разных языков.

На данный момент наиболее крупной платформой для оценки моделей разрешения неоднозначности считается мультиязычный проект XL-WSD [Pasini et al., 2021], предоставляющий аннотированные по значениям тестовые и обучающие выборки для 18 языков из 6 различных семей. Помимо этого, в рамках данного проекта с помощью моделей машинного перевода и размеченных обучающих корпусов для английского языка (SemCor и WNGT) были созданы обучающие данные для 15 языков (кроме китайского и корейского).

Основными источниками размеченных обучающих и тестовых данных на русском языке являются три разных набора данных, представленных на технологическом соревновании по автоматическому извлечению значений слов из неразмеченного корпуса текстов для русского языка RUSSE-2018[25] [Panchenko et al., 2018]. Первая коллекция состоит из контекстов, взятых из Национального корпуса русского языка. Второй набор данных содержит примеры из статей Википедии. Последняя выборка основана на Активном словаре русского языка [Апресян и др., 2017] и состоит из контекстов, взятых из раздела примеров и иллюстраций данного словаря.

В таблице 1 приведено количество многозначных слов и размеченных примеров из текстовых коллекций для некоторых языков из проекта XL-WSD

[25] https://nlpub.github.io/russe-wsi-kit/

(первая часть таблицы) и данные по коллекциям на русском языке, размеченным для оценки решений разработчиков на технологическом соревновании RUSSE-2018 (вторая часть таблицы).

Таблица 1. Количественные данные обучающих и тестовых наборов данных.

Язык	**Количество многозначных слов**		**Количество размеченных примеров**	
	Обучающая выборка	**Тестовая выборка**	**Обучающая выборка**	**Тестовая выборка**
Английский	24658	2199	840471	8062
Баскский	5294	525	197309	1580
Немецкий	2332	166	184952	862
Японский	581	2390	23217	7602
Словенский	1296	93	128395	2032
Русский (wiki-wiki)	4	7	439	638
Русский (bts-rnc)	30	51	3491	6556
Русский (active-dict)	85	168	2073	3729
Русский (active-rnc)	20	-	1829	-
Русский (active-rutenten)	21	-	3671	-
Русский (bts-rutenten)	11	-	956	-

Стоит отметить, что в таблице приведены не все языки из имеющихся в XL-WSD: те языки, что в ней продемонстрированы, были выбраны для того, чтобы показать, как варьируется размер данных. Название набора для русского языка,

указанное в скобках, отражает используемый инвентарь значений и корпус, из которого извлекались контексты для обучения и тестирования. Также для некоторых текстовых коллекций на русском языке, приведенных в таблице, отсутствуют тестовые данные. Статистика, представленная в таблице, наглядно демонстрирует, что для русского языка объем размеченных данных гораздо меньше, а также отсутствует унификация текстовых коллекций по инвентарю значений. Это еще раз подчеркивает актуальность решения проблемы нехватки размеченных данных в русском языке.

Отсутствие или недостаток аннотированных данных сильно тормозит развитие и исследование систем разрешения неоднозначности для многих языков, а ручная разметка таких корпусов требует много времени и трудозатрат. Для русского языка имеется очень мало размеченных обучающих коллекций, поэтому задача автоматического сбора обучающих корпусов актуальна для русского языка. О возможных способах преодоления проблемы нехватки размеченных данных речь пойдет в разделе 1.5.

1.3. Представление контекста

Как и во многих других задачах в области автоматической обработки текстов, для подготовки текстовых коллекций в качестве входных данных для обучения моделей разрешения неоднозначности необходимо сначала выполнить токенизацию текста, т.е. разбиение текста на токены, которые обычно соотносятся с отдельными словами. Среди необязательных шагов предварительной обработки текстов можно выделить процедуру удаления стоп-слов, лемматизацию (приведение всех словоформ в тексте к начальной (нормальной) форме), приписывание токенам в тексте меток частей речи и синтаксический парсинг. После первичной подготовки текстов для обучения моделей разрешения неоднозначности требуется тем или иным способом закодировать целевое слово и окружающей его контекст.

Чтобы извлечь контекст, в котором было употреблено неоднозначное слово, обычно используется один из двух подходов: контекст представляется в виде окна слов, окружающих слово, или контекст рассматривается с учетом каких-либо отношений, например, синтаксических. Чаще всего в работах используется локальный контекст в виде окна слов. Из подобного локального контекста обычно извлекаются признаки слов, окружающих целевое неоднозначное слово, которые могут быть лингвистическими, статистическими и структурными: например, часть речи, расстояние до целевого слова [Navigli, 2009: 12], частота встречаемости определенных слов и т.п. Подобные признаки образуют вектора, которые можно использовать в обучении систем разрешения неоднозначности методом обучения с учителем. Стоит отметить, что размер контекста может влиять на качество разрешения неоднозначности; данное свойство было подробно изучено в работах [Yarowsky, Florian, 2002] и [Cuadros, Rigau, 2006]. Размер оптимального контекста также может варьироваться между различными частями речи. В работе [Yarowsky, 1993: 270] отмечалось, что большой контекст лучше подходил для разрешения многозначности существительных, в то время как для глаголов и прилагательных более полезной оказалась информация о словах, близко расположенных к целевому многозначному слову.

Как уже говорилось ранее, существует несколько типов признаков, извлекаемых из контекста, в котором встретилось многозначное слово. Для обучения алгоритмов разрешения лексической неоднозначности могут использоваться признаки, характеризующие более широкий отрезок текста, чем просто окно определенного размера. Контекст может являться предложением, в которое входит целевое слово, абзацем или даже всем документом. Для представления информационных характеристик таких контекстов могут использоваться, например, вектора тем, полученные в ходе процедуры тематического моделирования [Boyd-Graber et al., 2007]. Помимо этого, в задаче разрешения лексической неоднозначности могут применяться синтаксические признаки, характеризующие зависимости между целевым словом и остальными

словами в контексте [Martínez et al., 2002; Lee, Ng, 2002]. В качестве признаков также используют информацию о словосочетаниях, например в работе [Yarowski, 1993] было показано, что в одинаковых коллокациях многозначное слово употребляется в одном и том же значении (этот принцип получил название “one sense per collocation”). Семантические признаки также широко используются для обучения моделей разрешения лексической неоднозначности [Melacci et al., 2018]. Например, в качестве этих признаков могут выступать значения слов, предшествующих и следующих за целевым. В работе [Азарова и др., 2008] представлены две процедуры автоматического разрешения неоднозначности «на базе значений морфологических категорий для группы частотных существительных, имеющих несколько вхождений в синсеты RussNet».

На смену признаковым моделям текста пришли сжатые векторные представления слов (**эмбеддинги**), получаемые в ходе обучения языковых моделей на основе нейронных сетей на неразмеченном корпусе. Преимуществом распределенных представлений слов является то, что они согласовываются с дистрибутивной гипотезой [Harris, 1954], которая заключается в следующем: слова, которые встречаются в похожих контекстах, как правило имеют схожие значения. Так как все вектора слов расположены в едином векторном пространстве, с помощью них можно вычислять меру семантической близости между словами. Обычно для этого используется косинусная мера: чем меньше угол между двумя векторами и, соответственно, больше косинус, тем больше слова похожи по смыслу. Выделяют различные типы моделей, с помощью которых можно получать векторные представления слов, например, word2vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], fastText [Bojanowski et al., 2016].

Статические вектора слов используются в работе [Taghipour, Ng, 2015b]. В статье [Rothe, Schütze, 2015] представлен метод получения сжатых представлений для синсетов и значений многозначных слов с помощью векторных представлений слов. В исследовании [Iacobacci et al., 2016] были протестированы

различные стратегии использования векторов слов в моделях разрешения неоднозначности, например, конкатенация и среднее арифметическое всех векторов слов предложения, а также взвешивание векторов слов пропорционально их расстоянию до ключевого слова.

В последнее время контекстуализированные вектора слов, такие как, например, context2vec [Melamud et al., 2016], ELMo [Peters et al., 2018], BERT [Devlin et al., 2019], применяются в работах по автоматическому разрешению неоднозначности всё чаще по сравнению со статическими представлениями слов. Их преимущество заключается в том, что вектор для слова не фиксирован, а зависит от контекста, в котором слово употреблено, поэтому такие представления слов лучше отражают полисемию. Некоторые архитектуры языковых моделей были разработаны специально для того, чтобы с их помощью получать контекстуализированные представления слов, подходящие для задачи разрешения лексической неоднозначности: [Bevilacqua, Navigli, 2019; Levine et al., 2019; Loureiro, Jorge 2019; Scarlini et al., 2020b].

1.4. Оценка методов

Сейчас оценка и сравнение моделей автоматического разрешения лексической многозначности производится в основном в рамках технологических соревнований (например, SemEval) или с помощью таких проектов, как XL-WSD, где таблица лучших моделей по качеству предсказаний постоянно обновляется с появлением новых подходов к разрешению неоднозначности. В статье [Raganato et al., 2017b] была предложена унифицированная система для тестирования систем снятия многозначности, которая включает в себя стандартизированные обучающие и тестовые корпуса в едином формате: аннотация всех наборов данных была приведена к формату, используемому в тезаурусе WordNet 3.0; все текстовые коллекции были предварительно обработаны с помощью одной и той же последовательности шагов.

Для определения качества разрешения многозначности обычно используются три параметра: точность, полнота и F-мера. **Полнота** (*recall*) для того или иного целевого значения многозначного слова – это отношение числа слов, для которых значение было выбрано правильно, к общему количеству слов с этим значением. **Точность** (*precision*) показывает, какая доля слов, которым классификатор приписал то или иное значение, действительно имеют это значение. **F-мера (F1)** – это гармоническое среднее между точностью и полнотой.

$$\mathrm{F} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

Помимо этих метрик также может использоваться *accuracy*, показывающая долю правильных ответов. Однако эта метрика менее информативная и не может использоваться, если в данных наблюдается дисбаланс классов.

Чаще всего качество алгоритмов по верхней границе сравнивается с самыми передовыми методами (*state-of-the-art* методами), которые на данный момент показывают наилучшее качество разрешения неоднозначности. Например, state-of-the-art решениями считаются системы, основанные на методах машинного обучения с учителем [Barba et al., 2021] и [Bevilacqua, Navigli, 2020].

1.5. Методы автоматизации построения обучающей коллекции

Экспертная разметка текстов для задачи разрешения неоднозначности требует много времени и трудозатрат. Большие корпуса, имеющие вручную сделанную разметку значений слов, сейчас доступны в основном для английского языка. Ограниченный доступ к данным с семантической аннотацией не позволяет масштабировать современные передовые модели разрешения неоднозначности на другие языки.

В данном разделе приведен обзор различных автоматических и полуавтоматических методов, направленных на решение проблемы нехватки размеченных данных в области автоматического разрешения лексической неоднозначности. Описываемые подходы направлены на генерацию обучающих

коллекций, а также пополнение и увеличение покрытия уже существующих коллекций. Все методы были условны разделены на группы, соответствующие основному принципу или источнику знаний, который в них используется. В основу данного раздела положена статья [Большина, 2022a].

1.5.1. Метод однозначных родственных слов

Существует большое число методов, основанных на разного рода заменах, которые не требуют экспертных ресурсов для разметки. Самым популярным считается подход, использующий информацию об однозначных родственных словах [Leacock et al., 1998].

Однозначные родственные слова – это те слова или словосочетания, которые связаны с целевым многозначным словом каким-либо отношением в семантическом графе и имеют при этом только одно значение, т.е. входят исключительно в один синсет. В качестве источника таких «родственников» обычно используются тезаурусы типа WordNet. **Расстояние** между значениями в семантической сети определяется числом отношений (ребер), соединяющих эти значения. Методы, 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. Во всех работах синонимы обязательно включаются в число однозначных «родственников» [Mihalcea, Moldovan, 1999]. Иногда помимо них включают гипонимы и гиперонимы [Przybyła, 2017] или, например, меронимы и холонимы [Seo et al., 2004].

В статье [Agirre, De Lacalle, 2004] примеры для всех неоднозначных существительных в WordNet собирались с помощью однозначных родственных слов, которые выступали в качестве запроса в Google. В обучающую коллекцию

добавлялись не документы целиком, а только сниппеты[26]. В работе в качестве однозначных «родственников» могли выступать следующие слова: гиперонимы, прямые и непрямые гипонимы, а также когипонимы. В исследовании [Agirre, Martinez, 2004] каждому однозначному родственному слову приписывался различный вес в зависимости от расстояния до многозначного слова. Данные коэффициенты затем использовались для того, чтобы определять порядок добавления примеров в обучающую коллекцию.

Система построения обучающего корпуса, описанная в [Martínez et al., 2006], была направлена на преодоление недостатков предыдущей работы. Во-первых, если у многозначного слова отсутствовал близко расположенный «родственник», то приходилось ориентироваться на более удаленный, чье значение меньше похоже на значение ключевого слова. Во-вторых, не все примеры, извлекаемые из сети Интернет с помощью однозначных родственных слов, одинаково хорошо подходят для репрезентации значения многозначного слова. Таким образом, в данном исследовании в качестве запроса в поисковый движок подавался контекст с однозначным родственным словом, которое заменило в нем целевое многозначное слово. В описываемом подходе в обучающий корпус включались только те примеры из Интернета, которые имеют высокую степень близости с контекстами употребления однозначных «родственников» на основании заранее заданных метрик. При ранжировании извлеченных из интернета примеров учитывались такие признаки, как количество токенов в исходном запросе, тип отношения, связывающий однозначное родственное слово и многозначное слово, расстояние от «родственника» до многозначного слова и количество найденных страниц по запросу. Интересно отметить, что в качестве родственных слов в статье также использовались многозначные существительные, так как авторы статьи предполагали, что в схожих контекстах такие «родственники» будут выступать в том, значении, которое совпадает у них с ключевым многозначным словом.

[26] Сниппет – это часть страницы сайта, которая выводится в поисковой выдаче.

Метод построения обучающей коллекции, описанный в статье [Martínez et al., 2008], по основным шагам похож на подходы из предыдущих работ: однозначные родственные слова подавались в Google в качестве запроса, затем в полученных сниппетах «родственник» заменялся на ключевое многозначное слово. Помимо этого, для каждого значения многозначного слова необходимое количество примеров набиралось в порядке приоритета родственных слов: сначала извлекались все примеры с синонимами, затем с прямыми гипонимами, после них с прямыми гиперонимами и удаленно расположенными гипонимами и, наконец, с когипонимами. Также в данной работе исследовался вопрос о том, какое должно быть количественное распределение примеров по значениям многозначного слова. Для того, чтобы оценить влияние этого параметра, авторы статьи исследовали четыре стратегии выбора количества примеров.

Метод однозначных родственных слов также использовался для создания автоматически размеченного обучающего корпуса для разрешения лексической неоднозначности в области биомедицины [Preiss, Stevenson, 2013]. Для этого применялся специализированный тезаурус The Unified Medical Language System (UMLS) [Humphreys et al., 1998].

Преимущество методов данной группы заключается в том, что они достаточно просты в реализации, а семантические сети, на которые они опираются при поиске однозначных родственных слов, доступны для большого числа языков (например, BabelNet).

1.5.2. Использование параллельных корпусов

Параллельные корпуса являются полезным источником, с помощью которого можно автоматически создавать семантически размеченные обучающие коллекции, в том числе для нескольких языков сразу. Впервые идея о том, что параллельные корпуса можно использовать для преодоления нехватки данных с разметкой значений, появилась в работе [Resnik, 1997].

В основе подхода, описанного в [Ng et al., 2003], и ряда других, последовавших за ним, лежит гипотеза о том, что разным значениями многозначного слова в английском будут соответствовать разные слова на китайском языке. Предложенный метод базировался на соответствии значений английских слов из инвентаря WordNet китайским словам. Из параллельного китайско-английского корпуса извлекались предложения со словами на китайском, которые были соотнесены с каким-либо значением из инвентаря WordNet. Соответствующее предложение на английском размечалось и добавлялось в обучающую коллекцию. Если у каких-то значений многозначного слова был один и тот же перевод на китайский, они группировались вместе. Метод, предложенный в [Chan, Ng, 2005], схож с уже описанным подходом, однако его отличие от предыдущего заключается в том, что здесь значения слов с одинаковым переводом не группировались. Если один и тот же перевод на китайский соответствовал нескольким значениям в WordNet, то перевод относился только к тому значению, чей порядковый номер в лексической базе был выше. В статье приведен такой пример: у слова *channel* в WordNet есть 7 значений, один и тот же перевод был присвоен значениям под номерами 1 и 7, в таком случае все английские слова, которые в параллельном корпусе соответствовали этому переводу, размечались значением под номером 1.

В исследовании [Wang, Carroll, 2005] по аналогии с предыдущими работами каждому значению многозначного слова в соответствие ставилось слово на китайском. Затем каждый такой переводной эквивалент подавался в качестве запроса либо в корпус китайского языка, либо в поисковую систему. После этого с помощью китайско-английского словаря эти размеченные тексты переводились на английский язык и добавлялись в обучающую коллекцию. Работа [Wang, Martínez, 2006] отличается от предыдущей работы тем, что здесь для перевода обучающих контекстов на английский язык использовался алгоритм машинного перевода.

Для построения размеченного корпуса OMSTI (*One Million Sense-Tagged Instances*) в работе [Taghipour, Ng, 2015a] использовался принцип, разработанный в статье [Chan, Ng, 2005]. За основу была взята англо-китайская часть корпуса MultiUN [Eisele, Chen, 2010], состоящего из протоколов заседаний ООН. Также для того, чтобы улучшить покрытие корпуса, авторы статьи добавили в полученную обучающую выборку примеры из корпуса SemCor и DSO.

Как уже говорилось ранее, преимуществом использования параллельных корпусов для автоматической генерации обучающих коллекций, является то, что с их помощью можно создавать мультиязычные аннотированные корпуса. Авторы статьи [Otegi et al., 2016] описывают методику сбора обучающих коллекций с разметкой значений для шести языков (баскского, английского, испанского, болгарского, чешского и португальского) на основе параллельных корпусов Europarl [Koehn, 2005] и QTLeap [Agirre et al., 2015], содержащих протоколы заседаний Европейского парламента и инструкции по устранению неполадок в программном и аппаратном обеспечении, соответственно. Для предсказания значений для каждого языка (или группы языков) в данной работе использовалась своя модель разрешения неоднозначности: например, для баскского, английского и испанского применялся метод UKB [Agirre, Soroa, 2009], базирующийся на графах.

В исследовании [Delli Bovi et al., 2017] за основу брался корпус Europarl и мультиязычный алгоритм снятия многозначности Babelfy [Moro et al., 2014], использующий векторные представления сущностей NASARI [Camacho-Collados et al., 2016]. На первом этапе из корпуса извлекалось предложение и все его возможные переводы, и осуществлялось разрешение неоднозначности для всех собранных контекстов. Авторы отмечали, что система Babelfy имеет тенденцию предсказывать наиболее частотное значение многозначного слова, поэтому векторные представления NASARI использовались в статье для того, чтобы исключить из выборки примеры с низким весом, размеченные на предыдущем этапе, или же исправить значения таких примеров.

В работе [Hauer et al., 2021] представлены три различных способа создания обучающих корпусов. Подход LABELPROP базировался на уже существующих корпусах с семантической разметкой. Размеченный корпус переводился на целевой язык с помощью алгоритма машинного перевода, а значения многозначных слов проецировались с аннотированного текста на переведенный. Сначала все автоматически приписанные значения проверялись с помощью базы знаний BabelNet: если в синсете, приписанном на предыдущем шаге, не было рассматриваемого ключевого многозначного слова, то такие аннотации не включались в финальную обучающую коллекцию. Также аннотации верифицировались с помощью системы разрешения неоднозначности: если спроецированное значение совпадало со значением, предсказанным алгоритмом, то такие примеры оставались в корпусе.

Для того, чтобы исключить зависимость алгоритма генерации обучающей текстовой коллекции от размеченных данных, авторы статьи предложили еще один метод – LABELSYNC. В рамках этого подхода обе части параллельного текста аннотировались с помощью системы разрешения неоднозначности UKB [Agirre et al., 2014], расширенной с помощью SyntagNet [Maru et al., 2019]. Затем предсказанные метки значений корректировались с помощью метода SOFTCONSTRAINT [Luan et al., 2020], а также проверялось, совпадают ли метки для соответствующих друг другу ключевых слов в текстах обоих языков.

Подход LABELGEN отличается от предыдущего тем, что в нем одна часть параллельного корпуса должна была быть на английском языке. За счет большего объема размеченных данных качество предсказаний значений многозначных слов для английского языка выше, чем для других языков. После автоматической разметки английской части корпуса поставленные метки проецировались на вторую часть параллельного корпуса по аналогии с тем, как это делалось в методе LABELPROP. Также выполнялись две процедуры фильтрации полученных значений на основании весов предсказаний и с помощью базы знаний BabelNet.

Основным ограничением на применимость методов данной группы является то, что далеко не для всех языков существуют параллельные корпуса. Кроме того, они могут не покрывать какие-то специализированные семантические области, интересующие исследователей, например, биомедицинскую тематику. И, наконец, в параллельных корпусах могут встречаться не все значения многозначных слов, а какие-то слова могут не употребляться в коллекциях вовсе.

1.5.3. Методы, использующие базы знаний

Существует ряд работ, в которых для автоматического построения размеченной обучающей коллекции используются семантические сети (например, WordNet). В работе [Pasini, Navigli, 2017] описан метод Train-O-Matic, предназначенный для сбора обучающей коллекции для разных языков с помощью неразмеченного корпуса и мультиязычного семантического графа BabelNet. Этот подход состоит из трех основных шагов. Для каждого значения многозначного слова на основе сети BabelNet подсчитывался персонализированный PageRank вектор [Haveliwala et al., 2003] (вариант оригинального алгоритма PageRank [Brin, Page, 1998]). Затем для каждого предложения, содержащего ключевое слово, оценивалась вероятность встретить то или иное значение данного слова в этом контексте. Таким образом, рассчитывалась вероятность $P(s|\sigma,w)$, где s – это значение ключевого слова, σ – предложение с многозначным словом w. Для получения значения вероятности использовались вектора, подсчитанные на первом шаге, и значение $P(s|w)$. Ключевому слову в предложении приписывалось наиболее вероятное из всех его значений. В финальную обучающую выборку для каждого значения многозначного слова отбирались только такие предложения, в которых с наибольшей вероятностью можно встретить ключевое слово в этом значении. Для извлечения предложений для разметки использовались два корпуса: Википедия и параллельный корпус ООН [Ziemski et al., 2016]. Благодаря тому, что в работе применялся мультиязычный семантический граф BabelNet, авторы статьи также показали применимость их алгоритма для генерации данных

на других языках (итальянском и испанском). В качестве неразмеченного корпуса в этом случае выступали соответствующие версии Википедии. Преимущества данного алгоритма заключаются в том, что он не требует размеченных данных и позволяет выбирать количество извлекаемых примеров для каждого значения многозначного слова.

Общедоступные интернет-ресурсы, например, Википедия и Викисловари, также широко применяются для автоматического создания корпусов с семантической разметкой. Метод, описанный в [Henrich et al., 2012], использовал соответствие семантической сети GermaNet немецкой версии Викисловаря для создания размеченной обучающей коллекции. В примерах употреблений слов в Викисловарях также встречались ссылки на внешние ресурсы, с помощью которых набирался дополнительный материал для корпуса. Похожий метод для арабского языка был описан в работе [Saif el al., 2018]. В данном исследовании для автоматической разметки корпуса использовалась Википедия и WordNet для арабского языка. Метод SEW (*Semantically Enriched Wikipedia*) [Raganato et al., 2016] вычислял эвристики, необходимые для разметки многозначных слов, с помощью семантического графа BabelNet, связей между страницами Википедии и их категориями. В работе [Camacho-Collados et al., 2019] описывалась процедура создания мультиязычного корпуса SenseDefs, который был составлен из аннотированных словарных определений многозначных слов, взятых из различных ресурсов: Викисловари, WordNet, Wikidata, Википедия и OmegaWiki.

Метод OneSeC, представленный в статье [Scarlini et al., 2019], позволяет автоматически создавать обучающие коллекции для разных языков. В рамках этого исследования Википедия использовалась и для извлечения обучающих примеров, и для их разметки. Предложенный подход основывается на предположении «одно значение для одной категории Википедии» (*One Sense per Wikipedia Category*), подразумевающем, что все употребления неоднозначного слова в рамках страниц Википедии, принадлежащих одной и той же категории, носят одно и то же значение. Метод, описанный в данной статье, позволяет

каждой категории Википедии приписать какое-либо значение многозначного слова. Таким образом, обучающая коллекция собиралась из предложений, содержащих ключевое слово и входящих в размеченные алгоритмом категории Википедии.

Плюсом методов данной группы является то, что ресурсы, на которые они опираются, имеются для достаточного большого числа языков и могут покрыть много целевых значений многозначных слов. Однако для того, чтобы с помощью этих источников генерировать хорошую разметку, требуется разработка довольно сложных эвристик и алгоритмов.

1.5.4. Алгоритм распространения меток, бутстрэппинг и активное обучение

Одной из первых работ в области компьютерной лингвистики, в которой было употреблено понятие бутстрэппинг (*bootstrapping*), была работа [Yarowsky, 1995]. Алгоритм состоит из следующих шагов: сначала классификатор обучается на небольшом объеме размеченных данных, затем с помощью уже обученного классификатора размечаются неаннотированные примеры. Далее он обучается на уже расширенной размеченной выборке, и процедура итеративно повторяется до тех пор, пока не будет размечена вся выборка. В статьях [Mihalcea, 2002a; Mihalcea, 2004; Pham et al., 2005] бутстрэппинг использовался для создания размеченного обучающего корпуса. В исследовании [Khapra et al., 2011] бутстрэппинг применялся для построения обучающих корпусов сразу для двух языков, в которых наблюдался недостаток размеченных данных.

Текстовые данные для задачи разрешения неоднозначности можно представлять в виде графа. В узлах такого графа содержатся аннотированные примеры из корпуса. Ребра, соединяющие узлы, имеют вес, определенный выбранной метрикой: например, если в качестве метрики используется косинусная мера близости двух слов, то вес ребра будет больше для слов, схожих по своему значению. В основе алгоритма распространения меток (*label*

propagation) лежит представление данных в виде графа, и его идея состоит в предсказании метки неразмеченного примера в графе, исходя из информации о классах, содержащихся в окружающих его узлах, и весах, приписанных им.

Подход, описанный в [Yuan et al., 2016], был основан на языковой модели с архитектурой LSTM, которая обучалась предсказывать одно пропущенное слово в предложении на большом неразмеченном корпусе. Предсказание, которое сеть выдавала для пропуска, использовалось как вектор контекста. В семантическом графе вершины, содержащие размеченные примеры, соединялись ребрами, вес которых был равен косинусной мере близости соответствующих контекстных векторов, вычисленных с помощью предварительно обученной в рамках данного исследования языковой модели. Авторы применили метод распространения меток, чтобы присвоить метку новым неаннотированным примерам на основании примеров, уже имеющих разметку. Авторы статьи [Le et al., 2018] воспроизводили метод, представленный в [Yuan et al., 2016], при этом использовали корпус меньшего размера.

Алгоритм MuLaN (Multilingual Label propagatioN) [Barba et al., 2020] был разработан для создания мультиязычных размеченных наборов данных и также основан на методе распространения меток. Авторы статьи использовали контекстуализированные представления слов, информацию из баз знаний и проекцию смысловых тегов с языка с большим количеством лингвистических ресурсов (*high-resource language*) на язык с недостаточным количеством размеченных текстовых данных (*low-resource language*). Для работы алгоритма требовался размеченный корпус на каком-либо языке и неразмеченный корпус на том языке, для которого необходимо было получить аннотированную обучающую выборку, при этом данный подход не задействовал параллельные корпуса. Описываемый метод – мультиязычный, поэтому в качестве инвентаря значений использовались значения из сети BabelNet. На первом шаге работы описываемого метода с помощью мультиязычной языковой модели Multilingual BERT (mBERT) [Devlin et al., 2019] для каждого предложения в обоих корпусах извлекались

контекстуализированные векторные представления. Затем для каждого предложения из размеченной выборки с помощью косинусной меры близости вычислялись 1000 ближайших соседей среди примеров неразмеченной выборки. Для того, чтобы при таком способе приписывания меток в генерируемую обучающую коллекцию не попадали нерелевантные примеры, авторы статьи исключали такие контексты на целевом языке, в которых многозначное слово не входило в синсет, спроецированный из размеченной выборки. Отобранные таким способом примеры попадали в список контекстов-кандидатов для финальной обучающей выборки. Затем эти предложения-кандидаты проходили процедуру обратной проверки: если исходное размеченное предложение попадало в список 1000 ближайших соседей того или иного неразмеченного предложения на целевом языке, то такой кандидат оставался в списке. В конце наиболее подходящие примеры-кандидаты отбирались с помощью маргинализированной косинусной меры близости (*marginalized cosine similarity, mcos*) [Artetxe, Schwenk, 2019], вычисленной для кандидата и размеченного предложения, к которому он относится. Контексты, отфильтрованные таким образом, получали метку того аннотированного примера, с которым они соотносились, и добавлялись в обучающую коллекцию.

Помимо уже описанных методов для решения проблемы нехватки размеченных данных используется такой подход, как активное обучение (*active learning*) [Fujii et al., 1998; Mihalcea, Chklovski, 2003; Chen et al., 2013; Alagić, Šnajder, 2015]. Алгоритм на основании определенных метрик выбирает наиболее информативные примеры из числа неразмеченных, и затем они размечаются экспертом. Эти аннотированные данные используются для обучения (или дообучения) модели, и после этого обновленная модель снова применяется для отбора наиболее ценных примеров из числа неразмеченных. Благодаря такому итеративному обучению модели, ее можно улучшить с помощью меньшего числа обучающих данных. Несмотря на то, что этот метод требует участия эксперта для разметки, он значительно сокращает её трудоёмкость.

Недостаток бутстрэппинга и алгоритма распространения меток состоит в том, что для их работы все равно необходимо некоторое количество размеченных данных. Минусом активного обучения является то, что этот метод подразумевает участие разметчика в генерации обучающих данных.

1.5.5. Подходы, направленные на увеличение покрытия значений и слов

Даже имеющиеся размеченные лингвистические ресурсы включают в себя не все многозначные слова из инвентаря значений и покрывают не все возможные значения многозначных слов. Например, один из самых больших аннотированных корпусов Semcor «содержит примерно 16% из всего инвентаря значений тезауруса WordNet» [Vial et al., 2018: 3]. Помимо этого, какие-то слова и их значения могут быть недостаточно представлены в корпусе. Как отмечалось в работе [Chen et al., 2021: 1774], «84% размеченных слов в корпусе SemCor имеют меньше 10 примеров». Ввиду этого, разрабатывается большое число алгоритмов, направленных на расширение существующих размеченных корпусов.

В исследовании [Vial et al., 2019] описывается метод компрессии словаря значений (*Sense Vocabulary Reduction*) для улучшения обобщающей способности моделей разрешения неоднозначности, обученных методом машинного обучения с учителем. Авторы статьи группировали вместе значения слов, основываясь на предположении о том, что то или иное значение многозначного слова и значения слов, ближайших к нему в семантической сети (например, гипонимы, гиперонимы и т.д.), относятся к какой-либо общей идее или концепту. В рамках исследования было сокращено количество различных значений в сети WordNet, и были оставлены только те, которые важны для различения значений слов. Без каких-либо дополнительных данных этот подход сокращает словарь меток, который используется для обучения модели, улучшает способность модели к обобщению и повышает качество предсказания слов, которых нет в обучающей выборке.

Работа [Loureiro, Jorge, 2019] посвящена созданию векторных представлений для значений слов, с помощью которых можно добиться полного покрытия значений из WordNet. Предложенный в статье метод с помощью связей между сущностями в семантической сети и уже вычисленными векторами значений генерировал представления для значений, которых нет в корпусе. Опираясь на данный метод вычисления векторов для отсутствующих в корпусе значений, в исследовании [Loureiro, Camacho-Collados, 2020] с помощью разметки значений однозначных слов в неаннотированном корпусе увеличивалось покрытие отсутствующих в размеченном корпусе значений.

Система EWISE (*Extended WSD Incorporating Sense Embeddings*) для решения задачи разрешения неоднозначности, описанная в работе [Kumar et al., 2019], опиралась на комбинацию размеченных обучающих корпусов, словарные определения и данные из лексических баз знаний. Данный метод в качестве целевых переменных в задаче классификации значений использовал векторные представления слов вместо привычных дискретных меток значений слов. Эта особенность позволяла подходу успешно предсказывать значения тех слов, которых не было в обучающей выборке. В исследовании [Bevilacqua, Navigli, 2020] демонстрировался алгоритм EWISER (*Enhanced WSD Integrating Synset Embeddings and Relations*), основанный на предыдущем методе EWISE. Для того, чтобы точнее предсказывать значения слов, которых не было в выборке, в нейронную архитектуру EWISER была добавлена информация из графа WordNet о связях входящих в него сущностей.

Для преодоления проблем, связанных с недостаточно полным покрытием семантических корпусов, в статье [Holla et al., 2020] предлагалось решать задачу разрешения неоднозначности с помощью метаобучения, которое неформально называют «учусь учиться» (*learning to learn*). Благодаря тому, что алгоритмы метаобучения обучаются решать множество связанных по смыслу задач, они могут легко адаптироваться к новым задачам даже при условии наличия для них малого количества обучающих данных. Авторы статьи применили три модели

метаобучения для решения задачи разрешения неоднозначности. Результаты их экспериментов показали, что данные системы можно успешно применять для предсказания значений слов в ситуации, когда обучающих примеров очень мало. Непараметрическая модель MetricWSD была предложена в работе [Chen et al., 2021] с целью уменьшить влияние дисбаланса классов в обучающей выборке на качество разрешения неоднозначности. Выучиваясь подсчитывать расстояния между значениями слова, алгоритм MetricWSD «переносил знания», а именно вычисленное метрическое пространство, от более частотных слов к менее частотным.

1.5.6. Результаты, достигаемые на автоматически порождаемых наборах данных

Для того, чтобы понять, могут ли автоматически размеченные коллекции заменить вручную аннотированные корпуса, были проанализированы имеющиеся метрики качества классификации моделей, обученных на данных с автоматической разметкой. Модели снятия многозначности, обученные на комбинации данных, аннотированных экспертами и автоматическими методами, могут показывать качество, сравнимое с моделями, обученными только на вручную размеченных текстах, а также могут и превосходить их. Результаты, представленные в единой системе оценки моделей [Raganato et al., 2017b], показывают, что обучение модели IMS [Zhong, Ng, 2010] на комбинации корпусов SemCor и OMSTI позволяет улучшить качество разрешения неоднозначности на наборе данных Senseval-2 [Edmonds, Cotton, 2001]. Системы, использующие только автоматически размеченные данные, могут достигать результатов, сопоставимых с теми, что были получены при обучении на данных, аннотированных экспертами, а также могут иметь более высокие оценки качества. Модель IMS, обученная на корпусе Train-O-Matic, на наборе Senseval-2 имеет значение метрики F1 всего лишь на 0.4% ниже, чем у модели, при обучении которой использовался корпус SemCor. Модель IMS, обученная на корпусе SEW,

достигает более высокого значения F1 на англоязычной части набора данных SemEval-2013 task 12 [Navigli et al., 2013], чем та же самая модель, для обучения которой применялся корпус SemCor.

В приложении 1 к диссертации приведена таблица, демонстрирующая качество разрешения неоднозначности методов, описанных в данном разделе, на различных тестовых наборах. В таблице приведены лишь те работы, в которых производилась какая-либо количественная валидация моделей, обученных на созданных в рамках исследования обучающих коллекциях.

1.6. История развития методов автоматического разрешения неоднозначности

Как уже говорилось ранее, задача разрешения неоднозначности возникла в конце 40-х гг. 20-го века как одна из побочных задач машинного перевода. В период с 50-х по 80-е гг. перед исследователями стояла большая проблема отсутствия машиночитаемых словарей и баз знаний, что, безусловно, тормозило развитие области. В 90-е гг. в связи с появлением обширных баз знаний типа WordNet и CyC [Lenat et al., 1985], а также Брауновского корпуса и корпуса Penn TreeBank[27] получили развитие различные статистические методы разрешения неоднозначности. В 2000-е годы наряду с методами обучения с учителем стали распространяться системы разрешения неоднозначности на основе методов обучения без учителя и с частичным привлечением учителя, а также методов, основанных на графах. В настоящее время одним из наиболее эффективных подходов к обучению моделей продолжает оставаться машинное обучение с учителем, особенно большой популярностью пользуются нейронные сети. Именно ввиду того, что все наиболее продвинутые решения по разрешению лексической неоднозначности основаны на методах обучения с учителем, сейчас так остро ощущается нехватка размеченных данных. В этом разделе будут

[27] https://catalog.ldc.upenn.edu/LDC99T42

подробно рассмотрены различные типы систем разрешения лексической неоднозначности.

1.6.1. Методы, основанные на знаниях

Самые первые системы разрешения неоднозначности были основаны на информации из лексических ресурсов, и одним из классических примеров является алгоритм Леска [Lesk, 1986]. Он состоит из следующих этапов: из толкового словаря извлекаются определения всех значений слов, которые есть в рассматриваемом тексте. Далее выявляется пересечение полученных толкований между собой, и многозначному слову приписывается такое значение, определение которого максимально пересекается с определениями слов в окружающем контексте. Одним из главных недостатков данного метода является его зависимость от формулировок словарных определений, а также отсутствие толкований для новых слов. Существует большое число работ, в которых продемонстрированы модификации оригинального алгоритма. Например, [Kilgarriff, Rosenzweig, 2000] предложили упрощенный алгоритм Леска, который определяет значение многозначного слова по пересечению толкований значений неоднозначных слов со соседними словами. В работе [Vasilescu et al., 2004] был проведен сравнительный анализ вариаций алгоритма Леска и оригинального метода. Исследование, описанное в [Basile et al., 2014], также дополняет алгоритм Леска. Подход основан на вычислении семантической близости слов, входящих в толкования и контекст, с помощью косинусного расстояния между представлениями слов в векторном пространстве.

Сейчас в основе методов, основанных на знаниях, лежат семантические графы, например, WordNet или BabelNet. В работе [Agirre et al., 2014] описывается подход UKB, основанный на случайных блужданиях (*random walks*) в семантических сетях и использовании персонализированного PageRank. Суть метода UKB заключалась в том, что исследователи дополняли семантический граф WordNet словами из контекста с многозначным словом и соединяли их со

всеми синсетами, в которых они встречались. Персонализированный PageRank оценивал важность узла в графе на основе связанных с ним вершин, таким образом, синсет, получивший наибольший вес, выбирался в качестве значения слова.

Также можно отметить подход Babelfy, представленный в [Moro et al., 2014]. Этот метод использовал семантический граф BabelNet и применялся для разрешения неоднозначности для всех языков и даже для именованных сущностей, представленных в данном лексико-семантическом ресурсе. Сначала для каждой сущности в графе создавалась семантическая сигнатура, т.е. набор тесно связанных с ней узлов в семантической сети. Затем формировалась семантическая интерпретация всего текста с помощью графов: возможные значения многозначных слов соединялись с ранее вычисленными семантическими сигнатурами. Основная идея данного подхода заключалась в том, что наиболее подходящие значения для каждого многозначного слова принадлежали самой плотной области построенного графа.

В исследовании [Chaplot, Salakhutdinov, 2018] система разрешения многозначности основывалась на латентном размещении Дирихле (LDA), которое обычно применяется в задаче тематического моделирования (*topic modeling*). В данной работе каждый текст представляется как дискретное распределение на множестве синсетов, таким образом, в качестве контекста для снятия неоднозначности с многозначного слова выступает целый текст. Исследование [Tripodi, Navigli, 2019] описывает модель разрешения неоднозначности Word Sense Disambiguation Games (WSDG), впервые представленную в работе [Tripodi, Pelillo, 2017]. Эта система была основана на принципах теории графов, которые используются для моделирования геометрии данных, и на теории игр, применяемой для реализации алгоритма разрешения лексической неоднозначности. В данном подходе слова выступали в роли игроков в некооперативной игре, а их значения – это стратегия, которую игроки могли выбирать. Все слова представлялись в виде графа, ребра которого отражали связи

между ними и несли информацию о сходстве слов. Отличие этого исследования от оригинальной статьи в том, что в работе [Tripodi, Navigli, 2019] использовались передовые модели векторных представления слов и их значений, например, ELMo, BERT, NASARI и т.п. Статья [Scozzafava et al., 2020] посвящена мультиязычной системе разрешения неоднозначности SyntagRank. В ее основе лежит применение алгоритма персонализированного PageRank к лексической базе знаний, состоящей из WordNet, корпуса WNGT и базы данных с коллокациями SyntagNet.

1.6.2. Методы машинного обучения с учителем

Для обучения систем разрешения лексической многозначности методами машинного обучения с учителем требуются семантически размеченные корпуса, где каждому многозначному слову в том или ином контексте приписано его значение. Наиболее простым способом разрешения неоднозначности является подход, который называется «Наиболее частотное значение» (*Most Frequent Sense – MFS*), согласно которому выбирается то значение многозначного слова, которое чаще всего встречается в обучающей выборке. Данный метод обычно используется в качестве базового решения (*baseline*, бэйслайна)[28], с которым сравниваются другие более сложные системы.

Обзор можно начать с базовых методов, например, наивного байесовского классификатора – это простой линейный классификатор, основанный на теореме Байеса. Байесовские методы имеют уже долгую историю, и сейчас они также пользуются популярностью у исследователей [Singh et al., 2014; Gosal, 2015; Gopal, Haroon, 2016].

Сейчас в работах по разрешению лексической неоднозначности нередко применяют метрический алгоритм классификации – метод k-ближайших соседей (*k-Nearest Neighbours, kNN*). Суть этого подхода довольно интуитивна: имеется признаковое пространство, в котором каким-либо образом расположены

[28] Бэйслайн – это наиболее простой алгоритм решения задачи, качество которого используется для сравнения с результатами других более сложных алгоритмов для оценки их производительности.

размеченные объекты, и, когда на вход классификатору приходит новый объект, он приписывает ему ту метку, которая чаще всего встречается среди k его соседей. Этот метод можно встретить в таких исследованиях, как [Rezapour et al., 2011], где признакам объектов были добавлены определенные веса; [Pandit, Naskar, 2015], в котором был описан опыт применения данного метода к задаче разрешения многозначности в бенгальском языке. В работе [Melamud et al., 2016] была описана модель context2vec, которая с помощью двунаправленной реккурентной нейронной сети LSTM (*bidirectional LSTM, biLSTM* [Hochreiter, Schmidhuber, 1997]) генерировала контекстные вектора для предложений. Для предсказания значений многозначных слов из тестовой выборки авторы с помощью меры семантической близости сравнивали контекстный вектор тестового примера и контекстные вектора, сгенерированные для предложений из размеченной выборки. Слову приписывалось значение того примера из обучающей выборки, к которому его контекстный вектор оказался ближе. Сами авторы утверждают, что, по сути, это простейшая форма алгоритма kNN, где $k = 1$. В работе [Wiedemann et al., 2019] для разрешения неоднозначности kNN применялся к признаковому пространству контекстуализированных векторов слов FLAIR [Akbik et al., 2018], ELMo [Peters et al., 2018] и BERT [Devlin et al., 2019]. Исследование [Loureiro, Jorge, 2019] посвящено методу генерации сжатых представлений для значений слов. Эксперименты, представленные в статье, показали, что алгоритм kNN, базирующийся на разработанных в рамках данного исследования контекстуализированных представлениях слов, превосходил по качеству алгоритмы разрешения неоднозначности, основанные на нейронных сетях.

Среди линейных моделей, которые используются для автоматического разрешения неоднозначности, можно отметить метод опорных векторов (*SVM, Support Vector Machines*). Принцип работы этого алгоритма, впервые предложенного в [Boser et al., 1992], заключается в нахождении гиперплоскости, которая бы разделила положительные и отрицательные примеры. При этом такая

плоскость должна максимизировать расстояние между самым близким положительным и отрицательным примером в выборке. Подход IMS, описанный в работе [Zhong, Ng, 2010], в основе которого лежал аппарат опорных векторов, до сих пор используется различными исследователями в качестве базового решения задачи разрешения неоднозначности. Этот метод опирался на набор вручную извлеченных признаков: например, слова, окружающие ключевое слово, метки частей речи окружающих слов и т.п.

Логистическая регрессия также используется для разрешения лексической неоднозначности. В статье [Kutuzov, Kuzmenko, 2019] предсказание значений многозначных слов в русском языке осуществлялось с помощью контекстуализированных представлений слов ELMo, которые подавались на вход логистической регрессии.

Несмотря на уже отмеченную проблему недостатка размеченных данных, сейчас самые передовые системы разрешения неоднозначности основаны на нейронных сетях различной архитектуры, требующих для своего обучения большого количества аннотированных данных. В исследовании [Kågebäck, Salomonsson, 2016] для разрешения неоднозначности применялась архитектура biLSTM на основе предварительно обученных представлений слов GloVe. В исследовании [Uslu et al., 2018] для предсказания значений слов использовалась простая нейронная сеть прямого распространения с одним скрытым слоем и предварительно обученная модель word2vec. Стоит отметить, что данные модели рассчитывали вероятностное распределение по значениям только для одного конкретного слова в последовательности.

В исследовании [Raganato et al., 2017a] авторы рассматривают задачу разрешения лексической неоднозначности как задачу классификации каждого токена в последовательности. Таким образом, нейронная сеть использовалась для перевода последовательности токенов в предложении в последовательность их значений. В этой работе наряду с biLSTM-моделями были предложены модели с архитектурой энкодер-декодер (*sequence-to-sequence*) [Sutskever et al., 2014].

Авторы исследования изучали данные типы моделей разрешения неоднозначности с учетом добавления в них различных дополнительных компонентов, например, слоя внимания (*attention layer*). Помимо этого, изучались мультизадачные модели, где задача разрешения неоднозначности решалась совместно с задачами предсказания части речи для каждого токена в последовательности и метки семантической категории, к которой относится синсет («45 семантических категорий, вручную соотнесенных со всеми синсетами в WordNet на основе как синтаксических, так и логических группировок, например, *существительное.местоположение* или *глагол.движение*» [Raganato et al., 2017a: 1161]). Было выявлено, что добавление в модель в качестве еще одной задачи предсказание семантической категории, дает прирост к F-мере на задаче разрешения неоднозначности, а классификация частей речи, напротив, практически не вносит вклад в итоговое качество модели.

На рисунке ниже представлена архитектура нейронной сети с энкодером (модулем, сжимающим входящую последовательность с помощью функции кодирования), декодером (модулем, который предназначен для восстановления входной последовательности с помощью декодирующей функции), двумя скрытыми реккурентными слоями LSTM и слоем внимания:

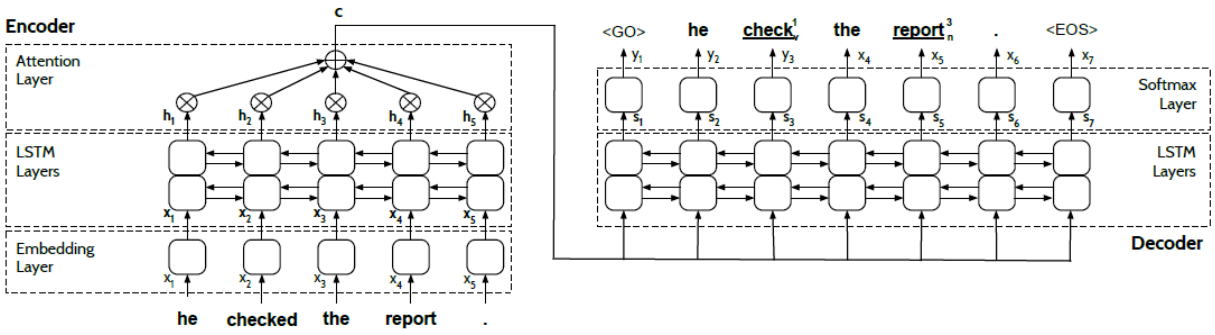

Рисунок 1. Архитектура нейронной сети для разрешения лексической многозначности [Raganato et al., 2017a: 1159].

Сейчас активно развиваются системы разрешения неоднозначности, которые основаны на переносе обучения (*transfer learning*). Принцип их работы заключается в следующем: из предварительно обученной на определенную задачу нейросети извлекаются знания, которые используются для решения уже другой

задачи. Использование уже обученных языковых моделей, например, ELMo и BERT, показывает свою эффективность для решения задачи разрешения неоднозначности. В работе [Hadiwinoto et al., 2019] контекстуализированные представления из языковой модели BERT подавались на вход полносвязной нейронной сети прямого распространения, а в исследовании [Vial et al., 2019] – на вход трансформеру.

Еще одно довольно популярное направление исследований связано с интеграцией внешних источников знаний в системы разрешения неоднозначности. Уже с самых первых работ в области разрешения неоднозначности словарные определения зарекомендовали себя как ценный источник информации [Lesk, 1986], поэтому и в современных системах исследователи используют дефиниции слов для снятия многозначности. Толкования значений могут кодироваться с помощью отдельного модуля в нейронной сети, предназначенной для разрешения неоднозначности, и использоваться для дальнейшего предсказания значения слова в контексте [Luo et al., 2018; Blevins, Zettlemoyer, 2020], могут быть представлены с помощью контекстуализированных векторов и внедрены в kNN-модель [Loureiro, Jorge, 2019], могут применяться в качестве одной из пар при классификации пар предложений для предсказания значения слова в контексте [Huang et al., 2019; Kohli, 2021].

Подход SensEmBERT [Scarlini et al., 2020b] был разработан для генерации векторных представлений для значений слов. Для этого использовались предварительно обученная модель BERT, вектора NASARI для концептов из BabelNet, контексты со словами, наиболее тесно связанными с целевым многозначным словом в сети BabelNet, и толкования значений слов. Метод ARES [Scarlini et al., 2020a] также предназначался для генерации векторных представлений для значений слов с помощью языковой модели BERT, размеченных примеров, имевшихся в корпусе, словарных определений для значений, а также неразмеченных примеров, извлеченных с помощью

специальных эвристик. Чтобы с помощью описанных подходов предсказать значение многозначного слова в том или ином примере, было достаточно получить его контекстуализированное представление и найти ближайший к нему вектор значения по косинусной мере близости. В модели EWISER, описанной в [Bevilacqua, Navigli, 2020], использовались векторные представления синсетов, информация о связях слов в сети WordNet и контекстуализированные представления из предварительно обученной модели BERT. Все эти компоненты применялись в полносвязной нейронной сети.

Таблица 2. Рейтинг моделей разрешения лексической неоднозначности для английского языка (F1-мера).

Набор данных / **Модель**	**Senseval 2**	**Senseval 3**	**SemEval 2007**	**SemEval 2013**	**SemEval 2015**
ESCHER [Barba et al., 2021]	**0.817**	0.778	**0.763**	**0.822**	**0.832**
EWISER [Bevilacqua, Navigli, 2020]	0.789	**0.784**	0.71	0.789	0.793
[Berend, 2020]	0.779	0.778	0.688	0.761	0.775
BEM [Blevins, Zettlemoyer, 2020]	0.794	0.774	0.745	0.797	0.817
ARES [Scarlini et al., 2020]	0.78	0.771	0.71	0.773	**0.832**
GlossBERT [Huang et al., 2019]	0.777	0.752	0.725	0.761	0.804
[Vial et al., 2019]	0.765	0.774	0.695	0.76	0.783

Работа [Berend, 2020] посвящена созданию разреженных контекстуализированных представлений слов (*sparse word representations*) с помощью предварительно обученной модели BERT для задачи разрешения лексической неоднозначности. Авторы статьи утверждают, что подобные вектора слов обладают более хорошей интерпретируемостью [Subramanian et al., 2018], а также могут улучшать качество предсказаний моделей в различных задачах [Berend, 2017]. Исследователи, разработавшие систему ESCHER [Barba et al.,

2021], рассматривали задачу разрешения неоднозначности как задачу извлечения отрезков текста. На вход модели подавалась последовательность, состоящая из предложения с многозначным словом и толкований для всех возможных значений многозначного слова, а вывести она должна была индексы, обозначающие начало и конец того отрезка из последовательности, который соответствовал значению целевого слова в данном контексте.

В таблице 2 приведены значения F1-меры на различных тестовых наборах данных для наиболее эффективных моделей из тех, что были описаны в данном подразделе.

1.6.3. Методы машинного обучения без учителя

Модели разрешения многозначности, основанные на методах машинного обучения с учителем, страдают от недостатка больших размеченных обучающих коллекций, а для методов машинного обучении без учителя (*unsupervised*), подобные ресурсы не требуются. Такие подходы направлены на автоматическое извлечение значений слов, и в англоязычной литературе данную область называют Word Sense Induction (WSI). Задача методов WSI состоит в том, чтобы разделить контексты употребления многозначного слова на ряд классов, в то время как классическая задача разрешения лексической неоднозначности заключается в приписывании слову в контексте конкретной метки значения [Navigli, 2009:26].

Подобные алгоритмы опираются на предположение о том, что слово в одном и том же значении встречается в похожих контекстах, следовательно, все употребления слова можно сгруппировать [Navigli, 2009:26]. В таких подходах не используются инвентари значений, поэтому нахождение соответствия между смысловыми кластерами, которые были выведены, и тем или иным набором значений слов представляется проблематичным. Кроме того, оценку и сравнение результатов работы подобных систем проводить гораздо сложнее.

В работе [Arefyev et al., 2019] было предложено условное деление всех методов автоматического извлечения значений слов на три группы: методы кластеризации контекста/вектора (*Context/Vector Clustering Methods*), методы кластеризации слова/графа (*Word/Graph Clustering Methods*) и методы скрытых переменных (*Latent Variable Methods*). В общих чертах подходы из первой группы работают следующим образом: все контексты с многозначными словами представляются с помощью векторов, а затем список этих контекстных векторов разбивается на группы каким-либо методом кластеризации. В работе [Amrami, Goldberg, 2019] для репрезентации контекстов использовались векторные представления из языковой модели BERT. Идея алгоритма заключалась в том, что если в предложениях, содержащих многозначное слово, оно употребляется в одном и том же значении, то такие контексты также имеют одинаковые контекстуальные замены (*in-context substitutes*). Языковая модель использовалась для вычисления таких замен для слов в контексте, которые затем кластеризовались.

Методы, основанные на кластеризации графов, давно используются в области автоматического извлечения значений слов [Di Marco, Navigli, 2013; Pelevina et al., 2016]. Их суть заключается в том, что сначала строится граф совместной встречаемости слов, в вершине которого располагается неоднозначное слово, а ребра между узлами обозначают, например, степень семантической близости многозначного слова и слов, которые употребляются с ним в одних контекстах. Затем узлы такого графа разбиваются на группы либо с помощью методов кластеризации графов, либо с помощью таких алгоритмов, как HyperLex [Véronis, 2004] или PageRank.

Методы скрытых переменных [Bartunov et al., 2016; Amplayo et al., 2019] рассматривают топики в тексте или значения слов в качестве скрытых переменных, каждый контекст по сути является смесью таких переменных, и, таким образом, каждое употребление многозначного слова в контексте можно отнести к одной из скрытых переменных. При таком рассмотрении задачи

подразумевается, что разные значения слова соотносятся с разными распределениями слов из контекста [Brody, Lapata, 2009: 103].

1.7. Исследования на материале русского языка

На материале русского языка существуют как исследования в области автоматического разрешения лексической неоднозначности, так и работы, посвященные автоматическому извлечению значений слов.

Обзор методов снятия лексической неоднозначности можно начать с серии исследований, посвященных лексико-семантической разметке НКРЯ. Из наиболее ранних статей можно отметить работу [Кобрицов, Ляшевская, 2004], в которой описывается опыт разрешения многозначности для Национального корпуса русского языка с помощью правил сочетаемости семантических классов. В исследовании [Кобрицов и др., 2005] предлагается разрешать неоднозначность поверхностными фильтрами, которые используют информацию из коллокаций: «данные (1) о лемме, (2) о частеречных, (3) словоклассифицирующих и (4) словоизменительных признаках составляющих, (5) об их исходной семантической разметке, а также (6) о некоторых грамматических и лексико-семантических характеристиках ближайшего контекста (например, "родительный падеж" для оборота типа кого-чего-л.)» [Кобрицов и др., 2005: 251]. В работе [Кустова и др., 2005] предлагалась система разрешения неоднозначности для слов в НКРЯ, основанная на вручную разработанных лингвистических правилах. Подобные фильтры задавали контексты и конструкции, в которых употребляются те или иные значения многозначных слов, и могли учитывать семантику, морфологию и лексику. Статья [Рахилина и др., 2006] посвящена принципам семантической разметки в НКРЯ. В работе обсуждалась «оптимизация исходного семантического словаря, а именно, установление иерархии значений» [Рахилина и др., 2006: 449]. Таким образом, в измененном семантическом словаре порядок значений многозначного слова отражал их частотность.

Статья [Кобрицов и др., 2007] описывает систему снятия неоднозначности с глаголов с помощью информации о моделях управления и семантических тегов. В работе [Толдова и др., 2008] изучались семантические фильтры для разрешения неоднозначности глаголов в НКРЯ. Проведенные эксперименты показали, что «грамматические характеристики актантов и сирконстант позволяют существенным образом понизить многозначность глаголов» [Толдова и др., 2008: 527]. Также было выявлено, что семантический класс актантов (одушевленность/неодушевленность, абстрактность/конкретность) «для одних глаголов может быть решающим, а для других – ни о чем не говорить» [Толдова и др., 2008: 526].

В исследовании [Lashevskaja, Mitrofanova, 2009] была разработана система разрешения неоднозначности для существительных, относящихся к материальным объектам и абстрактным понятиям. В работе использовалась статистическая модель, которая учитывала такие признаки, как лексические маркеры, встречающиеся в контексте, маркеры таксономии элементов контекста и морфологические метки контекста. Система автоматического разрешения неоднозначности, описанная в [Lyashevskaya et al., 2011], использовала признаки из разметки НКРЯ: лексические теги (теги лемм), морфологические теги, лексико-семантические теги, а также комбинации этих меток. Семантически и морфологически однозначные контексты, а также контексты, в которых многозначность была снята вручную, использовались в качестве обучающей выборки. Сначала для каждого значения многозначного слова с помощью частотности признаков, обозначенных выше, вычислялось его эталонное представление. Значение многозначного слова из тестовой выборки определялось с помощью косинусной меры близости: отбирались такие вектора значений, вычисленные на предыдущем шаге, которые были наиболее похожими на неоднозначный контекст.

В работе [Лукашевич, Чуйко, 2007] изучались различные параметры выбора значения слова (типы путей тезауруса, размер окна и т. д.) на основе тезауруса

русского языка РуТез. В диссертации [Турдаков, 2009] исследовался подход на основе лексических цепочек к распознаванию значений слов, описанных в Википедии. Исследование [Loukachevitch, Chetviorkin, 2015] было посвящено выявлению самого частого значения слова (*Most Frequent Sense*) с помощью тезауруса РуТез. Как уже говорилось ранее, данная эвристика часто используется как базовое решение в системах разрешения лексической неоднозначности. В исследовании [Lopukhin, Lopukhina, 2016] изучалось разрешение неоднозначности глаголов. Авторы статьи использовали инвентарь значений и примеры к ним из Активного словаря русского языка [Апресян и др., 2014]. Для каждого значения слова извлекались всевозможные примеры: контексты употребления слова, коллокации, синонимы и т.п. Затем для каждого из таких контекстов создавался вектор, который подавался на вход классификатору, предсказывающему значение контекста. Также стоит отметить работу [Kutuzov, Kuzmenko, 2019], которая уже упоминалась в разделе 1.6.2. настоящего диссертационного исследования. Описанная система снятия многозначности в русском языке использовала контекстуализированные представления слов из языковой модели ELMo, обученной на комбинации текстов русскоязычной части Википедии и НКРЯ, которые классифицировались с помощью логистической регрессии.

Помимо систем, предсказывающих значения многозначных слов, на материале русского языка также разрабатываются модели автоматического извлечения значений слов. В исследовании [Lopukhin et al., 2017] сравнивались несколько различных методов для автоматического извлечения значений слов: латентное размещение Дирихле (LDA), кластеризация контекстных векторов (усредненные word2vec вектора слов контекста) с помощью метода k-средних (*k-means*); кластеризация слов наиболее близких к многозначному с помощью метода k-средних; метод AdaGram. В ходе экспериментов подход AdaGram показал один из лучших результатов. Метод хорошо работал для разных частей речи и выделял около 63% значений из словарей. Более того, он позволял детектировать новые и специфичные для предметной области значения, которые

могли быть не включены лингвистические ресурсы с общей лексикой. В статье [Ustalov et al., 2018] описывалась система извлечения значений слов, которая была основана на вычислении семантической близости между данным предложением и синсетом, отражающим смысл целевого неоднозначного слова. Для решения задачи выделения значений слов в работе [Kutuzov, 2018] использовалось векторное представление контекстов, которое вычислялось из предварительно обученных моделей векторных представлений слов. Затем все полученные вектора кластеризовались с помощью алгоритма *Affinity Propagation*. В статье [Arefyev et al., 2018] описывался похожий подход, только здесь контекст, в котором было употреблено целевое слово, представлялся как взвешенное среднее векторов всех слов, которые в него входят. Стоит отметить, что подходы [Kutuzov, 2018] и [Arefyev et al., 2018], которые приведены в данном обзоре, были разработаны в рамках соревнования по извлечению лексических смыслов неоднозначных слов в русском языке RUSSE-2018.

Выводы к главе 1

В главе 1 описываются теоретические аспекты создания систем автоматического разрешения лексической неоднозначности, а также подходы к автоматической генерации размеченных обучающих коллекций для данной задачи; выполняется сопоставление всех автоматических методов сбора и аннотирования обучающих наборов данных с точки зрения качества разрешения неоднозначности, которое можно получить, используя их в обучении; сравниваются имеющиеся в языках наборы семантически размеченных обучающих данных. Помимо этого, в данной главе анализируются существующие алгоритмы разрешения неоднозначности, основанные на метрических и линейных моделях и нейронных сетях, а также приводится обзор решений в области разрешения неоднозначности и извлечения значений слов на материале русского языка.

Анализ предметной области показал, что существующие аннотированные обучающие коллекции для разрешения неоднозначности на русском языке обладают недостаточным объемом и степенью покрытия многозначных слов. Кроме того, для русского языка существует мало систем разрешения неоднозначности, основанных на методах машинного обучения с учителем, а имеющиеся, в свою очередь, базируются на далеко не самых передовых методах. Очевидно, что отсутствие семантически размеченного набора данных для русского языка является препятствием к развитию эффективных систем снятия многозначности. Именно поэтому главной целью настоящей диссертационной работы является создание большого семантически размеченного корпуса для русского языка.

Для достижения этой цели планируется использовать один из подходов, описанных в обзоре, – подход, основанный на однозначных родственных словах. Преимуществом данного метода является то, что он опирается на семантические сети, которые можно найти для большинства языков (например, граф BabelNet по состоянию на февраль 2021 года содержит в себе информацию для 500 языков).

Сравнительный анализ различных подходов к разрешению неоднозначности, использующих размеченные данные, показал, что методы, основанные на нейронных сетях, превосходят все прочие по своей обобщающей способности. Таким образом, в данной работе автоматически полученные семантические обучающие коллекции предназначены для создания моделей разрешения неоднозначности, базирующихся глубоких языковых моделях.

Глава 2. Автоматическое порождение корпуса с семантической разметкой на основе однозначных кандидатов

2.1. Описание метода

Как уже говорилось ранее, в настоящем диссертационном исследовании для преодоления нехватки семантически аннотированных данных в русском языке использовался метод однозначных родственных слов. Главные составляющие части разработанной системы автоматического порождения семантической разметки обсуждаются в наших статьях [Bolshina, Loukachevitch, 2020a, 2020c, 2020d, 2020e].

Основная идея подхода для автоматической генерации размеченной обучающей коллекции базируется на предположении, что однозначные «родственники» могут быть связаны с целевым многозначным словом не только с помощью прямых отношений таких, как синонимия, гиперонимия и гипонимия, но и с помощью слов, расположенных в семантическом графе дальше от целевого значения, например, когипонимов. Например, большинство контекстов для слова *крона* в значении ‘валюта’ похожи на контексты, в которых встречаются другие слова, обозначающие валюту, например, *английский фунт* ‘фунт стерлингов’, потому что у этих слов есть общий гипероним *валюта*.

Основные характеристики подхода состоят в следующем:

1. Учитываются не только ближайшие к значению целевого слова однозначные родственные слова, как это делалось во многих предыдущих работах, но и более удаленные.
2. Чтобы оценить, насколько хорошо кандидат может отражать значение целевого многозначного слова, применяется коэффициент схожести однозначного «родственника-кандидата» и синсетов, близких к целевому значению многозначного слова.
3. Было введено понятие ***гнездо синсета***, которое помогает оценить потенциал контекстов употребления однозначного кандидата для

отображения смысла многозначного слова. Для того чтобы измерить, насколько хорошо однозначный кандидат подходит для задачи, используется набор слов из тезауруса близких к целевому значению. 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.

4. Как для близких, так и для более удаленных однозначных родственных слов вычисляются значения семантической близости к гнезду синсета, так как слово, описанное в тезаурусе как однозначное, в корпусе может иметь несколько значений. Например, русское слово *ириска* может также обозначать прозвище футбольного клуба Эвертон (The Toffees) [Loukachevitch, 2019]. Именно поэтому контексты из корпуса для всех однозначных кандидатов должны быть дополнительно проверены.
5. Были предложены два различных способа сбора обучающих коллекций на основе рейтинга однозначных «родственников».

Значение целевого многозначного слова – это значение многозначного слова, которое в данный момент рассматривается в системе разрешения неоднозначности или генерации обучающей коллекции. **Однозначные кандидаты («родственники-кандидаты»)** – это однозначные слова или словосочетания, которые могут быть расположены в пределах четырех шагов от многозначного слова. Рассматриваются только те слова или словосочетания, которые встречаются в корпусе хотя бы 50 раз.

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

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

Предлагаемый метод извлечения однозначных родственных слов основан на сравнении дистрибутивных и тезаурусных мер близости. Модели векторных

представлений слов используются для выбора наиболее подходящих однозначных «родственников», чей контекст может служить хорошей репрезентацией значения целевого слова. В данной работе применялись модели векторных представлений слов word2vec [Mikolov et al., 2013], основанные на архитектуре нейронных сетей CBOW. Они используются в работе для извлечения наиболее близких слов для каждого однозначного слова в списке кандидатов. Эти извлеченные слова представляют собой дистрибутивный набор близких слов с соответствующими косинусными коэффициентами близости. Ввиду того, что модель word2vec предназначена отражать, как однозначные родственные слова употребляются в текстах, предпочтительнее использовать модель, обученную на том же корпусе, из которого будут извлекаться примеры для обучающей коллекции. Алгоритм отбора и ранжирования однозначных родственных слов, состоит из следующих шагов:

1. Извлекаются все однозначные кандидаты в пределах 4 шагов от целевого значения многозначного слова 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 список 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 $\boldsymbol{w^j_{k_1}, \dots, w^j_{k_i}}$, которые представляют этот синсет в пересечении.

6. Финальный вес однозначного кандидата $\boldsymbol{r_j}$ вычисляется как сумма весов всех синсетов из гнезда синсета $\boldsymbol{ns_j}$. 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:

$$Weight_{r_j} = \sum_{k=1}^{N_k} \max\left[cos(r_j, w^j_{k_1}), \dots, cos\left(r_j, w^j_{k_i}\right)\right] \quad (2)$$

На рисунке 2 приведена упрощенная схема алгоритма порождения семантически размеченных обучающих коллекций:

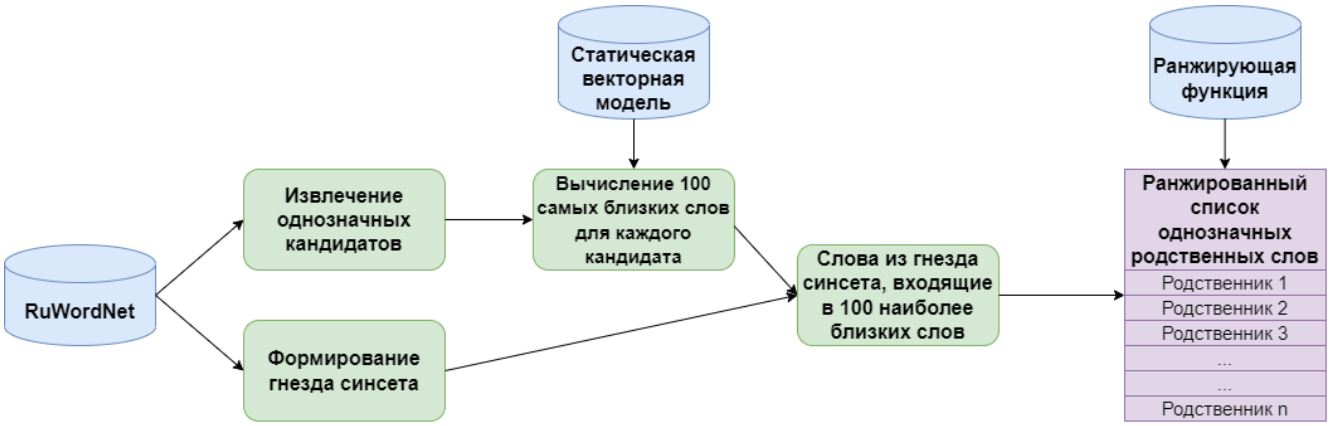

Рисунок 2. Метод сбора и разметки обучающих коллекций на основе однозначных родственных слов.

В примерах 3 и 4 приведены фрагменты списка однозначных родственных слов с соответствующими значениями близости (приведены в скобках), которые были получен для двух значений существительного *гвоздика*:

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

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

Примером пары слов, расположенных на небольшом расстоянии друг от друга в сети тезауруса, но получивших нулевой коэффициент близости по модели word2vec, являются слова *байка* и *марля*. *Марля* – когипоним для слова *байка* в значении ‘плотная ткань’, однако оно не было включено в список однозначных родственных слов, потому что его дистрибутивный набор близких слов не имеет пересечений с гнездом синсета.

В результате описанной процедуры все однозначные «родственники» сортируются в соответствии с весом, который они получили. Предполагается, что контексты с однозначными родственными словами, которые расположены выше в рейтинге, являются более подходящими примерами для репрезентации значения целевого слова. Алгоритм ранжирования однозначных кандидатов помогает определить, какие однозначные «родственники» ближе к целевому значению неоднозначного слова. Когда однозначные кандидаты отобраны и упорядочены по весам, из корпуса извлекаются контексты, в которых они встречаются. После этого в этих текстах однозначные родственные слова заменяются на целевые многозначные слова, и тексты добавляются в обучающую коллекцию.

2.2. Данные

В данном разделе будут описаны основные лингвистические ресурсы, которые используются в настоящем диссертационном исследовании для реализации всех компонентов метода однозначных слов. В качестве источника знаний и инвентаря значений используется семантическая сеть для русского языка RuWordNet. Эта семантическая сеть имеет WordNet-формат представления графов. Всего RuWordNet содержит 111.5 слов и словосочетаний для русского языка. RuWordNet содержит 29 326 синсетов для существительных, 7 509 синсетов для глаголов, 12 657 синсетов для прилагательных, 63 014 однозначных существительных, 21 051 однозначных глагола и 12 566 однозначных прилагательных. В таблице 3 приведены данные по общему числу многозначных слов, количеству значений, имеющихся для каждого многозначного слова для

трех частей речи, представленных в тезаурусе. На рисунке 3 приведены примеры описаний многозначных слов и их значений в тезаурусе[29].

Тезаурус RuWordNet использовался в работе:

- в качестве инвентаря значений;
- как источник однозначных родственных слов;
- для извлечения семантических отношений (например, синонимия, гипонимия и т.п.) между целевым значением многозначного слова и всеми словами (или словосочетаниями), связанными с ним;
- для вычисления расстояния между словами в графе.

Таблица 3. Количественные характеристики многозначных слов в RuWordNet.

Количество значений многозначного слова	**Количество существительных**	**Количество глаголов**	**Количество прилагательных**
2 значения	4 274	3 334	1931
3 значения	997	1 118	355
4 значения	399	532	83
5 значений	149	216	18
> 5 значений	76	124	5
Общее кол-во многозначных слов	5 895	5 324	2 392
Общее кол-во уникальных синсетов, в которые входят многозначные слова	10 779	4 916	4 453
Количество пар типа “многозначное слово:значение”	14 358	14 048	5 379

[29] https://ruwordnet.ru/ru

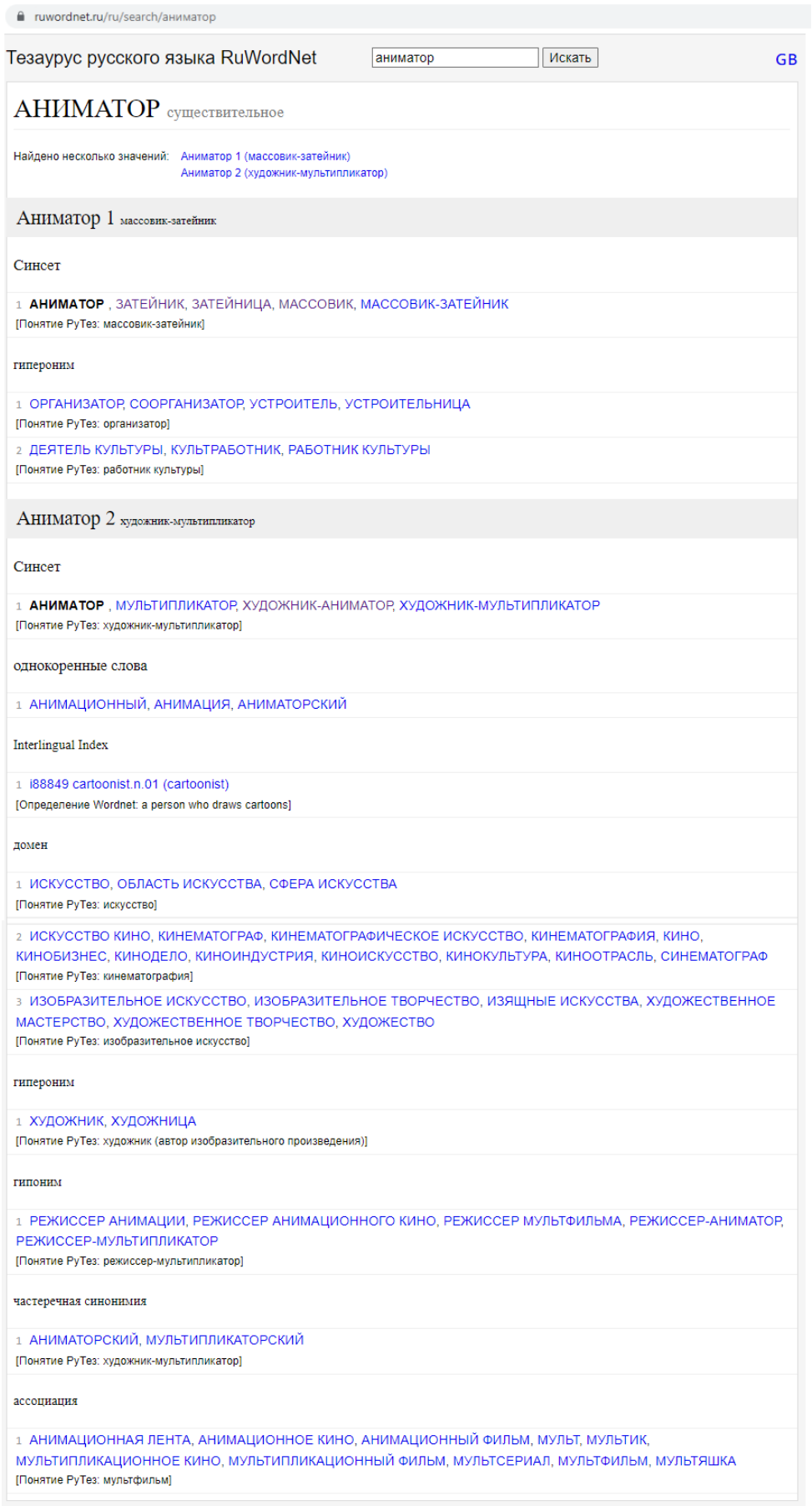

Рисунок 3. Описание многозначного слова *аниматор* в тезаурусе RuWordNet.

В работе использовалось два корпуса. Новостной корпус состоит из новостных статей, собранных из различных источников. Тексты были очищены от html-элементов и любой другой разметки. Второй корпус это – корпус «Тайга» [Shavrina, Shapovalova, 2017]. «Тайга» состоит из различных сегментов, в настоящем диссертационном исследовании использовался сегмент Проза.ру, представляющий собой набор текстов художественного стиля, а также новостной сегмент, в который вошли статьи из ресурсов Lenta.ru, Интерфакс, Комсомольская правда, Журнальный зал, Fontanka.ru. Эти два корпуса применялись для обучения статических word2vec моделей, а также для извлечения контекстов с однозначными родственными словами, которые затем добавлялись в обучающую коллекцию. Использованные в работе модели представлений слов word2vec, основаны на архитектуре нейронных сетей CBOW (с помощью этой архитектуры в языковой модели предсказывается текущее слово, исходя из окружающего его контекста) с размером окна равным трем. Всего в рамках диссертационного исследования были обучены две модели – на корпусе «Тайга»-Проза.ру и на новостном корпусе. В таблице 4 представлены характеристики всех описанных выше корпусов.

Таблица 4. Количественные характеристики корпусов, использованных в экспериментах.

	«Тайга»-Проза.ру	**«Тайга»-Новостной сегмент**	**Новостной корпус**
Количество предложений	32,8 миллиона	14,6 миллионов	24,2 миллиона
Количество лемм	246,8 миллиона	130,8 миллионов	288,1 миллиона
Количество уникальных лемм	2,1 миллиона	1,2 миллиона	1,4 миллиона

2.3. Подготовка обучающей коллекции с помощью однозначных родственных слов

В данном разделе на конкретных примерах из исходных данных будут продемонстрированы основные характеристики метода однозначных родственных слов. Для того чтобы проверить применимость данного метода к материалу RuWordNet, с помощью описанного алгоритма, но без применения фильтрации word2vec моделью, были найдены однозначные кандидаты для всех многозначных существительных в тезаурусе. Оказалось, что только два существительных из 5895 не имеют однозначных родственных слов в пределах 4 шагов в сети RuWordNet. Количественные характеристики извлеченных однозначных кандидатов представлены в таблице 5.

Таблица 5. Количественные характеристики однозначных кандидатов для целевых значений существительных в RuWordNet.

Расстояние до однозначного кандидата	**Количество целевых значений, которые имеют хотя бы одного однозначного кандидата на этом расстоянии**
0 (синсет)	9 818
1	13 095
2	14 129
3	14 021
4	13 768

Таблица 6. Целевые значения, которые имеют минимум 500 примеров употреблений своих однозначных родственных слов в корпусе.

	Количество целевых значений до фильтрации word2vec моделью	**Количество целевых значений после фильтрации word2vec моделью**
«Тайга» - Проза.ру	13 738	12 797
Новостной корпус	14 017	13 099

В работе [Taghipour et al., 2015a: 339] было отмечено что, 500 примеров для каждого значения является достаточным количеством для составления обучающей коллекции. В таблице 6 показано, сколько пар «многозначное существительное – значение» имеют по крайней мере 500 примеров со своими однозначными родственными словами в корпусе. Был также принят во внимание случай, когда однозначные кандидаты были отфильтрованы word2vec моделью. Данные таблиц показывают, что, применяя описанный метод к материалу RuWordNet, можно найти однозначных «родственников» почти для всех многозначных существительных в тезаурусе и, таким образом, создать обучающую коллекцию для обучения алгоритма автоматического разрешения неоднозначности.

Как говорилось ранее, метод однозначных родственных слов основан на заменах. После того, как однозначные родственники были найдены и ранжированы, контексты их употребления извлекаются из выбранного корпуса. Для сравнения качества моделей в настоящем диссертационном исследовании были созданы две основные отличающиеся по жанру обучающие коллекции: из текстов новостного корпуса и Проза.ру[30]. В таблице 7 приведены контексты, в которых однозначные родственные слова, найденные для двух значений слова *барометр*, были заменены на слово *барометр*. Такие предложения с заменами могут звучать неестественно, однако контекст, окружающий многозначное слово, помогает понять, в каком значении слово употреблено.

Помимо этого, использовались два способа сбора обучающей коллекции на основе рейтинга однозначных родственных слов. В соответствии с первым методом сбора обучающей коллекции была создана коллекция только с тем однозначным родственным словом, которое получило наибольший вес во время процедуры ранжирования. Планировалось получить для каждого целевого значения по 1000 примеров, но иногда это было невозможно ввиду того, что для однозначного родственного слова не было такого числа примеров в корпусе.

[30] В разделе 3.3.2 используется обучающая коллекция, созданная из текстов новостного корпуса и новостного сегмента корпуса «Тайга».

Поэтому в таких случаях также выбирались примеры со словами, идущими следующими в рейтинге, до тех пор, пока не набиралось 1000 примеров. Для удобства будем называть эту коллекцию 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. Эту коллекцию назовем сбалансированной, так как выбор обучающих примеров не был ограничен только одним однозначным родственным словом.

Таблица 7. Примеры контекстов употребления однозначных родственных слов с заменой их на многозначное слово.

Многозначное слово	**Однозначный «родственник»**	**Контекст с заменой**
барометр ‘индикатор’	*точка отсчета* (когипоним)	**Барометром** чиновничьего летоисчисления является создание нового органа власти – сената.
барометр ‘гидрометеорологический прибор’	*альтиметр* (гипоним)	Дополнительно учёные анализируют рельеф поверхности с помощью лазерного **барометра**.

В экспериментах были использованы две word2vec модели, обученные, соответственно, отдельно на новостном корпусе и корпусе Проза.ру с размером окна равным трем. На этапе предобработки данных корпус был разбит на отдельные предложения, затем каждое предложение токенизировалось, из него убирались все стоп-слова, знаки препинания (кроме дефисов) и числа, а все оставшиеся токены лемматизировались с помощью инструмента pymorphy2 [Korobov, 2015]. С помощью этих моделей для каждого однозначного кандидата извлекались 100 наиболее похожих слов по косинусной мере, которые затем использовались для нахождения пересечения со словами из гнезда синсета. Слова,

извлеченные из word2vec моделей, отфильтровывались – убирались те, что не были включены в словарь тезауруса.

Выводы к главе 2

В главе 2 описан разработанный в рамках диссертационного исследования новый подход для автоматического сбора и разметки обучающих коллекций на основе однозначных родственных слов. Также был продемонстрирован механизм ранжирования однозначных кандидатов слов на основе близости их векторных представлений со словами семантически близкими целевому значению. Помимо этого, были предложены два способа формирования обучающих коллекций с помощью рейтинга однозначных родственных слов.

Принимая во внимание недостатки работ других исследователей, которые в основном опираются на близко расположенных в семантическом графе однозначных «родственников», в предложенном методе учитываются однозначные «родственники», расположенные на удаленном расстоянии от целевого многозначного слова в семантическом графе, что позволяет ему находить однозначные родственные слова для большинства значений многозначных слов. Благодаря компоненту, фильтрующему однозначных кандидатов, в обучающие коллекции попадает меньше нерелевантных примеров.

Количественный анализ однозначных кандидатов для существительных из тезауруса RuWordNet показал, что больше всего однозначных родственных слов находятся на расстоянии 2-3 шагов от целевого синсета и являются его когипонимами. Это говорит о большом вкладе далеко расположенных «родственников» в формирование обучающей коллекции.

Глава 3. Снятие лексической многозначности

3.1. Разрешение лексической неоднозначности на наборе данных RUSSE-RuWordNet

Было проведено несколько экспериментов для оценки качества, которое могут давать модели разрешения неоднозначности, обученные на коллекциях, собранных по вышеописанному методу. Основные результаты представлены в наших статьях [Bolshina, Loukachevitch, 2020a, 2020c, 2020d, 2020e].

В данной серии экспериментов было использовано несколько источников данных, речь о которых пойдет дальше. Для оценки метода сбора и разметки обучающей коллекции использовались три разных набора данных из технологического соревнования RUSSE-2018. Из них были исключены некоторые многозначные слова: таблица 8 представляет наиболее частые причины исключения слов. Результирующий список целевых неоднозначных слов состоял из 30 существительных, у каждого из которых по два значения (см. приложение 2). Полученный набор данных будет обозначаться 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 было аннотировано 385 предложений с 27 целевыми неоднозначными словами, входящими в выборку RUSSE-RuWordNet. В этом тестовом наборе данных не было контекстов со словами *бор* ‘сосновый лес/бор’, *лук* ‘лук/лук’ и *гвоздика* ‘гвоздика/гвоздика’.

Также была создана небольшая размеченная выборка, которая состоит из определений значений слов и примеров их употреблений, взятых из Словаря Ожегова[31] для каждого целевого многозначного слова. Эти данные

[31] https://slovarozhegova.ru/

использовались для обучения базового решения (*baseline*) задачи разрешения лексической неоднозначности, с которым будут сравниваться остальные подходы. В этом наборе для каждого значения многозначного слова есть от одного до трех примеров. Количественные характеристики всех описанных наборов данных приведены в таблице 9.

В данной серии экспериментов, изучались два типа обучающих коллекций – сбалансированная и Корпус-1000. В таблице 10 представлены следующие количественные характеристики однозначных «родственников», включенных в две сбалансированные обучающие коллекции: отношения, соединяющие целевое значение многозначного слова и его однозначные родственные слова, расстояния между ними, а также пропорция однозначных родственных слов, выраженных словосочетанием. Количество примеров в коллекциях приведено в приложении 2.

Таблица 8. Случаи, при которых слово из RUSSE'18 не включалось в финальный набор данных RUSSE-RuWordNet.

Объяснение	**Кол-во слов**	**Пример**
У слова только одно значение в RuWordNet	34	Слово *двойник* имеет только одно значение в RuWordNet, в то время как в RUSSE'18 у него их 4.
Слово отсутствует в словаре RuWordNet	9	Слово *гипербола.*
Только одно общее значение слова для значений из RuWordNet и RUSSE'18	4	Слово *мандарин* имеет два значения в RUSSE'18: его значение 'фрукт' есть в тезаурусе, а значение 'чиновник' отсутствует.
Неоднозначные случаи соотнесения значений	29	Слово *демократ* имеет 2 значения: 'сторонник демократии' и 'член демократической партии'. Но в RUSSE'18 есть еще одно значение: 'человек демократичного образа жизни, взглядов'.
Недостаточно примеров в корпусе	2	Слова *карьер* и *шах*.
Слова с морфологической омонимией	1	Слово *суда* 'суд (Gen, Sg)/корабль (Nom, Pl)'. Эти два слова имеют разные леммы.

Таблица 9. Количественные характеристики наборов данных, использованных в экспериментах.

	RUSSE-RuWordNet	**Набор данных «Комсомольская правда»**	**Словарный корпус**
Количество предложений	2 103	385	144
Количество лемм	39 311	4 916	657
Количество уникальных лемм	12 110	2 558	475

Таблица 10. Количественные характеристики однозначных родственных слов, включенных в сбалансированную коллекцию.

Расстояние до целевого значения многозначного слова	**Пропорция употреблений в новостной коллекции**	**Пропорция употреблений в Проза.ру коллекции**
0 (синсет)	2%	4%
1	13%	9%
2	38%	37%
3	31%	34%
4	16%	16%
Отношение между целевым значением и однозначным родственным словом		
Синонимы	2%	4%
Гипонимы	13%	8%
Гиперонимы	11%	9%
Когипонимы	28%	28%
Когипонимы на расстоянии 3 шагов	24%	28%
Когипонимы на расстоянии 4 шагов	19%	22%
Другие	3%	1%
Словосочетания	48%	29%

Как и в [Wiedemann et al., 2019], в данном исследовании использовался легко интерпретируемый алгоритм классификации – метод ближайших соседей (kNN), объектами классификации которого будут контекстуализированные представления слов, извлеченные из языковых моделей ELMo [Peters et al., 2018]

и BERT [Devlin et al., 2019]. В экспериментах использовались две модели ELMo – одна обученная коллективом DeepPavlov на русской части корпуса WMT News, а другая модель RusVectōrēs [Kutuzov, Kuzmenko, 2016], обученная на лемматизированном корпусе «Тайга». В диссертационном исследовании изучались два способа извлечения представлений слов из модели ELMo: вычисление вектора для всего предложения, содержащего ключевое слово, или отдельный вектор для целевого неоднозначного слова. В работе также применялись две модели BERT: мультиязычная модель BERT-base-multilingual-cased от Google Research и RuBERT, который был обучен DeepPavlov [Kuratov, Arkhipov, 2019] на русскоязычной части Википедии и новостных материалах. Для извлечения контекстуализированных представлений из предварительно обученной модели BERT использовался метод, описанный в [Devlin et al., 2019] и [Wiedemann et al., 2019]: конкатенировались векторные представления слов из четырех последних слоев уже обученного трансформера. Таблицы 11 и 12 демонстрируют результаты, полученные с помощью разных обучающих коллекций, контекстуализированных представлений слов и параметров модели.

Таблица 11. Значения метрики F1 для моделей, основанных на векторах слов 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 12. Значения метрики F1 для моделей, основанных на векторах слов 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	

Данные, приведенные выше, показывают, что все системы смогли показать результаты лучше, чем базовое решение, для обучения которого использовались тексты словарных дефиниций и примеров употреблений слов. Это значит, что с помощью собранной обучающей коллекции можно обучить алгоритм разрешения неоднозначности, показывающий хорошее качество предсказаний на наборе данных RUSSE-RuWordNet.

Модели, обученные на Проза.ру, показывают более высокие результаты, чем модели, использовавшие материал новостного корпуса. Качественный анализ ошибок классификации, сделанных моделью, обученной на новостном корпусе, показал, что главная причина ошибок – это лексические и структурные различия между обучающим и тестовым корпусами. Примеры из тестового корпуса были взяты из Национального корпуса русского языка и Википедии, в то время как

обучающая коллекция состояла из новостных статей. В корпусе Проза.ру, напротив, содержатся различные художественные произведения, поэтому обучающие примеры из этой коллекции имеют более схожую репрезентацию с тестовыми.

Для того, чтобы проверить зависимость качества разрешения неоднозначности от жанра тестовой коллекции, был проведен еще один эксперимент. На этот раз был взят набор данных, состоящий из статей «Комсомольской правды». В этом эксперименте использовалась сбалансированная обучающая коллекция и языковая модель RusVectōrēs ELMo (вектор извлекался для целевого слова). Наилучший результат на этом тестовом наборе данных был получен с помощью модели, обученной на новостной коллекции, и составил **0.78** F1; метрика F1 для модели, обученной на Проза.ру, была равна **0.74.** Таким образом, эти показатели подтверждают предположение о том, что лучшая производительность в задаче разрешения неоднозначности достигается, когда жанр обучающей и тестовой коллекций совпадает.

Система, основанная на контекстуализированных представлениях слов ELMo от RusVectōrēs, превзошла по качеству все остальные модели, ее 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 контекстуализированных векторах слов от Multilingual BERT. 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. Это может говорить о том, что подход, использованный для сбора сбалансированной коллекции, больше подходит для текущей задачи. В Корпус-1000 не включены все возможные для целевого значения родственные слова, поэтому в этой коллекции отсутствует контекстное разнообразие. Сбалансированная коллекция, напротив, более репрезентативна в плане множества контекстов.

Как уже было упомянуто ранее, контекстуализированные представления слов из языковой модели ELMo можно использовать двумя разными способами, поэтому был проведен еще один эксперимент, чтобы понять, какой из них лучше подходит для задачи разрешения неоднозначности, а также для конкретной модели. В первых двух столбцах таблицы 13 показаны результаты классификации с помощью контекстуализированных представлений, извлеченных из языковых моделей ELMo от RusVectōrēs и DeepPavlov.

Таблица 13. Значения метрики F1 для моделей, основанных на векторах 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	-

В отличие от предыдущего эксперимента в данном случае из модели от RusVectōrēs для представления ключевого слова извлекался вектор всего предложения, а из модели от DeepPavlov брался только вектор самого ключевого слова. Полученные результаты демонстрируют, что эти способы извлечения данных из языковых моделей ухудшили предыдущие результаты классификации для обеих моделей. Таким образом, следующие методы использования контекстуализированных представлений из моделей ELMo для задачи разрешения лексической неоднозначности с помощью kNN-классификатора являются оптимальными: из модели от RusVectōrēs предпочтительнее извлекать вектор

только для ключевого неоднозначного слова, а из модели от DeepPavlov рекомендуется брать вектор для всего предложения, в котором встретилось ключевое слово.

Результаты исследования [Kutuzov, Kuzmenko, 2019] показали, что лемматизированные обучающие данные могут улучшить результаты, которые ELMo показывает на задаче разрешения неоднозначности для русского языка. В рамках диссертационной работы был проведен эксперимент, который показал, что это верно также и для автоматически сгенерированных обучающих коллекций. Для сравнения были взяты две модели ELMo от RusVectōrēs: модель, обученная на лемматизированном корпусе «Тайга», и модель, обученная на нелемматизированных НКРЯ и российском сегменте Википедии. В качестве обучающей коллекции использовалась сбалансированная коллекция Проза.ру в двух вариантах – лемматизированная и просто токенизированная. Результаты для нелемматизированной обучающей коллекции представлены в последнем столбце таблицы 13, а для лемматизированной – в таблице 12. Эксперимент показывает, что даже для сгенерированных обучающих коллекций ELMo модель, обученная на леммах, дает более высокое качество, чем модель обученная на нелемматизированных данных. Таким образом, для русского языка для задачи разрешения лексической неоднозначности является предпочтительным обучать модели на лемматизированных данных, так как они не содержат дополнительной морфологической информации, являющейся излишней для лексико-семантической задачи.

Другой эксперимент был направлен на оценку модели, обученной на автоматически сгенерированной обучающей коллекции, дополненной словарными дефинициями и примерами употребления ключевых слов. Начиная с самых ранних работ в области автоматического разрешения неоднозначности [Lesk, 1986], словарные дефиниции являлись ценным источником информации для моделей. В разделе 1.6.2 отмечалось, что в современных исследованиях определения слов также активно используются для улучшения качества снятия

неоднозначности. В рамках этого исследования словарные дефиниции и примеры употребления слов, уже использованные ранее в базовом решении, были добавлены в новостную сбалансированную обучающую коллекцию и коллекцию Проза.ру. После этого из этих расширенных обучающих данных были извлечены контекстуализированные представления ELMo RusVectōrēs для ключевых слов, к которым затем был применен kNN-классификатор. Результаты разрешения неоднозначности представлены в таблице 14, в скобках указан прирост качества. Несмотря на то, что количество новых примеров было невелико, можно наблюдать незначительные улучшения в качестве предсказаний для модели, обученной на Проза.ру, и увеличение F1 на 2% для новостной коллекции.

Таблица 14. Значения метрики F1 для моделей, основанных на векторах слов из языковой модели 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 (+0.007)	0.824 (+0.027)
3	0.835 (+0.002)	0.832 (+0.022)
5	0.847 (+0.002)	0.828 (+0.018)
7	**0.859** (+0.002)	0.834 (+0.019)
9	0.858(+0.002)	**0.842** (+0.021)

Помимо этого, было проведено сравнение качества предсказаний моделей, обученных на автоматически и вручную размеченных данных. В этом эксперименте также использовались контекстуализированные представления слов ELMo от RusVectōrēs. Для каждого целевого значения из набора данных RUSSE-RuWordNet создавались 5 случайных разбиений примеров на обучающие и тестовые в соотношении 2:1. Затем эти данные применялись для обучения и тестирования пяти различных моделей разрешения неоднозначности. Среди

результатов, полученных для каждого классификатора (то есть среди результатов для всех значений k), брался максимальный, а затем финальное значение F1 представлялось как среднее этих пяти значений. F1 при таком обучении составила **0.917**.

После этого было подсчитано значение F1 на пяти тестовых сетах, которые описывались выше, с помощью модели, обученной на новостном корпусе. Качество предсказаний модели, обученной на новостном корпусе, на этих 5 тестовых выборках составило **0.84 F1**. Затем оценивалось качество модели, обученной на комбинации новостной коллекции с каждым из обучающих подмножеств, описанных выше, ее F1 составила **0.94**. Таким образом, обучающая выборка, представленная комбинацией автоматически и вручную размеченных данных, показывает наилучший результат предсказаний (**0.94 F1**).

Эксперименты, представленные в этом разделе, показали, что наилучший результат разрешения неоднозначности достигается, когда жанры обучающей и тестовой коллекции совпадают. Помимо этого, было доказано, что модель, обученная на лемматизированной автоматически сгенерированной обучающей коллекции, по качеству превосходит модель, обученную на нелемматизированном корпусе. Кроме того, эксперимент продемонстрировал, что модель, обученная на комбинации автоматически и вручную размеченных данных, показывает наилучшее качество предсказаний.

3.2. Разрешение лексической неоднозначности для всех частей речи

В этой серии экспериментов с помощью разработанного метода генерации обучающих данных была создана обучающая коллекция для многозначных существительных, глаголов и прилагательных. На основе этой коллекции была обучена модель, с помощью которой делались предсказания значений всех многозначных слов в тексте. В качестве набора данных, который размечался моделью, были выбраны новостные статьи из секции экономики на ресурсе

Wikinews. Как уже упоминалось ранее, в методе отбора и ранжирования однозначных родственников предпочтительнее использовать word2vec модель, обученную на том же корпусе, из которого будут извлекаться примеры для обучающей коллекции. В дополнение к этому, результаты предыдущих экспериментов показали, что наилучшее качество разрешения неоднозначности достигается, когда жанры обучающей и тестовой коллекции совпадают. Ввиду этих факторов в данном эксперименте новостной корпус использовался для извлечения контекстов, а модель, обученная на нем, применялась для отбора и ранжирования однозначных кандидатов. Данный эксперимент и его результаты были описаны в нашей статье [Bolshina, Loukachevitch, 2020b].

3.2.1. Количественные характеристики многозначных слов и их однозначных родственных слов

В таблице 15 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. По сравнению с глаголами и прилагательными, у существительных выше количество уникальных однозначных родственников, найденных для их ключевых значений, что может означать, что существительные требуют более объемного корпуса, из которого будут извлекаться примеры, чтобы покрыть большинство отобранных алгоритмом родственников. Медианное значение количества однозначных родственников, приходящихся на одно значение, почти одинаковое для всех трех частей речи и колеблется от 10 до 16 слов. Среднее арифметическое однозначных родственников также довольно

высокое, что говорит о том, что обычно многозначное слово имеет несколько подходящих родственников. Этот факт свидетельствует о том, что обучающая коллекция, сгенерированная автоматически с помощью нашего метода, отличается лексическим многообразием, так как при ее составлении использовались различные однозначные родственники.

Таблица 15. Количественные характеристики многозначных слов и их однозначных родственных слов.

	Существительные	**Глаголы**	**Прилагательные**
Общее число однозначных слов в тезаурусе RuWordNet	63 014	21 051	12 566
Общее число многозначных слов в тезаурусе RuWordNet	5 895	5 324	2 392
Общее число уникальных значений, которые есть у многозначных слов	10 779	4 916	4 453
Количество пар типа “многозначное слово:значение”	14 358	14 048	5 379
Количество пар типа “многозначное слово:значение”, для которых не нашлось однозначных родственных слов	676	255	688
Количество многозначных слов, у которых есть однозначные родственные слова по крайней мере для двух значений	5 511	5 220	1 910
Количество многозначных слов, у которых есть однозначные родственные слова для всех значений	5265	5080	1813
Количество уникальных однозначных родственников	17173	7224	6391
Среднее число однозначных родственных слов на одно значение	36	42	78
Медианное значение однозначных родственных слов на одно значение	10	16	10

Таблицы 16 и 17 представляют характеристики самих однозначных родственных слов относительного того, к многозначному слову какой части речи они относятся.

Таблица 16. Характеристики отношений между целевыми многозначными словами и однозначным родственными словами.

Отношение	**Существительные**	**Глаголы**	**Прилагательные**
Синонимы	2%	3%	1%
Гипонимы	13%	13%	3%
Гиперонимы	5%	9%	2%
Когипонимы	29%	31%	34%
Когипонимы, расположенные на расстоянии 3 шагов	31%	27%	36%
Когипонимы, расположенные на расстоянии 4 шагов	18%	14%	23%
Другие	2%	3%	1%

Таблица 17. Расстояния между целевыми многозначными словами и однозначным родственными словами.

Расстояние	**Существительные**	**Глаголы**	**Прилагательные**
0 (синсет)	2%	3%	1%
1	5%	9%	2%
2	36%	40%	37%
3	36%	31%	36%
4	21%	17%	24%

Анализ представленных данных показывает, что большинство однозначных родственников расположены на более удаленных расстояниях от ключевого

слова. Можно заметить, что пропорция близких родственников, таких как гипонимы и гиперонимы, выше для существительных и глаголов, тогда как для прилагательных эти пропорции относительно низкие. Помимо этого, такие отношения как 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. Приведенные данные еще раз подтверждают гипотезу о том, что использование однозначным родственников, расположенных на удаленном расстоянии от ключевого слова, положительно влияет на качество генерируемой обучающей коллекции.

3.2.2. Полуавтоматическая разметка всех многозначных слов в тексте

Эксперимент по разметке всех многозначных слов в тексте с помощью модели, обученной на автоматически сгенерированной коллекции, состоял из следующих этапов: сначала создавалась обучающая коллекция для решения задачи разрешения неоднозначности, затем на полученной коллекции обучалась модель, с помощью которой делались предварительные предсказания на оценочном наборе данных. Наконец, эти предсказания были вручную проверены и скорректированы.

С помощью метода однозначных родственных слов был получен список однозначных родственников, из которого были удалены те слова, что имели нулевой вес. В настоящем исследовании принимались во внимание только те многозначные слова, у которых однозначные родственники были найдены для всех значений. Помимо этого, в эксперименте применялся сбалансированный способ сбора обучающей коллекции: результаты предыдущих экспериментов показали, что модель, обученная на корпусе, собранном с помощью этого способа, показала более высокое значение F1.

Для каждого значения многозначного слова из новостного корпуса с помощью описываемого метода автоматически извлекалось и размечалось около

30 примеров. Безусловно, для моделей разрешения неоднозначности, основанных на нейронных сетях, такого количества обучающих примеров недостаточно, но в текущем эксперименте использовался уже ранее описанный kNN-классификатор, базирующийся на контекстуализированных представлениях слов из языковой модели ELMo. Эта система уже применялась в предыдущих экспериментах и показала хорошее качество разрешения неоднозначности (0.857 F1). Кроме того, этот метод классификации не требует такого большого количества размеченных данных, как нейросети. Как и в экспериментах ранее (см. раздел 3.1), в текущем будет использована ELMo-модель от RusVectōrēs, обученная на лемматизированном корпусе «Тайга».

Таблица 18. Характеристика многозначных слов, представленных в оценочном наборе данных.

	Существительные	**Глаголы**	**Прилагательные**
Количество уникальных многозначных слов	224	159	84
Общее количество многозначных слов	400	208	149
Количество уникальных синсетов	243	156	93
Количество слов, для которых нет значения в тезаурусе	11	7	1
Количество слов, являющихся частью коллокации	14	6	1
F1 kNN-классификатора (количество ближайших соседей = 5)	**0.8**	**0.72**	**0.8**

Как было упомянуто ранее, в текущем исследовании в качестве оценочного набора данных использовались новостные статьи из секции экономика ресурса Wikinews. На этапе предобработки данных была проведена лемматизация, и из текстов были удалены стоп-слова. В совокупности, выборка состояла из 107 предложений, 1777 лемм и 1047 уникальных лемм. Более подробная характеристика многозначных слов из этого набора данных представлена в таблице 18.

Результаты, полученные с помощью kNN-классификатора, базирующегося на контекстуализированных представлениях слов из языковой модели ELMo, были вручную проверены. Значения F1 для трех частей речи представлены внизу таблицы 18. Несмотря на то, что качество предсказаний не очень высокое, предварительная аннотация, сделанная классификатором, облегчила ручную работу по разметке данных. Описываемый эксперимент еще раз продемонстрировал, что разметка значений слов в корпусе требует значительных усилий и много времени.

Для того, чтобы изучить способы еще большего уменьшения ручного труда в разметке данных, была произведена следующая оценка. В описываемых экспериментах применялась довольно простая и легко интерпретируемая модель, основной параметр в которой – это количество соседей (в данном эксперименте их число было равно 5). Была произведена оценка предсказаний для многозначных глаголов в том случае, когда 4 из 5 примеров определенного значения были ближайшими к ключевому слову, и в ходе нее было обнаружено, что в 80% этих случаев метка значения, которая была приписана моделью, оказалась правильной. Можно предположить, что модель с более сложной архитектурой будет давать более точную вероятностную оценку возможным значениям. Это свойство можно будет использовать для пополнения обучающей коллекции с помощью тех примеров, которым модель будет приписывать высокое значение вероятности. Такая вероятностная оценка значений также может облегчить затраты на ручную разметку обучающих коллекций.

3.2.3. Анализ ошибок

В ходе проверки размеченных предложений были обнаружены случаи, выбор метки в которых был не так очевиден. Например, значение слова *история* во фразе *переписать историю* должно быть выбрано из следующих значений: ‘исторические науки’; ‘рассказ, словесное изложение’; ‘история, ход развития, движения’; ‘историческое развитие’. Общим решением было принято разметить это употребление слова значением ‘историческое развитие’, хотя с первого взгляда это выбор этой метки был не таким однозначным. Другой пример связан с глаголом *встретить* в словосочетании *встретить понимание*, здесь среди всех возможных значений глагола было выбрано значение ‘встретиться в жизни, деятельности’. Для того, чтобы определиться с меткой значений, потребовалось проанализировать все остальные значения и примеры их употреблений, и методом исключения было принято финальное решение.

Помимо этого, были обнаружены случаи, где у слова были значения, которые не включены в инвентарь значений. Иногда это было вызвано тем, что это значение у слова появилось относительно недавно, например, слово *канал* в значении ‘канал YouTube’. Иногда правильное значение слова попросту отсутствует в инвентаре тезауруса, например, глагол *бросить* в значении ‘оставлять, покидать кого-то/что-то’ в контексте *бросить на растерзание*. Статистика по подобным случаям приведена в таблице 18.

Кроме того, в данных был ряд случаев, когда многозначное слово было частью коллокации, поэтому для него не требовалось выбирать какое-либо конкретное значение. Среди подобных примеров можно отметить следующие: прилагательное *большой* в словосочетании *по большому счету*; глагол *отдавать* во фразе *отдавать себе отчёт*; существительное *свет* в выражении *выйти в свет*. Количество случаев такого рода также отражено в таблице 18.

В ходе анализа выяснилось, что в некоторых примерах морфологическая неоднозначность пересекалась с лексической. Слово *стали* может быть, как формой прошедшего времени 3 лица множественного числа многозначного

глагола *стать*, так и однозначным существительным *сталь* в формах Р.п/Д.п./Пр.п Ед.ч или Им.п./В.п. Мн.ч. Текущая система разрешения неоднозначности работает на лемматизированных текстах, и в процессе приведения к нормальной форме словоформа глагола *стать* была некорректно приведена к лемме *сталь*, поэтому для данного ключевого слова в таком случае предсказания получены не были. Эта проблема может быть решена либо использованием другого инструмента для лемматизации (который бы мог разрешать морфологическую омонимию), либо с помощью применения другой модели разрешения лексической неоднозначности, которая бы делала предсказания на нелемматизированных текстах. Однако даже в текстах, не подвергнутых никакой предобработке, могут встречаться случаи с морфологической неоднозначностью.

В данных также был обнаружен случай, когда многозначное слово было частью имени собственного, поэтому оно также не требовало снятия неоднозначности: слово *новости* в названии новостного агентства РИА Новости.

3.2.4. Итоги эксперимента по предсказанию значений для всех частей речи

Основная цель этого эксперимента заключалась в том, чтобы оценить возможность разработанного подхода порождать семантически размеченную обучающую коллекцию для разных частей речи. Исследование продемонстрировало, что метод однозначных родственных слов подходит для создания обучающих коллекций для всех частей речи, имеющихся в тезаурусе RuWordNet. Была выдвинута гипотеза, что если обучить более сложную и мощную модель на большем числе примеров, то получится улучшить полученные результаты.

В этом эксперименте оценочный набор данных создавался полуавтоматически, а именно, в качестве основы выступали предсказания модели разрешения неоднозначности, которые вручную проверялись и исправлялись.

Используя данную тестовую выборку, был проведен анализ предсказаний модели, обученной на автоматически размеченной выборке. Выявленные ошибки были разбиты на несколько классов в зависимости от того, чем они были вызваны. Это поможет учесть их в будущих исследованиях.

3.3. Разрешение неоднозначности на основе псевдоаннотированной коллекции

Еще один способ решения проблемы недостатка аннотированных данных в русском языке – это использование псевдоразметки. Псевдоаннотированные коллекции – это такие коллекции, в которых разметка была получена с помощью модели, обученной на размеченных данных [Тесленко, Усталов, 2018; Rizve et al., 2021]. В данном разделе рассматривается подход для порождения псевдоразметки с использованием ансамбля моделей, базирующихся на слабо контролируемом обучении.

Для первичной разметки данных использовался метод однозначных родственных слов, разработанный в рамках диссертационного исследования. Очевидно, что предложенный подход не гарантирует стопроцентную точность разметки обучающих данных, к тому же ввиду того, что основополагающий принцип порождения обучающих контекстов – это лексические замены, генерируемые тексты не всегда звучат естественно. Несмотря на это, подобные «шумные» данные могут использоваться в такой парадигме обучения как слабо контролируемое обучение (*weak supervision*), которая подразумевает обучение моделей на коллекциях с неточными, зашумленными метками. Чаще всего такую разметку получают с помощью вручную заданных эвристик, внешних баз знаний, предсказаний предварительно обученных классификаторов и т.п. Слабо контролируемое обучение сейчас активно используется во многих областях автоматической обработки текстов для преодоления нехватки размеченных данных, например, в задаче распознавания сущностей [Lison et al., 2021], задаче связывания именованных сущностей [Le, Titov, 2019], извлечении отношений [Lin

et al., 2016; Li et al., 2021] и классификации текстов [Wang et al., 2019]. Таким образом, с помощью синтетических данных, сгенерированных методом однозначных родственных слов, были обучены три модели для разрешения неоднозначности. Так как метки, приписанные разработанным методом, могут быть неточными, было принято решение для получения значений ключевых многозначных слов использовать все модели в ансамбле. Эксперимент по порождению псевдоразметки и его результаты обсуждаются в статьях [Bolshina, Loukachevitch, 2021] и [Большина, 2022].

3.3.1. Используемые данные и модели

В экспериментах использовались два корпуса: новостной корпус и новостной сегмент корпуса «Тайга». Эти текстовые коллекции служили источником предложений с ключевыми многозначными словами, которые впоследствии размечались. Помимо этого, для извлечения однозначных родственных слов применялась word2vec модель, обученная на новостном корпусе.

Для данного эксперимента были отобраны многозначные слова, которые не входили в набор данных RUSSE-RuWordNet, при этом они должны были встретиться в новостном корпусе не менее 500 раз и не меньше, чем в 200 документах. Всего в исследовании рассматривался набор из 10 ключевых многозначных слов, для которых вручную был составлен и размечен тестовый набор данных, состоящий из статей Wikinews.

Толкования многозначных слов из различных лексических ресурсов часто используются в области разрешения неоднозначности [Luo et al., 2018; Huang et al., 2019; Blevins, Zettlemoyer, 2020]. В экспериментах, представленных в данном разделе, словарные дефиниции многозначных слов применялись для дополнения обучающих данных. Для каждого значения ключевого многозначного слова был собран набор словарных дефиниций и примеров употреблений, взятых из различных толковых словарей и Викисловаря. Эти данные добавлялись в

обучающую коллекцию с псевдоразметкой. Список выбранных ключевых многозначных слов, количество аннотированных контекстов для их значений в тестовом корпусе, а также число примеров со словарными толкованиями и примерами употреблений представлены в таблице 19.

Таблица 19. Количественные характеристики выбранных многозначных слов и их значений.

Многозначное слово и его значение	**Кол-во тестовых примеров**	**Кол-во толкований и примеров употреблений**
аниматор$_0$ ‘мультипликатор’	29	11
аниматор$_1$ ‘массовик-затейник’	28	4
барометр$_0$ ‘гидрометеорологический прибор’	24	17
барометр$_1$ ‘индикатор’	24	5
болячка$_0$ ‘заболевание’	21	9
болячка$_1$ ‘рана’	21	9
графит$_0$ ‘минерал’	24	5
графит$_1$ ‘стержень’	17	6
дичь$_0$ ‘охотничье-промысловые животные’	31	7
дичь$_1$ ‘вздор, ерунда’	17	9
зайчик$_0$ ‘солнечный зайчик’	11	6
зайчик$_1$ ‘заяц’	14	5
зародыш$_0$ ‘зародыш организма’	18	6
зародыш$_1$ ‘росток, первые признаки’	10	7
калейдоскоп$_0$ ‘круговорот событий’	12	6
калейдоскоп$_1$ ‘детский калейдоскоп’	14	6
колыбель$_0$ ‘колыбель для младенца’	16	5
колыбель$_1$ ‘родина возникновения’	13	6
колокольчик$_0$ ‘травянистое растение’	15	5
колокольчик$_1$ ‘звонок’	16	7

Сбалансированная коллекция, собранная с помощью метода однозначных «родственников», использовалась для обучения трех моделей; среднее количество обучающих примеров для каждого значения составило 706. Две модели использовали уже обученную языковую модель ruBERT от DeepPavlov, а другая базировалась 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*) базировалась на идеях, описанных в [Huang et al., 2019] и [Kohli, 2021]: с помощью модели ruBERT решалась задача классификации пар предложений, которые были представлены контекстом с ключевым многозначным словом и словарной дефиницией одного из его значений. Два предложения разделялись в обучающей выборке с помощью специального вспомогательного символа “[SEP]” из модели BERT. В самое начало каждого словарного определения (или примера употребления многозначного слова) ставилось многозначное слово, которое выступало в качестве сигнала для модели разрешения неоднозначности о том, какое целевое многозначное слово рассматривалось в данном примере. Каждая пара контекст-толкование помечалась как положительный класс (метка 1), если словарное определение соответствовало значению многозначного слова в данном контексте, все прочие пары имели метку 0. Эксперименты, описанные в разделе 3.1 и работе [Kutuzov, Kuzmenko, 2019], показали, что для решения задачи разрешения лексической неоднозначности в русском языке лучше подходят лемматизированные коллекции. По этой причине во всех моделях использовались обучающие и тестовые данные, в которых все слова были приведены к нормальной форме. Примеры лемматизированных пар контекст-толкование из автоматически размеченной обучающей коллекции приведены в таблице 20.

Таблица 20. Примеры из автоматически собранной обучающей коллекции, приведенные к формату необходимому для обучения модели context-gloss pair BERT.

Обучающий пример	**Метка**
кожа покраснение "болячка" припухлость круг глаз воспользоваться консилер [SEP] **болячка** : болезнь	0
кожа покраснение "болячка" припухлость круг глаз воспользоваться консилер [SEP] **болячка** : болезненный образование на тело	1
выключатель адреналин необходимый создание ингибитор "болячка" бета блокатор [SEP] **болячка** : болезненный образование на тело	0
выключатель адреналин необходимый создание ингибитор "болячка" бета блокатор [SEP] **болячка** : болезнь	1

Также, как и в работе [Kutuzov, Kuzmenko, 2019], в настоящем исследовании для предсказаний значений слов применялась логистическая регрессия, принимавшая на вход контекстуализированные представления слов из языковой модели ELMo в качестве признаков. Результаты предыдущих экспериментов, продемонстрированные в разделе 3.1, показали, что наиболее эффективный способ применения векторов слов из данной языковой модели для разрешения неоднозначности – это использовать вектор только для ключевого многозначного слова в контексте, взятый с самого последнего слоя предварительно обученной модели. В описываемых экспериментах также была реализована эта стратегия.

Следует отметить, что в текущем исследовании изучалась производительность одноязычных моделей разрешения неоднозначности, обученных на автоматически размеченных и псевдоаннотированных обучающих коллекциях, поэтому мультиязычные языковые модели не использовались. Несмотря на это, результаты, полученные в описанных экспериментах, могут быть использованы в качестве основы для сравнения в будущих работах по разрешению неоднозначности.

3.3.2. Метод порождения псевдоразметки текстов на основе ансамбля моделей

Как уже говорилось ранее, автоматически размеченные данные могут содержать то или иное количество «шума», из-за которого обученная на них система может допускать ошибки при классификации. В описываемых экспериментах для того, чтобы снизить влияние этого фактора на результирующие предсказания моделей, обученных в парадигме слабо контролируемого обучения, было решено использовать их в ансамбле, т.е. при вычислении финальной метки значения для примера из тестовой выборки учитывались вероятностные оценки всех трех систем. Помимо этого, в данном исследовании принимался во внимание уровень «уверенности» модели в том или ином предсказании.

Таким образом, алгоритм порождения псевдоразметки для неаннотированного корпуса состоял из следующих шагов: генерация обучающих данных с помощью разработанного в рамках диссертационного исследования метода однозначных родственных слов; обучение трех моделей на этой коллекции; предсказание меток значений для многозначных слов с помощью ансамбля моделей. Затем тексты с такой разметкой использовались для повторного обучения имеющихся систем разрешения неоднозначности. Подобная процедура дообучения моделей схожа с алгоритмом бутстрэппинга, о котором говорилось в разделе 1.5.4. Общая схема описанного процесса представлена на рисунке 4.

В экспериментах также учитывалась степень «уверенности» каждой из моделей. Прежде всего, для каждого классификатора определялся диапазон вероятностей, в котором он совершал наибольшее число ошибок. Эти пороговые значения вычислялись с помощью тестового набора данных и помогали фильтровать неточные прогнозы классификаторов. Вероятностные предсказания логистической регрессии и модели fine-tuned ruBERT на тестовой выборке анализировались для выявления областей, где модели совершают наибольшее

число ошибок. На рисунке 5 область, где модель допускала ошибки, выделена малиновым цветом. Таким образом, в случае данной модели вероятностный интервал от 0.45 до 0.6 будет исключаться при рассмотрении полученных предсказаний.

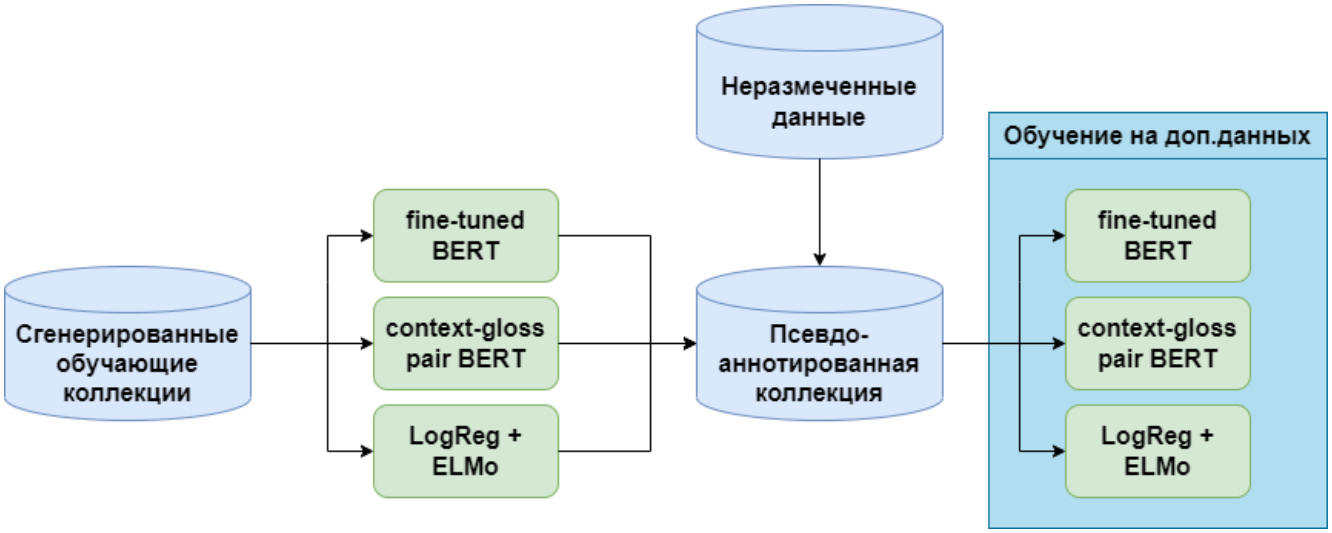

Рисунок 4. Схема эксперимента по порождению псевдоразметки и ее валидации.

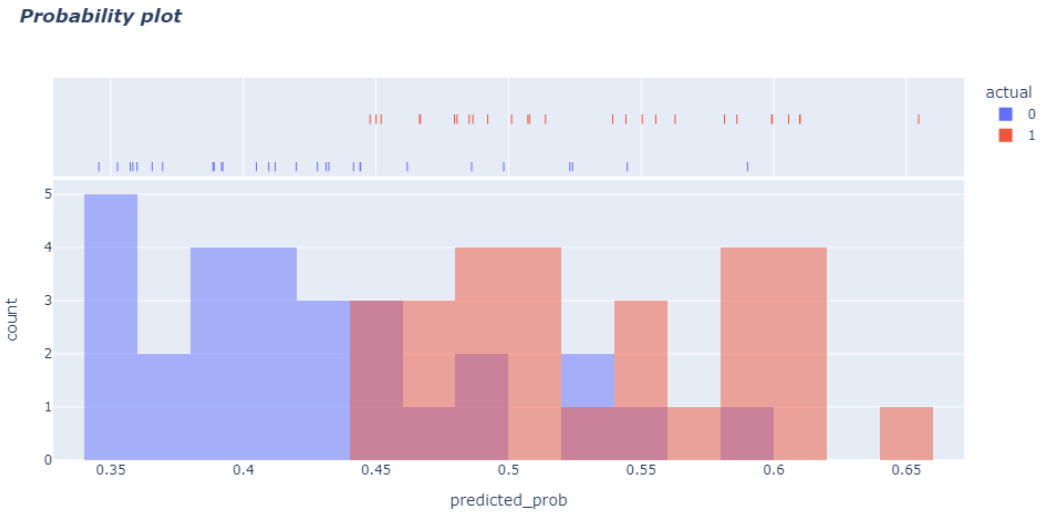

Рисунок 5. Вероятности для примеров со словом *аниматор* из тестовой выборки, предсказанные с помощью модели логистической регрессии.

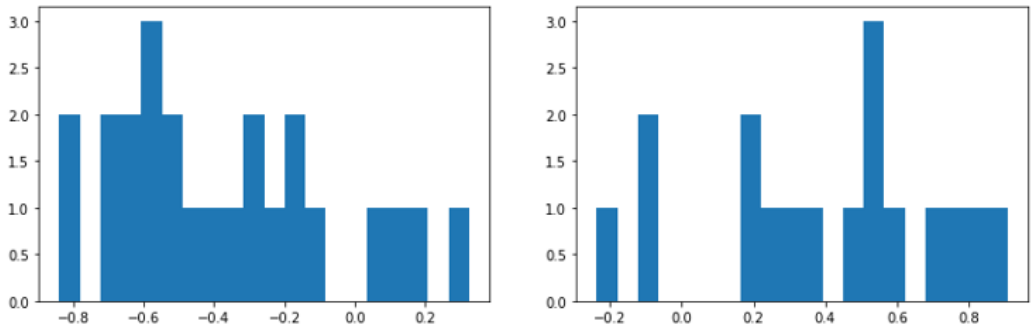

Рисунок 6. Различия в вероятностях, предсказанных моделью context-gloss pair BERT для слова *графит* в значениях ‘минерал’ и ‘стержень’, соответственно.

В случае модели context-gloss pair BERT для каждого значения целевого слова рассматривалась разница между вероятностями правильного и неправильного класса, предсказанными системой для примеров из тестовой выборки. Если эта разница была больше 0, значит модель предсказала правильную метку. В данном эксперименте 0,25-квантиль положительных значений разницы считался порогом «уверенности» для модели context-gloss pair BERT. На рисунке 6 представлены гистограммы различий в вероятностях предсказаний для двух значений слова *графит*. В соответствии с данными, продемонстрированными на этих графиках, пороговое значение для первого значения слова *графит* (‘минерал’) составило 0.1, а для второго (‘стержень’) – 0.3. Рассмотрим применение этого критерия на примере: классификатор предсказал значение ‘минерал’, а разница между вероятностью этой метки и вероятностью значения ‘стержень’ получилась меньше 0.1, поэтому такой прогноз модели не будет учитываться в ансамбле.

Чтобы получить окончательную метку класса из вероятностей, предсказанных моделями и соответствующих заранее заданным требованиям, к выходным данным моделей применялась весовая функция. Существуют различные схемы взвешивания вероятностных оценок, и в данном эксперименте применялась такая, где «вместо использования одного общего веса w_j для всех предсказаний каждому классу присваивается отдельный вес w_{ij}. Этот вес определяется как доля корректно предсказанных случаев для этого класса на

обучающих данных» [Large et al., 2019: 1678]. В разработанной в рамках диссертационного исследования 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 значения для примера с целевым словом. Эта схема оценки позволила учитывать только предсказания классификаторов с высокой степенью «уверенности».

В экспериментах применялись различные варианты разметки текстов. В некоторых коллекциях псевдоразметка корректировалась с помощью принципа «Одно значение на дискурс» (“One sense per discourse”) [Gale et al., 1992], который гласит следующее: «если многозначное слово встречается в дискурсе два или более раза, то с большей степенью вероятности все эти употребления имеют одно и то же значение». Из новостного корпуса и новостных сегментов «Тайги» извлекались все тексты с целевыми многозначными словами, в которых они встречались хотя бы 2 раза. С помощью алгоритма порождения псевдоразметки, описанного выше, для каждого неоднозначного слова предсказывалось его значение. Если принцип «Одно значение на дискурс» применялся, то финальная метка класса для всех примеров с тем или иным многозначным словом в рамках одного текста выбиралась в соответствии с большинством голосов: слову приписывалась та метка класса, которая была предсказана ансамблем моделей более, чем половине всех контекстов. Если наиболее частотное значение нельзя было определить, то значение многозначного слова определялось классом, у которого среднее значение предсказанных вероятностей максимальное среди всех имеющихся у слова значений. Если принцип «Одно значение на дискурс» не использовался, то каждому примеру с неоднозначным словом приписывалась та метка, которая была предсказана ансамблем моделей. Примеры предложений с псевдоразметкой приведены в таблице 21.

Таблица 21. Примеры предложений с псевдоразметкой.

Псевдометка	**Контекст**	**Корректность метки**
барометр ‘индикатор’	Обеспечение продовольственной безопасности служит **барометром** эффективности деятельности государственных учреждений.	✓
барометр ‘гидрометеорологический прибор’	Средством ведения метеорологических наблюдений является ртутный **барометр**.	✓
графит ‘минерал’	Оголенная древесина на коническом кончике потемнела, приобрела свинцово-сливовый оттенок, слившись с тупым мыском **графита**, чей слеповатый лоск один только и отличал его от дерева.	✗
графит ‘стержень’	Но прорыв в атомной отрасли был омрачен проблемами на реакторах типа РБМК. Изменение геометрии графитовой кладки (в этих реакторах в качестве замедлителя используется **графит**) началось "раньше прогнозируемого".	✗

Помимо этого, в данном эксперименте применялись различные подходы к дополнению текстовых данных. Максимальное число примеров, размеченное ансамблем моделей, составило 804 примера (всего для двух значений слова *колокольчик*), что является недостаточным для обучения моделей, основанных на языковой модели BERT. В дополнение к расширению псевдоаннотированной обучающей коллекции с помощью словарных дефиниций и примеров употреблений слов применялись также техники, описанные в работе [Wei, Zou, 2019]: замена случайных слов в предложении на синонимы, вставка синонима одного из слов в контексте на случайную позицию, случайная перестановка двух слов в предложении, удаление случайного слова. Данные методы были специально адаптированы для задачи разрешения неоднозначности в русском языке: было добавлено извлечение синонимов из RuWordNet, а также введено

ограничение на трансформации текстов, включающие целевое многозначное слово (например, запрет на его удаление или перемещение). С помощью подобных трансформаций текстов удалось увеличить псевдоаннотированную выборку для моделей, использующих языковую модель BERT, в 6 раз.

3.3.3. Результаты и выводы

В данном разделе представлены результаты оценки моделей, обученных на псевдоаннотированных данных и коллекциях, сгенерированных методом однозначных родственных слов. Стоит отметить, что оценка на тестовом наборе данных проводилась с учетом двух различных размеров контекстного окна: *win=1* подразумевает, что в качестве контекста использовалось одно предложение до и после того предложения, в котором содержалось многозначное слово; *win=0* означает, что учитывалось только предложение с многозначным словом. Усредненные значения F1-меры различных моделей для всех ключевых многозначных слов на тестовом наборе данных представлены в таблице 22. Более детальные метрики моделей по каждому многозначному слову представлены в таблицах 23, 24 и 25. Во всех таблицах меткой (1) обозначается логистическая регрессия, (2) – fine-tuned BERT; (3) – Context-gloss pair BERT. В ходе эксперимента были оценены три типа моделей, обученных на четырех разных вариантах псевдоаннотированных данных: (а) – обучающая коллекция, собранная без учета принципа «Одно значение на дискурс» и без дополнения данными из словарей; (б) – обучающая коллекция, собранная без учета принципа «Одно значение на дискурс» и с дополнением данными из словарей; (в) – обучающая коллекция, собранная с учетом принципа «Одно значение на дискурс» и без дополнения данными из словарей; (г) – обучающая коллекция, собранная с учетом принципа «Одно значение на дискурс» и с дополнением данными из словарей.

Таблица 22. 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 данных, 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

Таблица 23. Результаты классификации моделей разрешения неоднозначности, обученных на данных, размеченных с помощью метода однозначных родственных слов.

Модель / **Целевое многозначное слово**	**(1)**	**(2)**	**(3)**
аниматор win=1	0.73	0.68	0.75
аниматор win=0	**0.76**	0.64	0.67
барометр win=1	**0.96**	0.87	0.84
барометр win=0	0.94	0.77	0.84
болячка win=1	0.72	0.68	**0.77**
болячка win=0	0.76	0.7	**0.77**
графит win=1	0.57	0.61	0.56
графит win=0	0.62	**0.63**	0.59
дичь win=1	**0.97**	0.94	0.98
дичь win=0	0.87	0.94	0.92
зайчик win=1	0.92	0.94	0.78
зайчик win=0	**1**	0.88	0.78
зародыш win=1	**0.95**	**0.95**	0.83
зародыш win=0	0.82	0.8	0.76
калейдоскоп win=1	0.84	0.77	0.76
калейдоскоп win=0	**0.88**	0.82	0.73
колыбель win=1	**0.93**	0.86	0.77
колыбель win=0	0.9	0.87	0.84
колокольчик win=1	**0.97**	0.9	0.93
колокольчик win=0	0.93	0.9	**0.97**

Таблица 24. Результаты классификации моделей разрешения неоднозначности, обученных на псевдоаннотированных данных, размеченных без использования принципа «Одно значение на дискурс» (F1-мера).

Модель / **Целевое многозначное слово**	**(а)**			**(б)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
аниматор win=1	0.74	0.67	0.77	0.73	0.67	**0.76**
аниматор win=0	0.75	0.67	0.73	**0.76**	0.67	0.73
барометр win=1	**0.96**	0.88	0.88	**0.96**	0.93	0.88
барометр win=0	0.94	0.88	0.89	0.94	0.88	0.89
болячка win=1	0.7	0.68	0.72	0.7	0.68	0.68
болячка win=0	0.75	0.76	**0.77**	0.75	0.76	**0.77**
графит win=1	0.6	0.7	**0.76**	0.62	0.67	0.74
графит win=0	0.63	0.68	**0.76**	0.63	0.72	0.7
дичь win=1	0.97	0.97	0.96	0.97	**1**	0.96
дичь win=0	0.87	0.86	0.95	0.87	0.88	0.95
зайчик win=1	0.97	0.96	0.98	0.97	0.97	0.98
зайчик win=0	**1**	0.96	0.94	**1**	0.96	0.96
зародыш win=1	0.95	**1**	0.96	0.95	0.95	0.96
зародыш win=0	0.82	0.89	0.88	0.82	0.9	0.88
калейдоскоп win=1	0.82	0.84	0.82	0.82	0.87	0.82
калейдоскоп win=0	**0.9**	0.82	0.84	**0.9**	0.82	0.84
колыбель win=1	**0.93**	0.86	**0.93**	**0.93**	0.87	**0.93**
колыбель win=0	0.9	0.87	0.9	0.9	0.87	0.9
колокольчик win=1	**0.97**	**0.97**	0.95	**0.97**	**0.97**	**0.97**
колокольчик win=0	0.94	0.93	0.95	0.94	0.93	**0.97**
усредненная F1-мера для win=1	0.86	0.85	**0.87**	0.86	0.86	**0.87**
усредненная F1-мера для win=0	0.85	0.83	0.86	0.85	0.84	0.86

Таблица 25. Результаты классификации моделей разрешения неоднозначности, обученных на псевдоаннотированных данных, размеченных с использованием принципа «Одно значение на дискурс» (F1-мера).

Модель / **Целевое многозначное слово**	**(в)**			**(г)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
аниматор win=1	0.74	0.7	0.78	0.79	**0.89**	0.85
аниматор win=0	0.8	0.7	0.77	0.79	0.8	0.77
барометр win=1	0.96	0.93	0.88	**0.98**	0.93	0.88
барометр win=0	0.94	0.91	0.87	0.92	0.94	0.9
болячка win=1	0.73	0.69	0.61	0.76	0.68	0.7
болячка win=0	0.76	0.69	0.74	**0.77**	0.73	0.75
графит win=1	0.6	0.59	0.78	0.65	**0.79**	0.74
графит win=0	0.65	0.59	0.72	0.63	0.74	0.74
дичь win=1	**0.97**	**0.97**	0.96	0.97	**0.97**	0.96
дичь win=0	0.9	0.9	0.95	0.87	**0.97**	0.95
зайчик win=1	**1**	**1**	0.96	**1**	**1**	0.96
зайчик win=0	**1**	**1**	0.98	**1**	**1**	0.98
зародыш win=1	0.95	**1**	0.96	0.95	**1**	0.96
зародыш win=0	0.86	0.95	0.88	0.86	0.95	0.92
калейдоскоп win=1	0.85	0.74	0.79	0.85	0.74	0.79
калейдоскоп win=0	**0.88**	0.74	0.84	**0.88**	0.77	0.87
колыбель win=1	**0.93**	0.87	0.86	**0.93**	0.87	0.89
колыбель win=0	0.9	0.9	**0.93**	0.9	0.9	**0.93**
колокольчик win=1	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
колокольчик win=0	0.94	0.93	**0.97**	0.94	**0.97**	**0.97**
усредненная F1-мера для win=1	0.87	0.85	0.86	**0.89**	**0.89**	0.87
усредненная F1-мера для win=0	0.86	0.83	0.87	0.86	**0.88**	**0.88**

Результаты моделей, дообученных на новых текстах, которые были размечены ансамблями, показывают, что эта процедура улучшает качество моделей разрешения неоднозначности. В некоторых случая дополнительное обучение сильно повышает F1-меру, например, максимальное значение F1 для слова *аниматор* (win=1) при обучении модели context-gloss pair BERT на данных, сгенерированных методом однозначных родственных слов, составляло 0.75,

однако после дообучения классификатора на псевдоаннотированных данных метрика повысилась до 0.85. Иногда эффект от обучения на дополнительных данных менее выражен, например, результаты классификации логистической регрессии для слова *барометр* не сильно меняются при различных модификациях обучающих коллекций. Однако иногда наблюдались такие случаи, когда модели, обученные на псевдоаннотированных корпусах, показывали результаты хуже, чем оригинальные модели, где тексты были размечены методом однозначных родственных слов: например, F1 логистической регрессии для слова *болячка*, которая была дополнительно обучена на данных без использования принципа «Одно значение на дискурс», ниже, чем у первоначальной модели.

В большинстве случаев, если при разметке текстов с псевдометками учитывался принцип «Одно значение на дискурс», то качество разрешения неоднозначности моделей, которые были на них обучены, выше. Исключение составляют лишь некоторые случаи, например, системы, использовавшие языковую модель BERT для слова *калейдоскоп*.

Что касается дополнения обучающих данных словарными дефинициями и примерами употребления многозначных слов, данные демонстрируют, что дополнительные примеры либо никак не влияют на результирующее качество системы, либо улучшают его. Таким образом, лексикографические данные могут служить хорошим и легко доступным способом расширения размеченной текстовой коллекции. Полученные метрики также свидетельствуют о том, что размер контекстного окна оказывает различное влияние на качество предсказаний отдельных слов, однако усредненные значения F1 выше в тех случаях, где для предсказания использовался более широкий контекст. Кроме того, полученные данные не позволяют сказать, какая из моделей лучше подходит для задачи разрешения неоднозначности: иногда логистическая регрессия, базирующая на представлениях слов из ELMo, показывает самый высокий результат (*барометр* win=1, F1=0.98), иногда fine-tuned BERT (*аниматор* win=1, F1=0.89), а иногда context-gloss pair BERT (*колыбель* win=0, F1=0.93).

Эксперименты также показали, что все слова отличаются по степени сложности снятия неоднозначности. Для некоторых слов метрика F1 высокая, что говорит о том, что их значения легко разделимы, например, слово *зайчик*. F1-мера для слова *графит* достаточно низкая по сравнению с другими словами в выборке, что говорит о сложности различения значений этого слова. Возможно, это связано с метонимическим типом связи между двумя значениями – это связь, основанная на отношении «материал — предмет, сделанный из него».

Проведенные эксперименты доказали, что данные, автоматически размеченные с помощью метода однозначных родственных слов, могут использоваться для генерации псевдоразметки. Также было показано, что дополнительное обучение моделей на псевдоаннотированных данных позволяет улучшить качество разрешения неоднозначности. Предложенная стратегия взвешивания предсказаний из ансамбля моделей может быть использована в качестве вспомогательного средства для ручной семантической аннотации или для активного обучения.

3.4. Визуализация контекстуализированных представлений примеров из обучающей коллекции

В данном разделе представлен сравнительный анализ векторных представлений ключевых слов, получаемых из контекстов автоматически сгенерированной обучающей коллекции, представлений слов из оценочного набора данных RUSSE-RuWordNet, размеченного вручную, и представлений слов из корпуса, применяемого в базовом решении. Контекстуализированные представления также, как и ранее, извлекались из модели ELMo от RusVectōrēs. Все вектора, извлеченные из обучающей и тестовой коллекции, а также коллекции, использованной в базовом решении, были визуализированы с помощью алгоритма t-SNE [Van der Maaten, Hinton, 2008], чтобы изучить, как они расположены относительно друг друга в векторном пространстве.

Визуализации многозначных слов из обучающей коллекции продемонстрировали, что большинство значений формируют легко разделимые смысловые кластеры. Рисунки 7 и 8 показывают контекстуализированные представления, полученные для двух значений существительных – *акция* и *крона*. На них видно, что примеры с одинаковым значением ключевого слова группируются вместе. Однако не все слова имеют подобные четко выделяемые смысловые группы. Например, значения слова *гвоздика* перемешаны и не имеют никакой явной границы между ними, что демонстрирует рисунок 9.

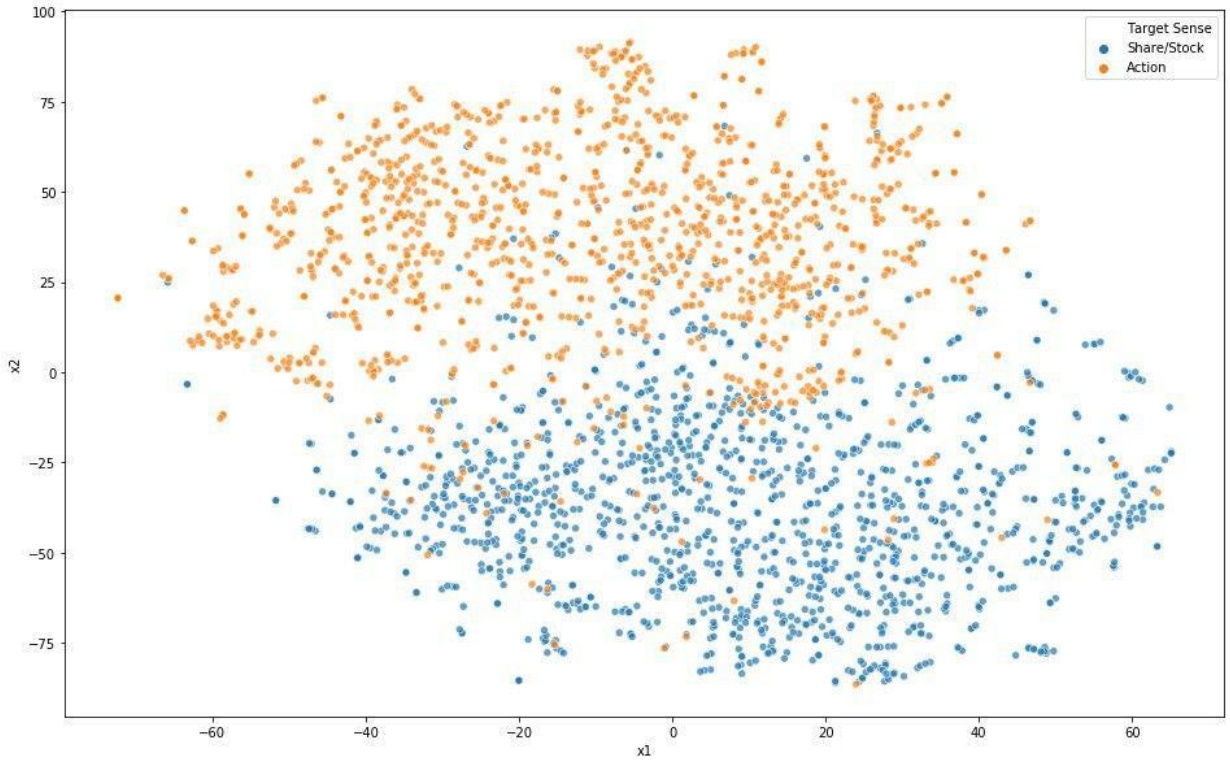

Рисунок 7. Представления для слова *акция*, извлеченные из RusVectōrēs ELMo модели, контексты были взяты из автоматически сгенерированный обучающей коллекции; визуализировано с помощью t-SNE.

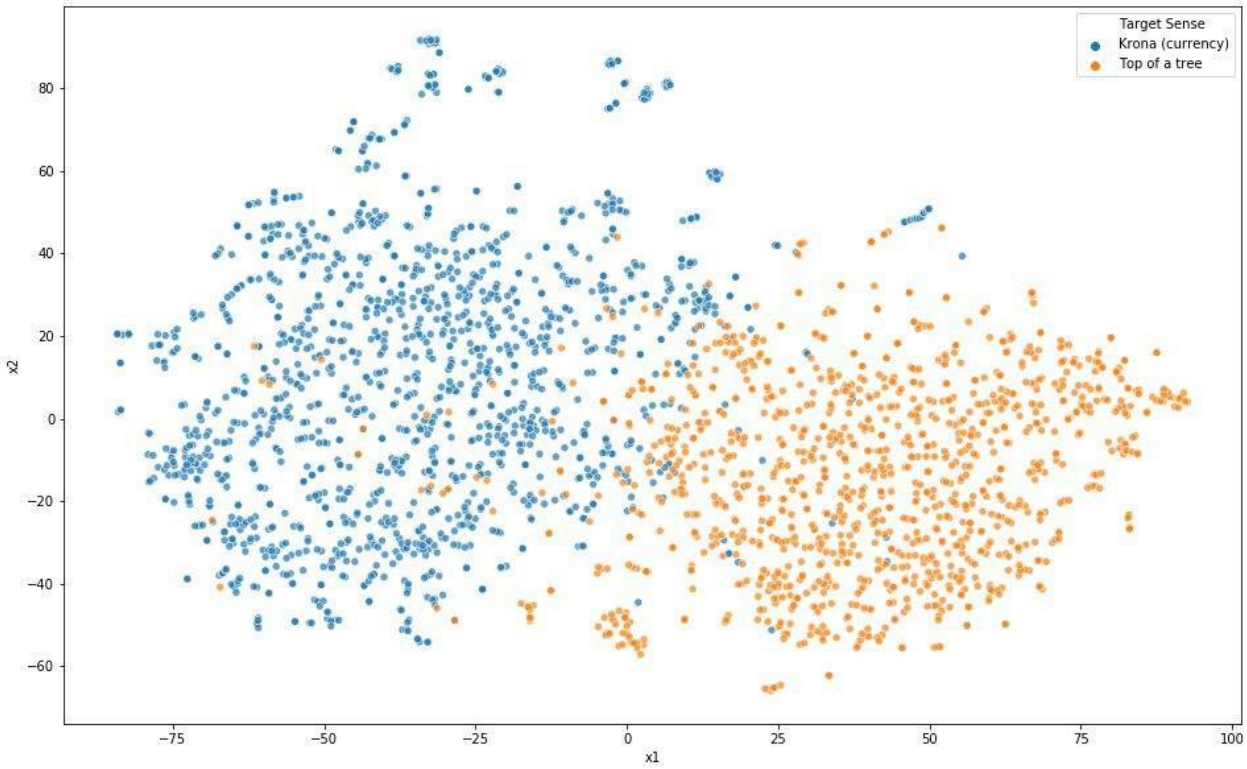

Рисунок 8. Представления для слова *крона*, извлеченные из RusVectōrēs ELMo модели, контексты были взяты из автоматически сгенерированный обучающей коллекции; визуализировано с помощью t-SNE.

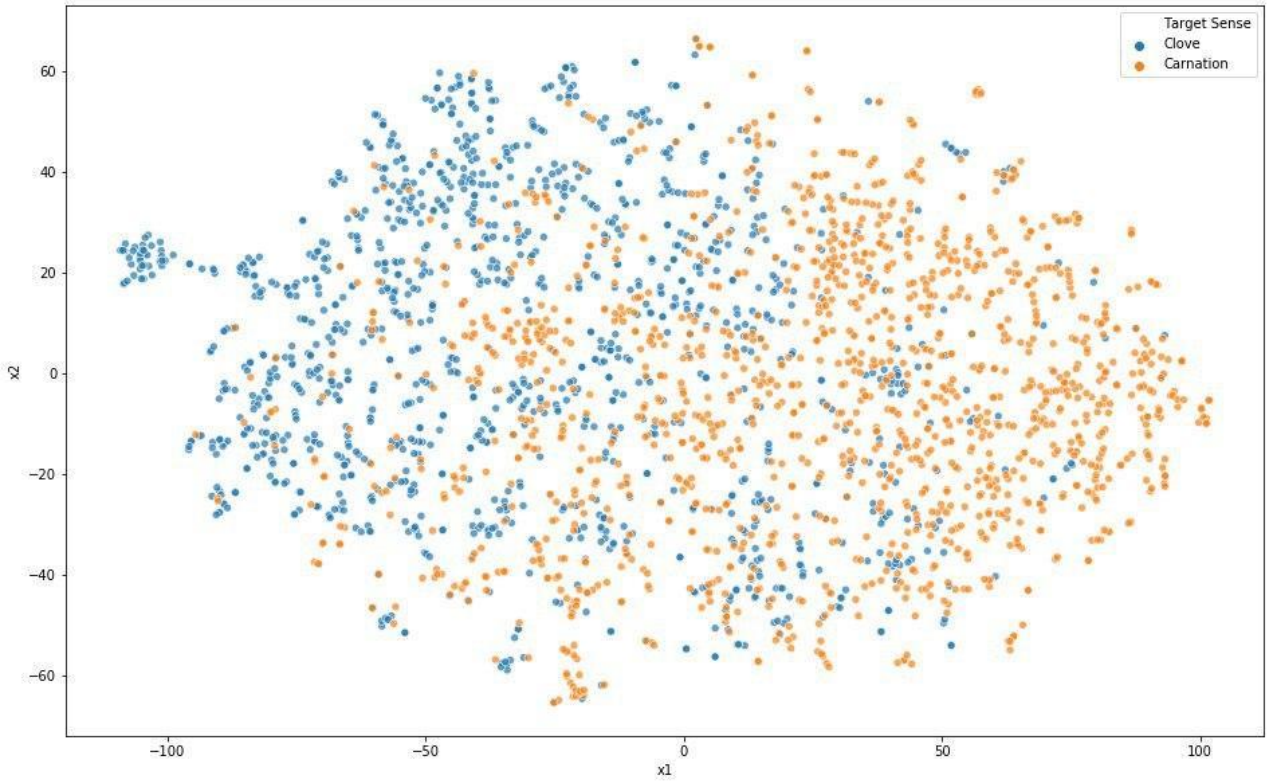

Рисунок 9. Представления для слова *гвоздика*, извлеченные из RusVectōrēs ELMo модели, контексты были взяты из автоматически сгенерированный обучающей коллекции; визуализировано с помощью t-SNE.

Сравнение контекстуализированных представлений, извлеченных из автоматически сгенерированной обучающей коллекции и из вручную размеченного корпуса, показало, что в некоторых случаях распределения значений отличались друг от друга, то есть группы значений занимали разные части векторного пространства. В то же время, были обнаружены случаи, когда кластеры значений из обучающей коллекции совпадали с кластерами из оценочного набора данных. На рисунке 10 изображены представления для слова *таз* в его двух значениях, которые были извлечены из обучающей коллекции (обозначены меткой “_train”) и из тестовой выборки (обозначены меткой “_test”).

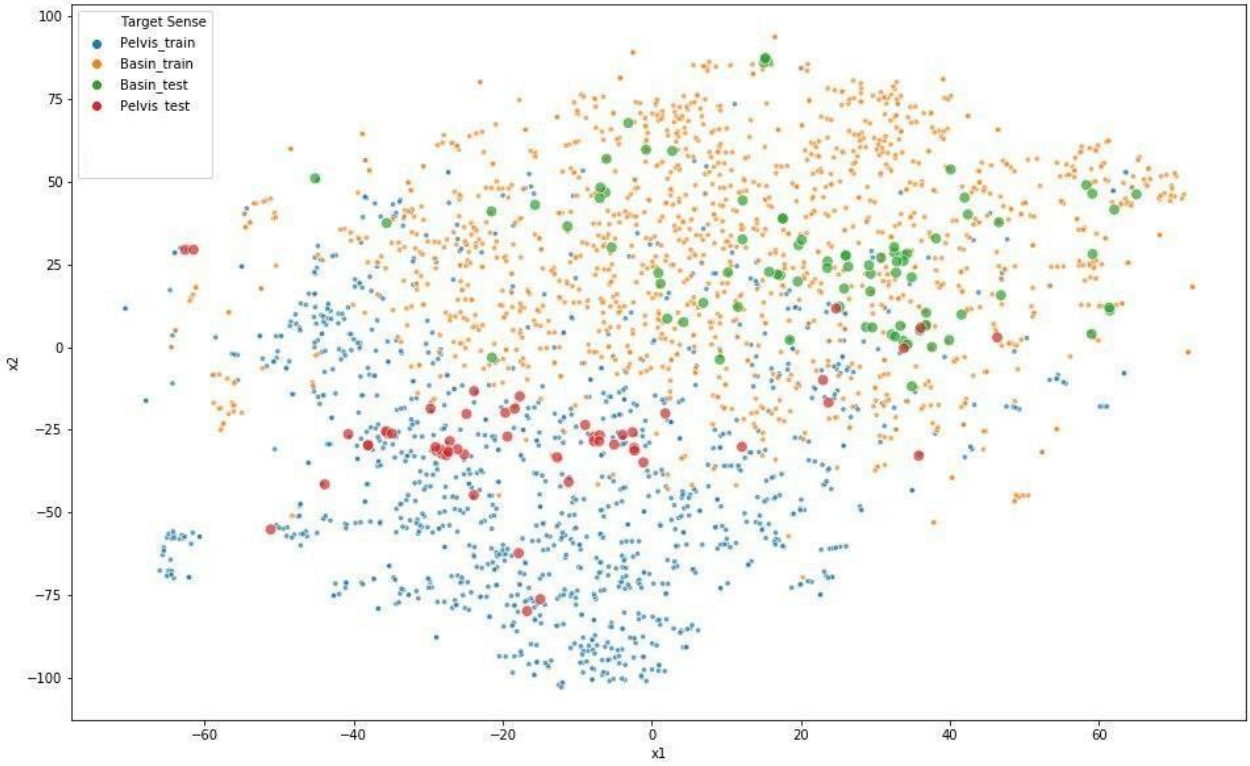

Рисунок 10. Представления для слова *таз*, извлеченные из RusVectōrēs ELMo модели; значения, маркированные символом “_train”, были взяты из автоматически сгенерированной обучающей коллекции; значения, маркированные символом “_test” были взяты из вручную размеченной.

Рисунки 11 и 12 показывают, что все примеры для одного и того же значения занимают одну и ту же часть векторного пространства. Примеры из словарного корпуса расположены у границы смысловых кластеров и в случае новостной обучающей коллекции, и в случае обучающей коллекции Проза.ру. Но

подобная конфигурация характерна не для каждого ключевого многозначного слова.

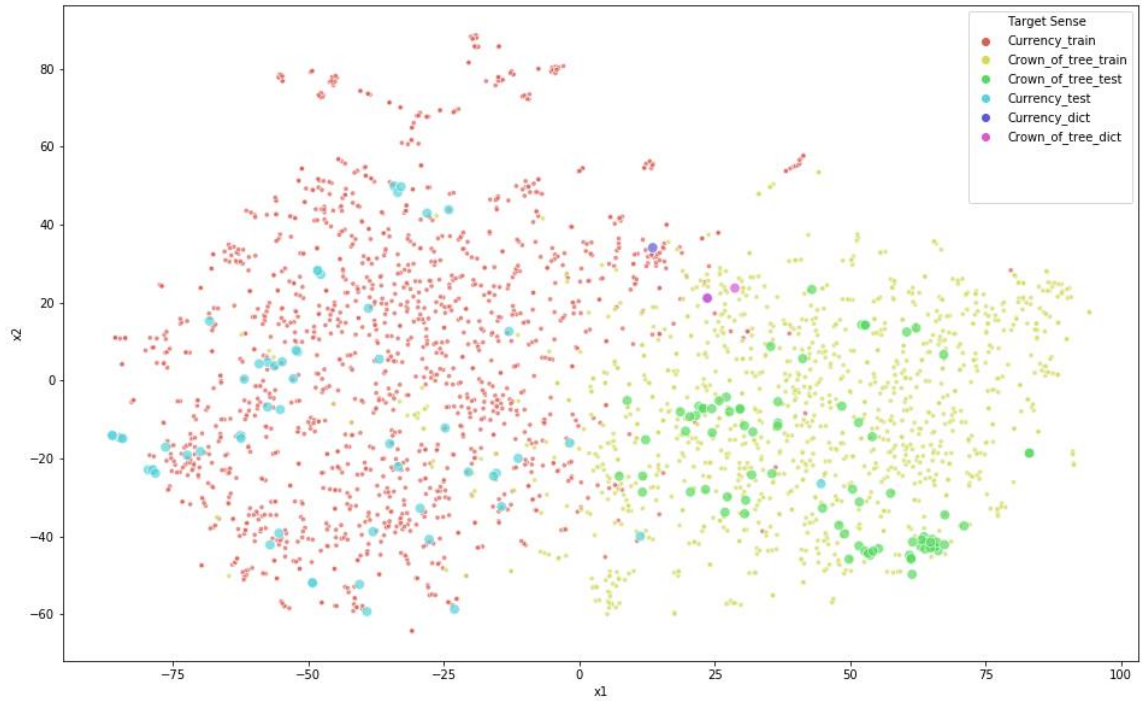

Рисунок 11. Представления для слова *крона*, извлеченные из RusVectōrēs ELMo модели; примеры, отмеченные тегом “_train”, были взяты из новостной обучающей коллекции (сбалансированной); примеры, отмеченные тегом “_test”, были взяты из тестовой выборки, размеченной вручную; примеры, отмеченные тегом “_dict”, были взяты из корпуса со словарными дефинициями и примерами употреблений.

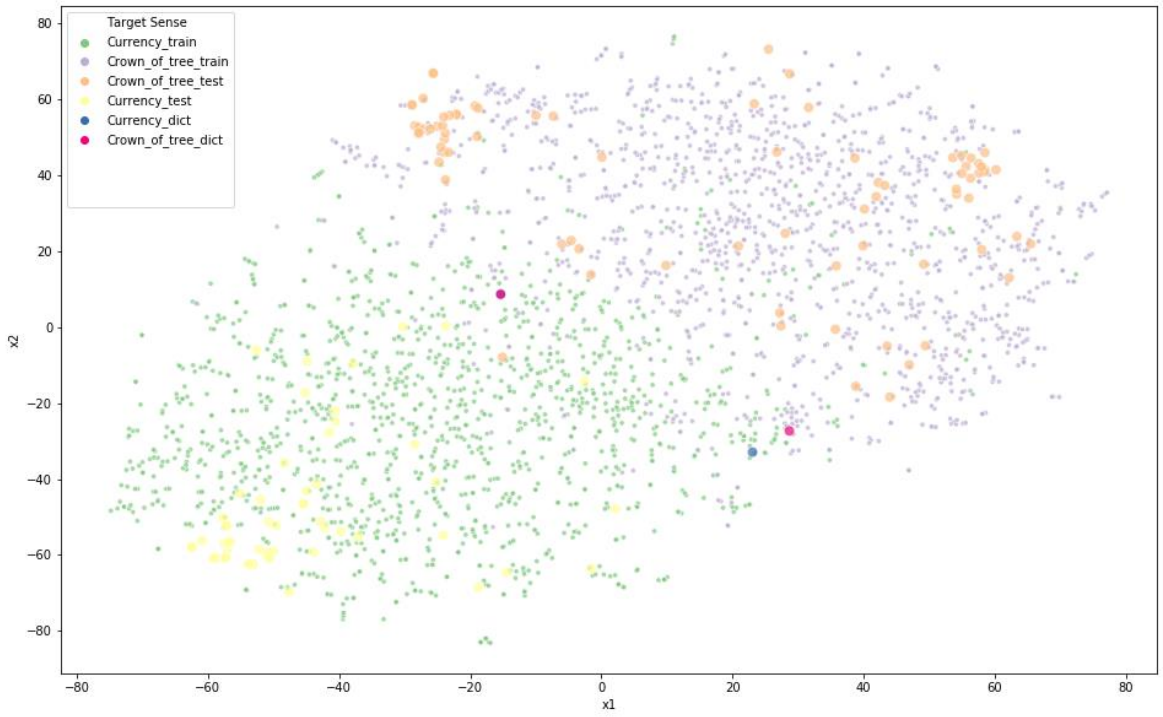

Рисунок 12. Представления для слова *крона*, извлеченные из RusVectōrēs ELMo модели; примеры, отмеченные тегом "_train", были взяты из обучающей коллекции (сбалансированной) Проза.ру; примеры, отмеченные тегом "_test", были взяты из тестовой выборки, размеченной вручную; примеры, отмеченные тегом "_dict", были взяты из корпуса со словарными дефинициями и примерами употреблений.

3.5. Задача Word-in-Context

Задача Word-in-Context (WiC) была представлена в статье [Pilehvar, Camacho-Collados, 2019] и частично напоминает задачу разрешения неоднозначности, но формулируется как задача бинарной классификации. 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.

Эксперимент, представленный в данном разделе, направлен на то, чтобы выяснить может ли коллекция, автоматически размеченная с помощью метода однозначных родственных слов, использоваться как обучающие данные для задачи Word-in-Context. Результаты исследования применимости метода

однозначных слов для решения задачи WiC представлены в нашей статье [Bolshina, Loukachevitch, 2020b].

Для оценки использовался набор данных для задачи WiC из проекта по оценке языковых моделей на задачах понимания текстов Russian SuperGLUE[32] [Shavrina et al., 2020]. Набор состоит из трех компонентов: обучающей, тестовой и валидационной выборки. Тестовая коллекция не содержит целевых меток, потому что она используется только для того, чтобы делать на ней предсказания, которые потом будут автоматически верифицироваться системой. Таким образом, в описываемом эксперименте применялся только аннотированный обучающий корпус из этого проекта. Стоит отметить, что инвентарь значений, используемый в этом наборе данных, отличается от того, что используется в настоящем исследовании, поэтому значения слов из Russian SuperGLUE были вручную приведены в соответствие со значениями, которые применяются в настоящей работе. В результате получилось, что не все многозначные слова и их значения из оригинального обучающего корпуса были включены в финальный набор данных, который в текущем эксперименте используется для оценки качества моделей. В общей сложности в собранном тестовом наборе данных содержалось 161 многозначное слово и 7006 пар контекстов. Сбалансированная обучающая коллекция для этой задачи была собрана с помощью метода однозначных родственных слов, всего в ней было 9105 примеров.

Для классификации данных в настоящем исследовании 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. Так как для оценки этой задачи в проекте Russian SuperGLUE использовалась метрика accuracy, в данном исследовании было принято решение также ее использовать. Стоит отметить, что дисбаланса классов в тестовой выборке не было, поэтому данную оценку применять можно.

[32] https://russiansuperglue.com/

Accuracy классификатора, обученного на автоматически размеченной выборке, составила **0.91**. Для того, чтобы сравнить качество моделей, обученных на автоматически и вручную размеченных контекстах, аннотированные обучающие примеры из проекта Russian SuperGLUE (на которых тестировалась предыдущая модель) были использованы для обучения классификатора. На этих данных была проведена 5-кратная кросс-валидация, а результирующее качество предсказаний считалось как среднее пяти значений метрики accuracy, полученных с помощью этих классификаторов, и 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.

Выводы к главе 3

Изучение применимости обучающих коллекций, размеченных с помощью метода однозначных родственных слов, для решения различных лексико-семантических задач, показало, что разработанный в рамках диссертационного исследования подход может успешно использоваться для решения задачи разрешения лексической неоднозначности (для всех частей речи), для задачи WiC и для порождения псевдоразметки.

Сравнительный анализ работы различных систем снятия неоднозначности на наборе данных RUSSE-RuWordNet позволил выявить, что kNN-классификатор достигает наилучшего качество предсказаний с помощью контекстуализированных представлений слов из языковой модели ELMo от RusVectōrēs, обученной на лемматизированном корпусе «Тайга». В дополнение к этому, проведенные эксперименты показали, что модели разрешения неоднозначности, обученные на автоматически размеченных коллекциях, работают лучше на лемматизированных текстах, и качество предсказаний модели зависит от того, совпадают ли жанры текстов обучающей и тестовой коллекций.

Для того, чтобы оценить качество обучающей коллекции, созданной для всех имеющихся в тезаурусе RuWordNet частей речи, в полуавтоматическом режиме был размечен набор данных. Полученная модель показала хорошее качество разрешения неоднозначности для существительных, глаголов и прилагательных, что свидетельствует о том, что предложенный подход пригоден для масштабирования на различные части речи.

Также было доказано, что обучающие данные, аннотированные с помощью разработанного метода, подходят и для задачи WiC. Модель, обученная на автоматически сгенерированной коллекции, показала более высокие результаты по сравнению с моделью, обученной на вручную размеченных данных. И, наконец, автоматически аннотированные коллекции использовались в задаче генерации псевдоразметки текстовых корпусов для бутстрэппинга моделей разрешения неоднозначности. Системы, повторно обученные на псевдоаннотированных данных, улучшили метрики качества оригинальных моделей разрешения неоднозначности.

Заключение

В настоящей работе была изучена проблема недостатка размеченных данных для задачи автоматического разрешения лексической неоднозначности в русском языке. Основная ***цель*** диссертационного исследования состояла в том, чтобы разработать метод автоматической семантической разметки корпуса русского языка для задачи разрешения лексической многозначности, а также создать его программную реализацию. Для достижения поставленной цели в рамках настоящей работы были поставлены и решены следующие ***задачи***:

1) Выполнен анализ теоретических аспектов создания систем автоматической генерации размеченных обучающих коллекций для задачи разрешения лексической неоднозначности.

2) Реализован метод автоматической разметки текстовых коллекций с использованием информации об однозначных родственных словах из лексико-семантического ресурса для русского языка RuWordNet.

3) Разработан метод фильтрации примеров, автоматически размеченных с помощью информации об однозначных родственных словах, который обеспечивает их семантическую близость к целевому значению многозначного слова.

4) Проведена оценка корректности семантической разметки, полученной методом однозначных родственных слов, с помощью анализа качества разрешения неоднозначности моделей, обученных на коллекциях с автоматической разметкой.

5) Обучены несколько моделей разрешения лексической многозначности для русского языка с применением полученного размеченного обучающего множества.

В настоящем диссертационном исследовании были получены следующие основные ***результаты***:

1) Разработан **подход к автоматическому сбору и разметке корпуса** для решения задачи разрешения лексической неоднозначности на основе 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. Еще одним преимуществом данного метода по сравнению с остальными способами порождения автоматически размеченных коллекций состоит в том, что для его работы требуется только семантический граф и неразмеченный корпус. Помимо этого, с помощью предложенного подхода можно собирать обучающие коллекции любого объема, всё зависит лишь от размера неразмеченного корпуса, из которого берутся примеры с однозначными родственными словами. Разработанный метод также может успешно применяться для генерации размеченных данных для других задач, связанных с семантическим анализом текстов, например, задачи Word-in-Context.
2) Реализован **механизм 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. Все однозначные родственные слова ранжируются с помощью коэффициента схожести однозначного «родственника-кандидата» и синсетов, близких к целевому значению многозначного слова. Подобный рейтинг однозначных «родственников» удобно использовать в различных стратегиях сбора обучающих коллекций.

3) Разработаны и обучены **системы автоматического разрешения лексической многозначности для русского языка** на материале автоматически собранных корпусов. Данные модели хорошо обучаются на неточных метках значений и показывают качество разрешения неоднозначности на уровне с моделями, обученными на вручную размеченных корпусах. Были разработаны модели, обладающие различной архитектурой: метрические (kNN), линейные (логистическая регрессия) и на основе нейронных сетей (тонко настроенный BERT и BERT, использующийся для классификации пар предложений контекст/толкование).
4) 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. Например, было доказано, что метрика F1 выше у тех моделей, которые были обучены на лемматизированных текстах. Также результаты экспериментов показали, что для задачи разрешения неоднозначности лучше всего подходят языковая модель ELMo от RusVectōrēs, обученная на лемматизированном корпусе «Тайга», и RuBERT от DeepPavlov.

На основе указанных результатов и полученных данных были сформулированы положения, выносимые на защиту и приведенные выше в разделе «Введение».

Дальнейшие исследования могут быть связаны с изучением новых методов поиска однозначных родственных слов (например, с помощью метода лексических замен (*lexical substitution*)) и разработкой более сложного многокомпонентного механизма их фильтрации и ранжирования. Кроме того, за рамками настоящей работы остались методы генерации размеченных обучающих коллекций, использующие параллельные корпуса. С появлением всё более сложных и мощных архитектур нейронных сетей возрастает потребность в

больших объемах аннотированных текстовых коллекций. Направления исследований, посвященные автоматической разметке текстов, кажутся перспективными ввиду растущего интереса в академической сфере и индустрии к разработке синтетических обучающих наборов данных.

Список литературы

1. *Азарова, И. В.* Автоматическое разрешение лексической неоднозначности частотных существительных (в терминах структурных единиц RussNet) [Текст] / И. В. Азарова, С. В. Бичинева, Вахитова Д. Т. // Труды Международной конференции «Корпусная лингвистика–2008». — 2008. — С. 5—8.
2. *Апресян, Ю. Д.* Избранные труды, том I. Лексическая семантика: 2-е изд., испр. и доп. [Текст] / Ю. Д. Апресян. — М.: Школа «Языки русской культуры», Издательская фирма «Восточная литература» РАН, 1995.
3. Активный словарь русского языка. Т. 1-2: А-Г [Текст] / Ю. Д. Апресян [и др.]; под ред. Ю. Д. Апресян. — М.: Языки славянской культуры, 2014.
4. Активный словарь русского языка (Т. 3, Д-З) [Текст] / В. Ю. Апресян [и др.]; под ред. В. Ю. Апресян, Б. Л. Иомдин, И. В. Галактионова. — М.: Издательство Нестор-История, 2017.
5. Интерактивное разрешение неоднозначности различных типов в машинном переводе [Текст] / И. М. Богуславский [и др.] // Труды международной конференции Диалог. — 2005.
6. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие [Текст] / Е. И. Большакова [и др.]. — М.: НИУ ВШЭ, 2017.
7. *Большина, А. С.* Методы автоматического формирования семантически размеченных корпусов [Текст] / А. С. Большина // Вестник Московского университета. Сер. 9. Филология. – 2022а. – № 2. – С. 173–183.
8. *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.

9. *Епрев, А. С.* Применение разрешения лексической многозначности в классификации текстовых документов [Текст] / А. С. Епрев // Машиностроение и компьютерные технологии. — 2010. — №10.
10. *Зализняк, А. А.* Феномен многозначности и способы его описания [Текст] / А. А. Зализняк // Вопросы языкознания. — 2004 — Т.2. — С. 20—45.
11. *Зализняк, А. А.* Многозначность в языке и способы ее представления [Текст] / А. А. Зализняк. — М.: Языки славянских культур, 2006.
12. *Иомдин, Б. Л.* Многозначные слова в контексте и вне контекста [Текст] / Б. Л. Иомдин // Вопросы языкознания. — 2014. — № 4. — С. 87—103.
13. *Кобозева, И. М.* Лингвистическая семантика: Учебник для студентов филологического профиля [Текст] / И. М. Кобозева. — М.В. Ломоносова. Филологический факультет. М.: Эдиториал УРСС, 2000.
14. *Кобрицов, Б. П.* Автоматическое разрешение семантической неоднозначности в Национальном корпусе русского языка [Текст] / Б. П. Кобрицов, О. Н. Ляшевская // Кобозева И. М., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог. — 2004.
15. *Кобрицов, Б. П.* Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе [Текст] / Б. П. Кобрицов, О. Н. Ляшевская, О. Ю. Шеманаева // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции “Диалог. — 2005. — С. 250—255.
16. *Кобрицов, Б. П.* Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных

толковых словарей [Текст] / Б. П. Кобрицов, О. Н. Ляшевская, С. Ю. Толдова // Электронная публикация: http://download.yandex.ru/IMAT2007/kobricov.pdf. 2007.

17. *Кустова, Г. И.* Типы производных значений и механизмы языкового расширения [Текст] / Г. И. Кустова. — М.: Языки славянской культуры, 2004.
18. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы [Текст] / Г. И. Кустова [и др.] // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — 2005. — С. 155—174.
19. *Лукашевич, Н. В.* Автоматическое разрешение лексической многозначности на базе тезаурусных знаний [Текст] / Н. В. Лукашевич, Д. С. Чуйко // Интернет-математика 2007.—Екатеринбург, 2007. — 2007.
20. *Лукашевич, Н. В.* Тезаурусы в задачах информационного поиска [Текст] / Н. В. Лукашевич. — М.: Издательство МГУ, 2011.
21. *Марчук, Ю. Н.* Контекстное разрешение лексической многозначности [Текст] / Ю. Н. Марчук // Вестник Московского государственного областного университета. Серия: Лингвистика. — 2016. — №1. — С.26—32.
22. *Митрофанова, О. А.* Статистическое разрешение лексико-семантической неоднозначности в контекстах для предметных имён существительных [Текст] / О. А. Митрофанова, О. Н. Ляшевская, П. В. Паничева // Компьютерная лингвистика и интеллектуальные технологии. — 2008. — Т. 7. — С. 368—375.
23. *Падучева, Е. В.* Динамические модели в семантике лексики [Текст] / Е. В. Падучева. — М.: Языки славянской культуры, 2004.
24. *Пашук, А. В.* Анализ методов разрешения лексической многозначности в области биомедицины [Текст] / А. В. Пашук, А. Б. Гуринович, Н. А.

Волорова, А. П. Кузнецов // Доклады БГУИР. — № 5(123). — 2019.

25. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка [Текст] / Е. В. Рахилина [и др.] // Компьютерная лингвистика и интеллектуальные технологии. — 2006. — С. 445—451.
26. *Тесленко, Д. А*. Разрешение лексической многозначности при помощи частичного обучения [Текст] / Д. А. Тесленко, Д. А. Усталов // Компьютерная лингвистика и интеллектуальные технологии. Студенческая сессия. — 2018.
27. *Толдова, С. Ю.* Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы [Текст] / С. Ю. Толдова, Г. И. Кустова, О. Н. Ляшевская // Труды конференции «Диалог». — 2008. — С. 522—529.
28. *Турдаков, Д.* Устранение лексической многозначности терминов Википедии на основе скрытой модели Маркова [Текст] / Д. Турдаков // XI Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — 2009. — 2009.
29. *Agirre, E.* Publicly Available Topic Signatures for all WordNet Nominal Senses [Текст] / E. Agirre, O. L. De Lacalle // LREC. — 2004.
30. *Agirre, E.* Unsupervised WSD based on automatically retrieved examples: The importance of bias [Текст] / E. Agirre, D. Martinez // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. — 2004. — С. 25—32.
31. *Agirre, E*. Personalizing PageRank for word sense disambiguation [Текст] / E. Agirre, A. Soroa // Proceedings of EACL. — 2009. — С. 33—41.
32. *Agirre, E.* Random walks for knowledge-based word sense disambiguation [Текст] / E. Agirre, O. Lopez de Lacalle, A. Soroa // Computational Linguistics. — 2014. — Т. 40, № 1. — С. 57—84.

33. Europarl QTLeap WSD/NED Corpus [Текст] / E. Agirre [и др.] // LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. — 2015.

34. FLAIR: An easy-to-use framework for state-of-the-art NLP [Текст] / A. Akbik [и др.] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). — 2019. — С. 54—59.

35. *Alagić, D.* Experiments on active learning for Croatian word sense disambiguation [Текст] / D. Alagić, J. Šnajder // The 5th Workshop on Balto-Slavic Natural Language Processing. — 2015. — С. 49—58.

36. *Alexeyevsky, D.* Word sense disambiguation features for taxonomy extraction [Текст] / D. Alexeyevsky // Computación y Sistemas. —Т. 22, №. 3. — 2018.

37. *Amplayo R. K.* Autosense model for word sense induction [Текст] / R. K. Amplayo, S. Hwang, M. Song // Proceedings of the AAAI Conference on Artificial Intelligence. — 2019. — Т. 33, № 1. — С. 6212—6219.

38. *Amrami, A.* Towards better substitution-based word sense induction [Текст] / A. Amrami, Y. Goldberg // arXiv preprint arXiv:1905.12598. — 2019.

39. *Arefyev, N.* Combining Lexical Substitutes in Neural Word Sense Induction [Текст] / N. Arefyev, B. Sheludko, A. Panchenko // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria: INCOMA Ltd., 09.2019. — С. 62—70. — URL: https://aclanthology.org/R19-1008.

40. *Arefyev, N.* How much does a word weight? Weighting word embeddings for word sense induction [Текст] / N. Arefyev, P. Ermolaev, A. Panchenko // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. — 2018. — С. 68—84.

41. *Artetxe, M.* Margin-based parallel corpus mining with multilingual sentence embeddings [Текст] / M. Artetxe, H. Schwenk // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 3197—3203.

42. *Bar-Hillel, Y.* The present status of automatic translation of languages [Текст] / Y. Bar-Hillel // Advances in computers. — 1960. — Т. 1. — С. 91—163.

43. *Barba, E.* ESC: Redesigning WSD with Extractive Sense Comprehension [Текст] / E. Barba, T. Pasini, R. Navigli // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 06.2021. — С. 4661—4672. — URL: https://aclanthology.org/2021.naacl-main.371.

44. MuLaN: Multilingual Label propagatioN for word sense disambiguation [Текст] / E. Barba [и др.] // Proceedings of IJCAI. — 2020. — С. 3837—3844.

45. Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project [Текст] / F. Barreto [и др.] // Proceedings of the 5th International Conference on Language Resources. — 2006.

46. Breaking sticks and ambiguities with adaptive skip-gram [Текст] / S. Bartunov [и др.] // International Conference on Artificial Intelligence and Statistics (AISTATS). — PMLR. 2016. — С. 130—138.

47. *Basile, P.* An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model [Текст] / P. Basile, A. Caputo, G. Semeraro // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. — Dublin, Ireland: Dublin City University, Association for Computational Linguistics, 08.2014. — 2014. — С. 1591—1600. — URL: https://aclanthology.org/C14-1151.

48. Corpus as language: from scalability to register variation [Текст] / V. Belikov [и др.] // Komp'juternaja lingvistika i intellektual'nye tehnologii. — 2013. — № 12. — С. 84—95.

49. *Benko, V.* Aranea: Yet another family of (comparable) web corpora [Текст] / V. Benko // International Conference on Text, Speech, and Dialogue. — Springer. 2014. — С. 247—256.

50. *Berend, G.* Sparse coding of neural word embeddings for multilingual sequence labeling [Текст] / G. Berend // Transactions of the Association for Computational Linguistics. — 2017. — Т. 5. — С. 247—261.

51. *Berend, G.* Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations [Текст] / G. Berend // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2020. — С. 8498—8508.

52. *Bevilacqua, M.* Quasi Bidirectional Encoder Representations from Transformers for word sense disambiguation [Текст] / M. Bevilacqua, R. Navigli // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — 2019. — С. 122—131.

53. *Bevilacqua, M.* Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information [Текст] / M. Bevilacqua, R. Navigli // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — С. 2854—2864.

54. *Blevins, T.* Moving down the long tail of word sense disambiguation with gloss-informed biencoders [Текст] / T. Blevins, L. Zettlemoyer // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — С. 1006—1017.

55. IR like a SIR: Sense-enhanced Information Retrieval for Multiple Languages [Текст] / R. Blloshmi [и др.] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. —

2021. — С. 1030—1041.

56. Enriching Word Vectors with Subword Information [Текст] / P. Bojanowski [и др.] // Transactions of the Association for Computational Linguistics. — Т. 5. — 2016. — С. 135—146.
57. *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.
58. *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.
59. *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.
60. *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.
61. *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.
62. *Bolshina, A.* 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.
63. *Boser, B. E.* A training algorithm for optimal margin classifiers [Текст] / B.

E. Boser, I. M. Guyon, V. N. Vapnik // Proceedings of the fifth annual workshop on Computational learning theory. — 1992. — С. 144—152.

64. *Boyd-Graber, J.* A topic model for word sense disambiguation [Текст] / J. Boyd-Graber, D. Blei, X. Zhu // Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). — 2007. — С. 1024—1033.
65. *Brin, S.* The anatomy of a large-scale hypertextual web search engine [Текст] / S. Brin, L. Page // Computer networks and ISDN systems. — 1998. — Т. 30, № 1—7. — С. 107—117.
66. *Brody, S.* Bayesian Word Sense Induction [Текст] / S. Brody, M. Lapata // Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). — Athens, Greece: Association for Computational Linguistics, 03.2009. — С. 103—111. — URL: https://aclanthology.org/E09-1013.
67. *Camacho-Collados, J.* NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities [Текст] / J. Camacho-Collados, M. T. Pilehvar, R. Navigli // Artificial Intelligence. — 2016. — Т. 240. — С. 36—64.
68. SenseDefs: a multilingual corpus of semantically annotated textual definitions [Текст] / J. Camacho-Collados [и др.] // Language Resources and Evaluation. — 2019. — Т. 53, № 2. — С. 251—278.
69. *Chan, Y. S.* Scaling up word sense disambiguation via parallel texts [Текст] / Y. S. Chan, H. T. Ng // AAAI. Т. 5. — 2005. — С. 1037—1042.
70. *Chaplot, D. S.* Knowledge-based word sense disambiguation using topic models [Текст] / D. S. Chaplot, R. Salakhutdinov // Proceedings of the AAAI conference on artificial intelligence. Т. 32. — 2018.
71. Bllip 1987-89 wsj corpus release 1 [Текст] / E. Charniak [и др.] // Linguistic Data Consortium, Philadelphia. — 2000. — Т. 36.
72. Applying active learning to supervised word sense disambiguation in

MEDLINE [Текст] / Y. Chen [и др.] // Journal of the American Medical Informatics Association. — 2013. — Т. 20, № 5. — С. 1001—1006.

73. *Chen, H.* Non-Parametric Few-Shot Learning for Word Sense Disambiguation [Текст] / H. Chen, M. Xia, D. Chen // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — С. 1774—1781.
74. *Cuadros, M.* Quality assessment of large-scale knowledge resources [Текст] / M. Cuadros, G. Rigau // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. — 2006. — С. 534—541.
75. *Delli Bovi, C.* Large-scale information extraction from textual definitions through deep syntactic and semantic analysis [Текст] / C. Delli Bovi, L. Telesca, R. Navigli // Transactions of the Association for Computational Linguistics. — 2015. — Т. 3. — С. 529—543.
76. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text [Текст] / C. Delli Bovi [и др.] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — 2017. — С. 594—600.
77. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Текст] / J. Devlin [и др.] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2019.
78. *Di Marco, A.* Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction [Текст] / A. Di Marco, R. Navigli // Computational Linguistics. — Cambridge, MA, 2013. — Сент. — Т. 39, № 3. — С. 709—754. — URL: https://aclanthology.org/J13-3008.
79. Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation

[Текст] / Dušek O. [и др.] // Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015). — 2015. — С. 82–90.

80. *Edmonds, P.* SENSEVAL-2: Overview [Текст] / P. Edmonds, S. Cotton // Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems. — Toulouse, France: Association for Computational Linguistics, 07.2001. — С. 1—5. — URL: https://aclanthology.org/S01-1001.
81. *Eisele, A.* MultiUN: A Multilingual Corpus from United Nation Documents. [Текст] / A. Eisele, Y. Chen // LREC. — 2010.
82. *Fellbaum, C.* A semantic network of English verbs [Текст] / C. Fellbaum // WordNet: An electronic lexical database. — 1998. — Т. 3. — С. 153—178.
83. *Fillmore, C. J.* Frame semantics for text understanding [Текст] / C. J. Fillmore, C. F. Baker // Proceedings of WordNet and Other Lexical Resources Workshop, NAACL. Т. 6. — 2001.
84. Selective sampling for example-based word sense disambiguation [Текст] / Fujii [и др.] // Computational Linguistics. — 1998. — Т. 24, №. 4. — С. 573—597.
85. *Gale, W. A.* One sense per discourse [Текст] / W. A. Gale, K. Church, D. Yarowsky // Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992. — 1992.
86. *Gonzales, A. R.* Improving word sense disambiguation in neural machine translation with sense embeddings [Текст] / A. R. Gonzales, L. Mascarell, R. Sennrich // Proceedings of the Second Conference on Machine Translation. — 2017. — С. 11—19.
87. *Gopal, S.* Malayalam word sense disambiguation using Naïve Bayes classifier [Текст] / S. Gopal, R. P. Haroon // 2016 International Conference on Advances in Human Machine Interaction (HMI). — IEEE. 2016. — С. 1—4.

88. *Gosal, G. P. S.* A Naïve Bayes Approach for Word Sense Disambiguation [Текст] / G. P. S. Gosal // International Journal. — 2015. — Т. 5, № 7.

89. *Hadiwinoto, C.* Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations [Текст] / C. Hadiwinoto, H. T. Ng, W. C. Gan // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China: Association for Computational Linguistics, 11.2019. — С. 5297—5306. — URL: https://aclanthology.org/D19-1533.

90. Prague Čzech-English Dependency Treebank 2.0 [Текст]: тех. отч. / J. Hajič [и др.] // Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC2012T08. — 2012.

91. *Harris, Z. S.* Distributional structure [Текст] / Z. S. Harris // Word. — 1954. — Т. 10, № 2/3. — С. 146—162.

92. Semi-Supervised and Unsupervised Sense Annotation via Translations [Текст] / B. Hauer [и др.] // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). — 2021. — С. 504—513.

93. *Haveliwala, T.* An analytical comparison of approaches to personalizing PageRank [Текст]: тех. отч. / T. Haveliwala, S. Kamvar, G. Jeh; Stanford. — 2003.

94. *Henrich, V.* WebCAGe–A Web-harvested corpus annotated with GermaNet senses [Текст] / V. Henrich, E. Hinrichs, T. Vodolazova // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. — 2012. — С. 387—396.

95. *Hochreiter, S.* Long short-term memory [Текст] / S. Hochreiter, J. Schmidhuber // Neural computation. — 1997. — Т. 9, № 8. — С. 1735—1780.

96. *Hogenboom, A.* The impact of word sense disambiguation on stock price

prediction [Текст] / A. Hogenboom, A. Brojba-Micu, F. Frasincar // Expert Systems with Applications. — 2021. — С. 115568.

97. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation [Текст] / N. Holla [и др.] // Findings of the Association for Computational Linguistics: EMNLP 2020. — 2020. — С. 4517—4533.
98. OntoNotes: the 90% solution [Текст] / E. Hovy [и др.] // Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers. — 2006. — С. 57—60.
99. *Hristea, F.* The long road from performing word sense disambiguation to successfully using it in information retrieval: An overview of the unsupervised approach [Текст] / F. Hristea, M. Colhon // Computational Intelligence. — 2020. — Март. — Т. 36, № 3. — С. 1026—1062.
100. GlossBERT: BERT for word sense disambiguation with gloss knowledge [Текст] / L. Huang [и др.] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — С. 3509—3514.
101. The unified medical language system: an informatics research collaboration [Текст] / B. L. Humphreys [и др.] // Journal of the American Medical Informatics Association. — 1998. — Т. 5, № 1. — С. 1—11.
102. *Iacobacci, I.* Embeddings for word sense disambiguation: An evaluation study [Текст] / I. Iacobacci, M. T. Pilehvar, R. Navigli // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2016. — С. 897—907.
103. *Jimeno-Yepes A.J.* Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation [Текст] / A.J. Jimeno-Yepes, B.T. McInnes, A.R. Aronson // BMC bioinformatics. — 2011. — Т.12, №1. — С. 1-14.
104. *Kågebäck, M.* Word Sense Disambiguation using a Bidirectional LSTM

[Текст] / M. Kågebäck, H. Salomonsson // Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V). — 2016. — С. 51—56.

105. Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual WSD [Текст] / M. M. Khapra [и др.] // Proceedings of the 23rd International Conference on Computational Linguistics. — 2010.
106. Together we can: Bilingual bootstrapping for WSD [Текст] / M. M. Khapra [и др.] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — 2011. — С. 561—569.
107. *Kilgarriff, A.* Framework and Results for English SENSEVAL [Текст] / A. Kilgarriff, J. Rosenzweig // Computers and the Humanities. — 2000. — Т. 34. — С. 15—48.
108. *Koehn, P.* Europarl: A parallel corpus for statistical machine translation [Текст] / P. Koehn // MT summit. Т. 5. — Citeseer. 2005. — С. 79—86.
109. *Kohli, H.* Transfer Learning and Augmentation for Word Sense Disambiguation [Текст] / H. Kohli // European Conference on Information Retrieval. — 2021. — С. 303—311.
110. *Korobov, M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages [Текст] / M. Korobov // Analysis of Images, Social Networks and Texts. Т. 542 / под ред. M. Y. Khachay [и др.]. — Springer International Publishing, 2015. — С. 320—332. — (Communications in Computer and Information Science). — URL: http://dx.doi.org/10.1007/978- 3-319-26123-2_31.
111. Zero-shot word sense disambiguation using sense definition embeddings [Текст] / S. Kumar [и др.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 5670—5681.
112. *Kuratov, Y.* Adaptation of Deep Bidirectional Multilingual Transformers for

Russian Language [Текст] / Y. Kuratov, M. Arkhipov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. — 2019. — С. 333—339.

113. *Kutuzov, A.* WebVectors: a toolkit for building web interfaces for vector semantic models [Текст] / A. Kutuzov, E. Kuzmenko // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2016. — С. 155—161.
114. *Kutuzov, A.* Russian Word Sense Induction by Clustering Averaged Word Embeddings [Текст] / A. Kutuzov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. — 2018. — С. 391—403.
115. *Kutuzov, A.* To lemmatize or not to lemmatize: how word normalisation affects ELMo performance in word sense disambiguation [Текст] / A. Kutuzov, E. Kuzmenko // Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing. — 2019. — С. 22—28.
116. *Large, J.* A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates [Текст] / J. Large, J. Lines, A. Bagnall // Data mining and knowledge discovery. — 2019. — Т. 33, № 6. — С. 1674—1709.
117. *Lashevskaja, O. N.* Disambiguation of taxonomy markers in context: Russian nouns [Текст] / O. N. Lashevskaja, O. Mitrofanova // Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009). Т. 4. — 2009. — С. 111—117.
118. A Deep Dive into Word Sense Disambiguation with LSTM [Текст] / M. Le [и др.] // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 08.2018. — С. 354—365. — URL: https://aclanthology.org/C18- 1030.
119. *Leacock, C.* Using corpus statistics and WordNet relations for sense

identification [Текст] / C. Leacock, M. Chodorow, G. A. Miller // Computational Linguistics. — 1998. — Т. 24, № 1. — С. 147—165.

120. *Le, P.* Boosting Entity Linking Performance by Leveraging Unlabeled Documents [Текст] / P. Le, I. Titov // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (long papers), Vol. 1. — 2019. — С. 1935—1945.
121. *Lee, Y. K.* An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation [Текст] / Y. K. Lee, H. T. Ng // Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). — 2002. — С. 41—48.
122. *Lenat, D. B.* CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks [Текст] / D. B. Lenat, M. Prakash, M. Shepherd // AI magazine. — 1985. — Т. 6, № 4. — С. 65—65.
123. *Lesk, M.* Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone [Текст] / M. Lesk // Proceedings of the 5th annual international conference on Systems documentation. — 1986. — С. 24—26.
124. SenseBERT: Driving some sense into BERT [Текст] / Y. Levine [и др.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 4656—4667.
125. Selective kernel networks for weakly supervised relation extraction [Текст] / Z. Li [и др.] // CAAI Transactions on Intelligence Technology. — 2021. — Т. 6, № 2. — С. 224—234.
126. Neural Relation Extraction with Selective Attention over Instances [Текст] / Y. Lin [и др.] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Berlin, Germany: Association for Computational Linguistics, 08.2016. — С. 2124—2133. — URL: https://aclanthology.org/P16-1200.
127. *Lison, P.* skweak: Weak Supervision Made Easy for NLP [Текст] / P.

Lison, J. Barnes, A. Hubin // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. — Online: Association for Computational Linguistics, 08.2021. — С. 337—346. — URL: https://aclanthology.org/2021.acl-demo.40.

128. *Liu, F.* Handling Homographs in Neural Machine Translation [Текст] / F. Liu, H. Lu, G. Neubig // Proceedings of NAACL-HLT. — 2018. — С. 1336—1345.

129. *Lopukhin, K. A.* Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries [Текст] / K. A. Lopukhin, A. A. Lopukhina // Компьютерная лингвистика и интеллектуальные технологии. — 2016. — С. 393—405.

130. *Lopukhin, K. A.* Word sense induction for Russian: deep study and comparison with dictionaries [Текст] / K. A. Lopukhin, B. L. Iomdin, A. Lopukhina // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog». — 2017. — С. 121—134.

131. Creating Russian WordNet by Conversion [Текст] / N. V. Loukachevitch [и др.] // Proceedings of Conference on Computational linguistics and Intellectual technologies Dialog-2016. — 2016. — С. 405—415.

132. *Loukachevitch, N.* Corpus-based Check-up for Thesaurus [Текст] / N. Loukachevitch // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 5773—5779.

133. *Loukachevitch, N.* Determining the most frequent senses using Russian linguistic ontology RuThes [Текст] / N. Loukachevitch, I. Chetviorkin // Proceedings of the Workshop on Semantic Resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015. — 2015.

134. *Loureiro, D.* Don't Neglect the Obvious: On the Role of Unambiguous

Words in Word Sense Disambiguation [Текст] / D. Loureiro, J. Camacho-Collados // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online: Association for Computational Linguistics, 11.2020. — С. 3514—3520. — URL: https://aclanthology. org / 2020.emnlp-main.283.

135. *Loureiro, D.* Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation [Текст] / D. Loureiro, A. Jorge // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 5682—5691.
136. Improving Word Sense Disambiguation with Translations [Текст] / Y. Luan [и др.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2020. — С. 4055—4065.
137. Incorporating Glosses into Neural Word Sense Disambiguation [Текст] / F. Luo [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2018. — С. 2473—2482.
138. Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation [Текст] / O. Lyashevskaya // Text, Speech and Dialogue. — 2011. — С. 80—90.
139. *Mallery, J. C.* Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers [Текст] / J. C. Mallery // Master's thesis, MIT Political Science Department. — Citeseer. 1988.
140. *Martínez, D.* Syntactic features for high precision word sense disambiguation [Текст] / D. Martínez, E. Agirre, L. Màrquez // COLING 2002: The 19th International Conference on Computational Linguistics. — 2002.
141. *Martínez, D.* Word relatives in context for word sense disambiguation [Текст] / D. Martínez, E. Agirre, X. Wang // Proceedings

of the Australasian Language Technology Workshop 2006. — 2006. — С. 42—50.

142. *Martínez, D.* On the use of automatically acquired examples for all- nouns word sense disambiguation [Текст] / D. Martínez, O. L. de Lacalle, E. Agirre // Journal of Artificial Intelligence Research. — 2008. — Т. 33. — С. 79—107.
143. *Martínez, D.* Word sense disambiguation for event trigger word detection in biomedicine [Текст] / D. Martínez, T. Baldwin // BMC bioinformatics. Т. 12. — Springer. 2011. — С. 1—8.
144. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations [Текст] / M. Maru [и др.] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP). — 2019. — С. 3525—3531.
145. *Melacci, S.* Enhancing modern supervised word sense disambiguation models by semantic lexical resources [Текст] / S. Melacci, A. Globo, L. Rigutini // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.
146. *Melamud, O.* context2vec: Learning generic context embedding with bidirectional LSTM [Текст] / O. Melamud, J. Goldberger, I. Dagan // Proceedings of the 20th SIGNLL conference on computational natural language learning. — 2016. — С. 51—61.
147. *Mihalcea, R.* An automatic method for generating sense tagged corpora [Текст] / R. Mihalcea, D. I. Moldovan // AAAI/IAAI. — 1999. — С. 461—466.
148. *Mihalcea, R.* Bootstrapping Large Sense Tagged Corpora [Текст] / R. Mihalcea // LREC. — 2002a.
149. *Mihalcea R.* Instance based learning with automatic feature selection applied to Word Sense Disambiguation [Текст] / R. Mihalcea //

Proceedings of the 19th International Conference on Computational Linguistics (COLINGACL 2002). — 2002b.

150. *Mihalcea, R.* Open mind word expert: Creating large annotated data collections with web users' help [Текст] / R. Mihalcea, T. Chklovski // Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003. — 2003.
151. *Mihalcea, R.* The role of non-ambiguous words in natural language disambiguation [Текст] / R. Mihalcea // Proceedings of the Conference on Recent Advances in Natural Language Processing, RANLP. — 2003.
152. *Mihalcea, R.* Co-training and self-training for word sense disambiguation [Текст] / R. Mihalcea // Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. — 2004. — С. 33—40.
153. Distributed representations of words and phrases and their compositionality [Текст] / T. Mikolov [и др.] // Advances in neural information processing systems. — 2013. — С. 3111—3119.
154. Using a semantic concordance for sense identification [Текст] / G. A. Miller [и др.] // In Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics. — 1994. — С. 240—243.
155. *Miller, G. A.* WordNet: a lexical database for English [Текст] / G. A. Miller // Communications of the ACM. — 1995. — Т. 38, № 11. — С. 39—41.
156. *Moro, A.* Entity linking meets word sense disambiguation: a unified approach [Текст] / A. Moro, A. Raganato, R. Navigli // Transactions of the Association for Computational Linguistics. — 2014. — Т. 2. — С. 231—244.
157. *Moro, A.* Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking [Текст] / A. Moro, R. Navigli //

Proceedings of SemEval-2015. — 2015.

158. *Navigli, R.* Semeval-2007 task 07: Coarse-grained English all-words task [Текст] / R. Navigli, K. C. Litkowski, O. Hargraves // Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). — 2007. — С.30–35.

159. *Navigli, R.* SemEval-2013 Task 12: Multilingual Word Sense Disambiguation [Текст] / R. Navigli, D. Jurgens, D. Vannella // Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). — Atlanta, Georgia, USA: Association for Computational Linguistics, 06.2013. — С. 222—231. — URL: https://aclanthology.org/S13-2040.

160. *Navigli, R.* Word sense disambiguation: A survey [Текст] / R. Navigli // ACM computing surveys (CSUR). — 2009. — Т. 41, № 2. — С. 1—69.

161. *Navigli, R.* BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network [Текст] / R. Navigli, S. P. Ponzetto // Artificial intelligence. — 2012. — Т. 193. — С. 217—250.

162. *Ng, H. T.* Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach [Текст] / H. T. Ng, H. B. Lee // Proceedings of the 34th annual meeting on Association for Computational Linguistics. — 1996. — С. 40—47.

163. *Ng, H. T.* Exploiting parallel texts for word sense disambiguation: An empirical study [Текст] / H. T. Ng, B. Wang, Y. S. Chan // Proceedings of the 41st annual meeting of the Association for Computational Linguistics. — 2003. — С. 455—462.

164. Qtleap wsd/ned corpora: Semantic annotation of parallel corpora in six languages [Текст] / A. Otegi [и др.] // Proceedings of the Tenth International Conference on Language Resources and Evaluation

(LREC’16). — 2016. — С. 3023—3030.

165. RUSSE’2018: A Shared Task on Word Sense Induction for the Russian Language [Текст] / A. Panchenko [и др.] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. — 2018. — С. 547—564.
166. *Pandit, R.* A memory-based approach to word sense disambiguation in Bengali using k-NN method [Текст] / R. Pandit, S. K. Naskar // 2015 IEEE 2nd international conference on recent trends in information systems (ReTIS). — IEEE. 2015. — С. 383—386.
167. *Pasini, T.* Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data [Текст] / T. Pasini, R. Navigli // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — 2017. — С. 78—88.
168. *Pasini, T.* The knowledge acquisition bottleneck problem in multilingual word sense disambiguation [Текст] / T. Pasini // Proceedings of the Twenty-eighth International Joint Conference on Artificial Intelligence, IJCAI-20, Yokohama, Japan. — 2020.
169. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation [Текст] / T. Pasini, A. Raganato, R. Navigli [и др.] // Proceedings of the AAAI Conference on Artificial Intelligence. — AAAI Press. 2021.
170. Making Sense of Word Embeddings [Текст] / M. Pelevina [и др.] // Proceedings of the 1st Workshop on Representation Learning for NLP. — Berlin, Germany: Association for Computational Linguistics, 08.2016. — С. 174—183. — URL: https://aclanthology.org/W16-1620.
171. *Pennington, J.* GloVe: Global Vectors for Word Representation [Текст] / J. Pennington, R. Socher, C. D. Manning // Empirical Methods in Natural Language Processing (EMNLP). — 2014. — С. 1532—1543. — URL: http://www.aclweb.org/anthology/D14-1162.

172. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data [Текст] / A. Pesaranghader [и др.] // Journal of the American Medical Informatics Association. — 2019. — Т. 26, № 5. — С. 438—446.

173. Deep contextualized word representations [Текст] / M. E. Peters [и др.] // Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2018. — С. 2227—2237.

174. *Pham, T. P.* Word sense disambiguation with semi-supervised learning [Текст] / T. P. Pham, H. T. Ng, W. S. Lee // Proceedings of the National Conference on Artificial Intelligence. Т. 20. — Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2005. — С. 1093.

175. *Pilehvar, M. T.* WiC: the word-in-context dataset for evaluating context-sensitive meaning representations [Текст] / M. T. Pilehvar, J. Camacho-Collados // Proceedings of NAACL-HLT. — 2019. — С.1267—1273.

176. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation [Текст] / A. Piperski [и др.] // Proceedings 8th Web as Corpus Workshop (WAC-8). — 2013. — С. 24—29.

177. *Preiss, J.* DALE: A word sense disambiguation system for biomedical documents trained using automatically labeled examples [Текст] / J. Preiss, M. Stevenson // Proceedings of the 2013 NAACL HLT Demonstration Session. — 2013. — С. 1—4.

178. Narodowy Korpus Języka Polskiego [Текст] / A. Przepiórkowski [и др.] // Wydawnictwo Naukowe PWN, Warszawa. — 2012.

179. *Przybyła, P.* How big is big enough? Unsupervised word sense disambiguation using a very large corpus [Текст] / P. Przybyła // arXiv preprint arXiv:1710.07960. — 2017.

180. SemEval-2007 task-17: English lexical sample, SRL and all words [Текст] / S. Pradhan // Proceedings of SemEval. — 2007. — С. 87–92.

181. Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation [Текст] / X. Pu [и др.] // Transactions of the Association for Computational Linguistics. — 2018. — Дек. — Т. 6. — С. 635—649.

182. *Raganato, A.* Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. [Текст] / A. Raganato, C. D. Bovi, R. Navigli // IJCAI. — 2016. — С. 2894—2900.

183. *Raganato, A.* Neural sequence learning models for word sense disambiguation [Текст] / A. Raganato, C. D. Bovi, R. Navigli // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — 2017a. — С. 1156—1167.

184. *Raganato, A.* Word sense disambiguation: A unified evaluation framework and empirical comparison [Текст] / A. Raganato, J. Camacho-Collados, R. Navigli // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. — 2017b. — С. 99—110.

185. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation [Текст] / A. Raganato, Y. Scherrer, J. Tiedemann [и др.] // Fourth Conference on Machine Translation Proceedings of the Conference (Volume 2: Shared Task Papers, Day 1). — The Association for Computational Linguistics. 2019.

186. Question answering via Bayesian inference on lexical relations [Текст] / G. Ramakrishnan [и др.] // Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering. — 2003. — С. 1—10.

187. *Resnik, P.* A perspective on word sense disambiguation methods and their evaluation [Текст] / P. Resnik // Tagging Text with Lexical Semantics: Why, What, and How? — 1997.

188. *Rezapour, A. R.* Applying weighted KNN to word sense disambiguation

[Текст] / A. R. Rezapour, S. M. Fakhrahmad, M. H. Sadreddini // Proceedings of the world congress on engineering. Т. 3. — 2011. — С. 6—8.

189. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning [Текст] / M. N. Rizve [и др.] // International Conference on Learning Representations. — 2021.
190. *Rothe, S.* Autoextend: Extending word embeddings to embeddings for synsets and lexemes [Текст] / S. Rothe, H. Schütze // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — 2015. — С. 1793—1803.
191. *Sabbir, A. K. M.* Knowledge-based biomedical word sense disambiguation with neural concept embeddings [Текст] / A. K. M. Sabbir, A. Jimeno-Yepes, R. Kavuluru // 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). — IEEE. 2017. — С. 163—170.
192. Building Sense Tagged Corpus Using Wikipedia for Supervised Word Sense Disambiguation [Текст] / A. Saif [и др.] // Procedia Computer Science. — 2018. — Т. 123. — С. 403—412.
193. *Scarlini, B.* Just “OneSeC” for producing multilingual sense-annotated data [Текст] / B. Scarlini, T. Pasini, R. Navigli // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 699—709.
194. *Scarlini, B.* With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation [Текст] / B. Scarlini, T. Pasini, R. Navigli // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2020a. — С. 3528—3539.
195. *Scarlini, B.* SensEmBERT: Context-enhanced sense embeddings for

multilingual word sense disambiguation [Текст] / B. Scarlini, T. Pasini, R. Navigli // Proceedings of the AAAI Conference on Artificial Intelligence. T. 34. — 2020b. — C. 8758—8765.

196. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation [Текст] / F. Scozzafava [и др.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. — 2020. — C. 37—46.
197. *Searle, J. R.* Minds, brains, and programs [Текст] / J. R. Searle // Behavioral and brain sciences. — 1980. — Т. 3, № 3. — С. 417—424.
198. Unsupervised word sense disambiguation using WordNet relatives [Текст] / H.-C. Seo [и др.] // Computer Speech & Language. — 2004. — Июль. — Т. 18, № 3. — С. 253—273.
199. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark [Текст] / T. Shavrina [и др.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online: Association for Computational Linguistics, 11.2020. — 2020. — С. 4717—4726. — URL: https://www.aclweb.org/anthology/2020.emnlp- main.381.
200. *Shavrina, T.* To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser [Текст] / T. Shavrina, O. Shapovalova // Proceedings of the “Corpora-2017”. — 2017. — С. 78—84.
201. *Shimura, K.* Text categorization by learning predominant sense of words as auxiliary task [Текст] / K. Shimura, J. Li, F. Fukumoto // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — С. 1109—1119.
202. *Singh, S.* Naïve Bayes Classifier for Hindi Word Sense Disambiguation [Текст] / S. Singh, T. J. Siddiqui, S. K. Sharma // Proceedings of the 7th ACM India Computing Conference. — 2014.
203. *Snyder B.* The English all-words task [Текст] / B. Snyder, M. Palmer //

Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), Barcelona, Spain. — 2004. — C.41–43.

204. Spine: Sparse interpretable neural embeddings [Текст] / A. Subramanian [и др.] // Proceedings of the AAAI Conference on Artificial Intelligence. Т. 32. — 2018.
205. *Sutskever, I.* Sequence to sequence learning with neural networks [Текст] / I. Sutskever, O. Vinyals, Q. V. Le // Advances in neural information processing systems. — 2014. — С. 3104—3112.
206. *Taghipour, K.* One million sense-tagged instances for word sense disambiguation and induction [Текст] / K. Taghipour, H. T. Ng // Proceedings of the nineteenth conference on computational natural language learning. — 2015a. — С. 338—344.
207. *Taghipour, K.* Semi-supervised word sense disambiguation using word embeddings in general and specific domains [Текст] / K. Taghipour, H. T. Ng // Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies. — 2015b. — С. 314—323.
208. *Tripodi, R.* A game-theoretic approach to word sense disambiguation [Текст] / R. Tripodi, M. Pelillo // Computational Linguistics. — 2017. — Т. 43, № 1. — С. 31—70.
209. *Tripodi, R.* Game theory meets embeddings: a unified framework for word sense disambiguation [Текст] / R. Tripodi, R. Navigli // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — С. 88—99.
210. fastsense: An efficient word sense disambiguation classifier [Текст] / T. Uslu [и др.] // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

211. An unsupervised word sense disambiguation system for under-resourced languages [Текст] / D. Ustalov [и др.] // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018. — С. 1018—1022.

212. *Van der Maaten, L.* Visualizing data using t-SNE [Текст] / L. Van der Maaten, G. Hinton // Journal of machine learning research. — 2008. — Т. 9, № 11.

213. *Vasilescu, F.* Evaluating Variants of the Lesk Approach for Disambiguating Words [Текст] / F. Vasilescu, P. Langlais, G. Lapalme // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). — Lisbon, Portugal: European Language Resources Association (ELRA), 05.2004. — URL: http://www.lrec-conf.org/proceedings/lrec2004/ pdf/219.pdf.

214. *Véronis, J.* Hyperlex: lexical cartography for information retrieval [Текст] / J. Véronis // Computer Speech & Language. — 2004. — Т. 18, № 3. — С. 223—252.

215. *Vial, L.* Improving the coverage and the generalization ability of neural word sense disambiguation through hypernymy and hyponymy relationships [Текст] / L. Vial, B. Lecouteux, D. Schwab // arXiv preprint arXiv:1811.00960. — 2018.

216. *Vial, L.* Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation [Текст] / L. Vial, B. Lecouteux, D. Schwab // Proceedings of the 10th Global Wordnet Conference. — 2019. — С. 108—117.

217. *Vrandečić, D.* Wikidata: A new platform for collaborative data collection [Текст] / D. Vrandečić // Proceedings of the 21st international conference on world wide web. — 2012. — С. 1063—1064.

218. *Wang, X.* Word sense disambiguation using sense examples automatically acquired from a second language [Текст] / X. Wang, J. A. Carroll // Proceedings of Human Language Technology Conference and Conference

on Empirical Methods in Natural Language Processing. — 2005. — С. 547—554.

219. *Wang, X.* Word sense disambiguation using automatically translated sense examples [Текст] / X. Wang, D. Martínez // Proceedings of the Cross-Language Knowledge Induction Workshop. — 2006.
220. A clinical text classification paradigm using weak supervision and deep representation [Текст] / Y. Wang [и др.] // BMC medical informatics and decision making. — 2019. — Т. 19, № 1. — С. 1—13.
221. Translation [Текст] / W. Weaver [и др.] // Machine translation of languages. — 1955. — Т. 14, № 15—23. — С. 10.
222. *Wei, J.* EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [Текст] / J. Wei, K. Zou // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China: Association for Computational Linguistics, 11.2019. — С. 6382—6388. — URL: https://aclanthology.org/D19- 1670.
223. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings [Текст] / G. Wiedemann [и др.] // Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers. — Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019. — С. 161—170.
224. *Yarowsky, D.* One sense per collocation [Текст] / D. Yarowsky // Proceedings of the workshop on Human Language Technology. — 1993. — С. 266—271.
225. Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French [Текст] / D. Yarowsky // Proceedings of the 32nd Annual Meeting of the Association for

Computational Linguistics. — 1994.

226. *Yarowsky, D.* Unsupervised word sense disambiguation rivaling supervised methods [Текст] / D. Yarowsky // 33rd annual meeting of the association for computational linguistics. — 1995. — С. 189—196.
227. *Yarowsky, D.* Evaluating sense disambiguation across diverse parameter spaces [Текст] / D. Yarowsky, R. Florian // Natural Language Engineering. — 2002. — Т. 8, № 4. — С. 293—310.
228. Semi-supervised word sense disambiguation with neural models [Текст] / D. Yuan [и др.] // Proceedings of COLING. — 2016.
229. *Zhong, Z.* It makes sense: A wide-coverage word sense disambiguation system for free text [Текст] / Z. Zhong, H. T. Ng // Proceedings of the ACL 2010 system demonstrations. — 2010. — С. 78—83.
230. *Zhong, Z.* Word Sense Disambiguation Improves Information Retrieval [Текст] / Z. Zhong, H. T. Ng // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2012. — С. 273—282.
231. *Ziemski, M.* The united nations parallel corpus v1. 0 [Текст] / M. Ziemski, M. Junczys-Dowmunt, B. Pouliquen // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — С. 3530—3534.
232. https://www.wikipedia.org // Домашняя страница Википедии (дата обращения: 22.04.2022)
233. https://www.wiktionary.org // Домашняя страница Викисловарей (дата обращения: 22.04.2022)
234. https://ru.wikinews.org/wiki/Заглавная_страница // Домашняя страница портала Wikinews (дата обращения: 22.04.2022)
235. https://ushakovdictionary.ru/ // Онлайн-словарь Ушакова (дата обращения: 22.04.2022)
236. https://slovarozhegova.ru/ // Онлайн-словарь Ожегова (дата обращения:

22.04.2022)

237. https://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/acref-9780199571123 // Oxford Dictionary of English (3 ed.) (дата обращения: 22.04.2022)
238. https://www.ldoceonline.com/ // Longman Dictionary of Contemporary English Online (дата обращения: 22.04.2022)
239. https://eur-lex.europa.eu/browse/eurovoc.html // Eurovoc thesaurus homepage (дата обращения: 22.04.2022)
240. https://www.congress.gov/browse/legislative-indexing-vocabulary/106th-congress // Legislative Indexing Vocabulary (LIV) Terms (дата обращения: 22.04.2022)
241. https://uni-tuebingen.de/en/142806 // GermaNet homepage (дата обращения: 22.04.2022)
242. https://cst.ku.dk/english/projects/dannet/ // DanNet homepage (дата обращения: 22.04.2022)
243. https://framenet.icsi.berkeley.edu/fndrupal/ // Official website for the FrameNet Project (дата обращения: 22.04.2022)
244. https://babelnet.org/about // Official website of BabelNet (дата обращения: 22.04.2022)
245. http://compling.hss.ntu.edu.sg/omw/ // Open Multilingual Wordnet (дата обращения: 22.04.2022)
246. http://www.omegawiki.org/ // OmegaWiki Project (дата обращения: 22.04.2022)
247. https://www.wikidata.org/wiki/Wikidata:Main_Page // Wikidata homepage (дата обращения: 22.04.2022)
248. http://www.natcorp.ox.ac.uk/ // British National Corpus online (дата обращения: 22.04.2022)
249. https://www.anc.org/ // The Open American National Corpus (дата обращения: 22.04.2022)

250. https://ruscorpora.ru/new/ // Национальный корпус русского языка (дата обращения: 22.04.2022)
251. http://korpus.uib.no/icame/brown/bcm.html // BROWN CORPUS MANUAL (дата обращения: 22.04.2022)
252. https://wordnetcode.princeton.edu/glosstag.shtml // Princeton WordNet Gloss Corpus (дата обращения: 22.04.2022)
253. https://nlpub.github.io/russe-wsi-kit/ // A Participant's Kit for RUSSE 2018 Word Sense Induction and Disambiguation Shared Task (дата обращения: 22.04.2022)
254. https://catalog.ldc.upenn.edu/LDC99T42 // Treebank-3 (дата обращения: 22.04.2022)
255. https://ruwordnet.ru/ru // Тезаурус русского языка RuWordNet (дата обращения: 22.04.2022)
256. https://russiansuperglue.com/ // RUSSIAN SUPERGLUE BENCHMARK (дата обращения: 22.04.2022)
257. http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz // ссылка для скачивания набора данных EuSemcor (дата обращения: 22.04.2022)

ПЕРЕЧЕНЬ ТАБЛИЦ

СПИСОК РИСУНКОВ

ПРИЛОЖЕНИЕ 1. Результаты оценки моделей, обученных на автоматически сгенерированных коллекциях.

Список сокращений:

- Prec. – Precision;
- Rec. – Recall;
- Acc. – Accuracy.

Источник	**Классификатор**	**Тестовый корпус**	**Оценка**	**Примечание**
[Mihalcea, Chklovski, 2003]	Semantic Tagger with Active Feature Selection (STAFS) [Mihalcea, 2002b]	Senseval-2, fine-grained, английский язык, существительные	0.645 Prec.	В обучении модели также использовалась обучающая выборка из Senseval-2.
[Ng et al., 2003]	Наивный байесовский классификатор	Senseval-2 lexical-sample, английский язык, существительные	0.72 Acc.	Тестировалось на 22 словах из набора данных. Всего в наборе данных 29 существительных.
[Agirre, Martinez, 2004]	Decision Lists [Yarowsky, 1994]	Senseval-2 lexical-sample, английский язык, существительные	0.5 Rec.	
[Mihalcea, 2004]	Наивный байесовский классификатор + smoothed co-training	Senseval-2 lexical-sample, английский язык, существительные	0.58 Prec.	В качестве начального набора обучающих данных использовалась обучающая выборка из набора данных Senseval-2 lexical-sample. Из тестового набора данных удалялись коллокации, содержащие ключевые

				многозначные слова.
[Chan, Ng, 2005]	Наивный байесовский классификатор	Senseval-2 all-words, английский язык, существительные	0.77 Acc.	Тестировалось на 437 словах из данного набора данных.
[Wang, Carroll, 2005]	Наивный байесовский классификатор	Senseval-2 lexical-sample, английский язык, существительные	0.52 Acc.	
[Pham et al., 2005]	Наивный байесовский классификатор + SGT-Cotraining	Senseval-2 lexical-sample, английский язык, существительные	0.65 Acc.	В качестве начального набора обучающих данных использовалась обучающая выборка из набора данных Senseval-2 lexical-sample.
	Наивный байесовский классификатор + SGT-Cotraining	Senseval-2 all-words, английский язык; существительные, глаголы и прилагательные	0.56 Acc.	В данной работе корпус SemCor использовался в качестве начального набора обучающих данных.
[Wang, Martinez, 2006]	Vector Space Model	Набор данных TWA [Mihalcea, 2003], английский язык, существительные	0.82 Rec.	Набор данных составлен для 6 существительных.
	Vector Space Model	Senseval-3 lexical sample [Snyder, Palmer, 2004], английский язык, существительные и прилагательные	0.39 Rec.	
[Martinez et al., 2008]	Decision Lists [Yarowsky, 1994]	Senseval-2 lexical-sample, английский язык, существительные	0.57 Rec.	

		Senseval-3 all-words, английский язык, существительные	0.65 Rec.	
[Khapra et al., 2011]	Bilingual Bootstrapping	Hindi-Health, язык хинди	0.58 F1	Все наборы данных описаны в работе [Khapra et al., 2010].
	Bilingual Bootstrapping	Marathi-Health, язык маратхи	0.65 F1	
	Bilingual Bootstrapping	Hindi-Tourism, язык хинди	0.61 F1	
	Bilingual Bootstrapping	Marathi-Tourism, язык маратхи	0.62 F1	
[Chen et al., 2013]	The pool-based active learning approach+Least confidence active learning algorithm+SVM WSD classification model	Набор данных MSH WSD [Jimeno-Yepes et al., 2011]	0.94 Acc.	Тестирование проводилось на 197 словах из этого набора данных.
[Taghipour, Ng, 2015]	IMS	Senseval-2 all-words, английский язык	0.64 Acc.	
	IMS	Senseval-3 all-words, английский язык	0.61 Acc.	
	IMS	SemEval-2007 fine-grained task 17 [Pradhan et al., 2007], английский язык	0.53 Acc.	
	IMS	SemEval-2007 coarse-grained task 7 [Navigli et al., 2007], английский язык	0.79 Acc.	
[Alagić, Šnajder, 2015]	The pool-based active learning strategy using	Вручную составленный набор данных с	0.89 Acc.	

	uncertainty sampling + SVM as the core classifier	аннотацией значений, состоящий из шести многозначных слов для хорватского языка		
[Yuan et al., 2016]	LSTMLP[33] (T:OMSTI, U:1K)	Senseval-2 all-words, английский язык	0.74 F1	Метка «T» обозначает корпус, который использовался для обучения модели. Метка «U» обозначает корпус, который использовался в качестве неразмеченных данных при обучении с частичным привлечением учителя. Корпус «1K» в исследовании состоит из 1000 предложений для каждой леммы, которые были случайным образом извлечены из сети Интернет.
	LSTMLP (T:SemCor, U:1K)	Senseval-3 all-words, английский язык	0.72 F1	
	LSTMLP (T:SemCor, U:OMSTI)	SemEval-2007 task 17, английский язык	0.64 F1	
	LSTMLP (T:SemCor, U:OMSTI)	SemEval-2007 task 07 coarse-grained English all-words task	0.84 F1	
	LSTMLP (T:SemCor, U:1K)	SemEval-2013, task 12, английский язык	0.69 F1	
	LSTMLP (T:SemCor, U:1K)	SemEval-2015, task 13 [Moro, Navigli, 2015], английский язык	0.73 F1	
[Raganato et al., 2016]	IMS	SemEval-2013, task 12, английский язык	0.81 F1	
	IMS	SemEval-2015, task 13, английский язык	0.88 F1	
[Otegi et al., 2016]	UKB [Agirre, Soroa, 2009]	Набор данных EPEC-EuSemcor[34],	0.56 F1	

[33] “LSTM language model with label propagation” [Yuan et al., 2016: 1]

		баскский язык		
	С помощью статистики встречаемости биграмм и гипотезы “One sense per discourse” [Gale et al., 1992]	The BulTreeBank-DB WSD gold standard, болгарский язык	0.66 F1	
	The verbal WSD approach [Dušek et al., 2015]	Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2012], чешский язык	0.80 F1	
	UKB	SemEval-2007 coarse-grained, английский язык	0.80 F1	
	Инструмент для разрешения лексической неоднозначности на португальском языке LX-WSD, который основан на UKB	Международный корпус португальского языка CINTIL [Barreto et al., 2006]	0.65 F1	
	UKB	SemEval-2007 task 09, испанский язык	0.79 F1	
[Pasini, Navigli, 2017]	IMS	Senseval-2 all-words, английский язык, существительные	0.70 F1	
	IMS	Senseval-3 all-words, английский язык, существительные	0.67 F1	
	IMS	SemEval-2007 task 17, английский язык, существительные	0.6 F1	
	IMS	SemEval-2013, task 12,	0.65 F1	

[34] http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz

		английский язык		
	IMS	SemEval-2015, task 13, английский язык	0.69 F1	
[Delli Bovi et al., 2017]	IMS	SemEval-2013 all-words, task 12; английский, французский, немецкий, испанский, итальянский языки	0.66 F1	В данной работе корпус SemCor использовался в качестве начального набора обучающих данных. Затем из корпуса EUROSENSE, созданного в рамках этого исследования, выбиралось около 500 примеров для каждого значения слова. Наконец, модель IMS обучалась на этой дополненной обучающей выборке.
	IMS	SemEval-2015 all-words, task 13; английский, испанский и итальянский языки	0.69 F1	
[Przybyła, 2017]	Наивный байесовский классификатор	Национальный корпус польского языка [Przepiórkowski et al., 2012]	0.78 Acc.	
[Scarlini et al., 2019]	IMS	Senseval-2 all-words, английский язык, существительные	0.71 F1	
	IMS	Senseval-3 all-words, английский язык, существительные	0.64 F1	
	IMS	SemEval-2007 task 17, английский язык, существительные	0.63 F1	
	IMS	SemEval-2013,	0.61 F1	

		task 12, английский язык		
	IMS	SemEval-2015, task 13, английский язык	0.67 F1	
	Bi-LSTM	SemEval-2013, task 12, итальянский язык	0.68 F1	
	Bi-LSTM	SemEval-2013, task 12, испанский язык	0.72 F1	
	Bi-LSTM	SemEval-2013, task 12, французский язык	0.75 F1	
	Bi-LSTM	SemEval-2013, task 12, немецкий язык	0.75 F1	
	Bi-LSTM	SemEval-2015, task 13, итальянский язык	0.62 F1	
	Bi-LSTM	SemEval-2015, task 13, испанский язык	0.63 F1	
[Barba et al., 2020]	mBERT	SemEval-2013, task 12, итальянский язык	0.78 F1	
	mBERT	SemEval-2013, task 12, испанский язык	0.81 F1	
	mBERT	SemEval-2013, task 12, французский язык	0.82 F1	
	mBERT	SemEval-2013, task 12, немецкий язык	0.82 F1	
	mBERT	SemEval-2015, task 13, итальянский язык	0.72 F1	
	mBERT	SemEval-2015, task 13, испанский язык	0.69 F1	

[Hauer et al., 2021]	mBERT+ LABELGEN	SemEval-2013, task 12, итальянский язык	0.78 F1	
	mBERT+ LABELGEN	SemEval-2013, task 12, испанский язык	0.80 F1	
	mBERT+ LABELGEN	SemEval-2013, task 12, французский язык	0.81 F1	
	mBERT+ LABELGEN	SemEval-2013, task 12, немецкий язык	0.75 F1	
	mBERT+ LABELSYNC	SemEval-2015, task 13, итальянский язык	0.71 F1	
	mBERT+ LABELSYNC	SemEval-2015, task 13, испанский язык	0.66 F1	
	mBERT+ LABELSYNC	Senseval-2 all-words, английский язык	0.7 F1	
	mBERT+ LABELSYNC	Senseval-3 all-words, английский язык	0.66 F1	
	mBERT+ LABELSYNC	SemEval-2007 task 17, английский язык	0.55 F1	
	mBERT+ LABELSYNC	SemEval-2013, task 12, английский язык	0.71 F1	
	mBERT+ LABELSYNC	SemEval-2015, task 13, английский язык	0.75 F1	

ПРИЛОЖЕНИЕ 2. Количество примеров для слов из набора данных RUSSE-RuWordNet в сбалансированной обучающей коллекции и Корпус-1000.

Многозначное слово	**Значение**	**Корпус-1000**	**Сбалансированная коллекция**
акция$_1$	Ценная бумага	1000	1239
акция$_2$	Действие	1000	1314
байка$_1$	Выдумка, вымысел	1000	1227
байка$_2$	Байковая ткань	245	245
гвоздика$_1$	Цветок	1000	1314
гвоздика$_2$	Пряность	1000	1154
гусеница$_1$	Гусеница бабочки	1000	1295
гусеница$_2$	Гусеничная лента	1000	1153
капот$_1$	Капот машины	1000	918
капот$_2$	Пеньюар	1000	1084
крона$_1$	Крона дерева	1000	1131
крона$_2$	Валюта	1000	1314
рок$_1$	Рок-музыка	1000	1016
рок$_2$	Несчастливая судьба	1000	938
слог$_1$	Слог слова	1000	1047
слог$_2$	Литературный стиль	1000	1137
стопка$_1$	Стопка предметов	1000	1258
стопка$_2$	Стакан (сосуд)	1000	1005
таз$_1$	Часть скелета	1000	1124
таз$_2$	Сосуд	1000	1314
такса$_1$	Расценки на услуги	1000	1300
такса$_2$	Собака	1000	1069

замок$_1$	Средневековый замок	1000	1078
замок$_2$	Замок для запирания	1000	947
лук$_1$	Оружие	1000	1286
лук$_2$	Растение	1000	1267
бор$_1$	Химический элемент	1000	1292
бор$_2$	Хвойный лес	1000	675
дар$_1$	Дарование, талант	1000	1117
дар$_2$	Подарок	1000	1169
двигатель$_1$	Мотор	1000	1310
двигатель$_2$	Катализатор, двигатель изменений	1000	1305
дедушка$_1$	Старый мужчина	1000	1299
дедушка$_2$	Дед (родственник)	1000	1231
декрет$_1$	Отпуск по беременности и родам	128	128
декрет$_2$	Правовой документ	1000	1300
дерево$_1$	Древесное растение	1000	1309
дерево$_2$	Древесина (материал)	1000	966
диалог$_1$	Разговор	1000	1278
диалог$_2$	Обмен мнениями	1000	1300
диплом$_1$	Награда	1000	1253
диплом$_2$	Дипломная работа	1000	1246
доктор$_1$	Врач	1000	1310
доктор$_2$	Доктор наук	1000	1300
доля$_1$	Часть	1000	1300
доля$_2$	Участь, судьба	1000	1300
достижение$_1$	Достичь, добиться	1000	1300
достижение$_2$	Достичь степени,	1000	1309

	уровня		
жестокость$_1$	Беспощадность	1000	801
жестокость$_2$	Жестокое обращение	1000	1313
жребий$_1$	Жеребьевка, решение по жребию	1000	1280
жребий$_2$	Участь, судьба	1000	1300
затея$_1$	Развлечение (занятие)	1000	1308
затея$_2$	Предприятие (вид деятельности)	1000	1309
застой$_1$	Застояться (испортиться)	1000	758
застой$_2$	Застой в развитии	1000	1235
затишье$_1$	Период снижения активности	1000	1206
затишье$_2$	Безветрие	1000	1300
затмение$_1$	Одурение, одурь	1000	1300
затмение$_2$	Затмение небесного тела	1000	1002