

АВТОМАТИЗИРОВАННОЕ СОЗДАНИЕ СЕМАНТИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА СЛОВСОЧЕТАНИЙ

Д.А. Зарипова

Московский государственный университет имени М.В. Ломоносова, Москва, Россия;
diana.ser.sar96@gmail.com

Н.В. Лукашевич

Московский государственный университет имени М.В. Ломоносова, Москва, Россия;
louk_nat@mail.ru

Аннотация: Задача автоматического разрешения многозначности (Word Sense Disambiguation, WSD) является первым и ключевым этапом семантического анализа текста. Она заключается в выборе одного из значений многозначного слова в контексте и вызывает затруднения даже у людей-аннотаторов. Кроме того, качество моделей WSD зависит от предметной области текста, от выбранных словарей, инвентаря значений. Для обучения и тестирования моделей на основе машинного обучения, в том числе нейронных сетей, необходимы большие объёмы данных с семантической разметкой.

Ручная разметка по значениям оказывается трудоёмкой, дорогой и занимает много времени. Поэтому важно разрабатывать и тестировать подходы к автоматической и полуавтоматической семантической разметке. Среди возможных источников информации для такой разметки родственные слова (синонимы, гиперонимы, гипонимы), а также коллокации, в которые входит слово.

В статье описывается процесс автоматического создания корпуса коллокаций с семантической разметкой на материале русского языка, принципы отбора словосочетаний в корпус. Для разрешения многозначности слов в пределах коллокаций используются родственные слова с опорой на тезаурус RuWordNet. Этот же тезаурус выступает источником инвентарей значений. Описанные методы позволяют достичь F1-меры 80% и добавить порядка 23% коллокаций с неснятой многозначностью в корпус.

Семантически размеченные корпуса коллокаций, созданные в автоматическом режиме, позволяют упростить подготовку размеченных данных для обучения и оценки моделей WSD, а также могут использоваться как источник знаний в модели автоматического разрешения многозначности на основе знаний.

Ключевые слова: автоматическое разрешение многозначности; «одно значение на коллокацию»; коллокация; родственные слова.

Для цитирования:

AUTOMATIC GENERATION OF SEMANTICALLY ANNOTATED COLLOCATION CORPUS

D.A. Zaripova

Lomonosov Moscow State University, Moscow, Russia; diana.ser.sar96@gmail.com

N.V. Loukachevitch

Lomonosov Moscow State University, Moscow, Russia; louk_nat@mail.ru

Abstract: Word Sense Disambiguation (WSD) is a crucial initial step in automatic semantic analysis. It involves selecting the correct sense of an ambiguous word in a given context, which can be challenging even for human annotators. Furthermore, the accuracy of WSD models depends on various factors, such as the specific task and subject area, the chosen dictionary, and the sense inventory used.

Supervised machine learning models, such as Artificial Neural Networks, require large datasets with semantic annotation to be effective. However, manual sense labeling can be a

costly, labor-intensive, and time-consuming task. Therefore, it is crucial to develop and test automatic and semi-automatic methods of semantic annotation. Information about semantically related words, such as synonyms, hypernyms, hyponyms, and collocations in which the word appears, can be used for these purposes.

In this article, we describe our approach to generating a semantically annotated collocation corpus for the Russian language. Our goal was to create a resource that could be used to improve the accuracy of WSD models for Russian. This article outlines the process of generating a semantically annotated collocation corpus for Russian and the principles used to select collocations for the corpus. To disambiguate words within collocations, semantically related words defined based on RuWordNet are utilized. The same thesaurus is also used as the source of sense inventories. The methods described in the paper yield an F1-score of 80% and help to add approximately 23% of collocations with at least one ambiguous word to the corpus.

Automatically generated collocation corpora with semantic annotation can simplify the preparation of datasets for developing and testing WSD models. These corpora can also serve as a valuable source of information for knowledge-based WSD models.

Key words: word sense disambiguation; one sense per collocation; collocation; related words.

For citation:

Введение

Задача *автоматического разрешения лексической неоднозначности* (Word Sense Disambiguation, WSD) играет важную роль в *автоматической обработке естественного языка* (Natural Language Processing, NLP). Результаты WSD влияют на качество решения таких более высокоуровневых задач, как *машинный перевод* [Pu et al., 2018], *информационный поиск* [Blloshmi et al., 2021], анализ тональности [Seifollahi, Shajari, 2019]. Однако для обучения и тестирования моделей WSD необходимы объёмные корпуса с семантической разметкой, создание которых является трудоёмким и длительным процессом. Корпуса коллокаций, размеченных по значениям, могут значительно упростить разметку. Теоретическим основанием такого подхода выступает гипотеза «одного значения на коллокацию». В таком случае нет необходимости принимать решение о разметке каждого слова по отдельности, можно сразу проставлять теги для лексем коллокации, найденной в корпусе. Самым известным примером размеченного корпуса коллокаций является SyntagNet ([Maru et al., 2019]), состоящий из более 80 тысяч словосочетаний и доступный на пяти языках.

В статье описывается процесс создания такого корпуса, теоретические основания, а также разные подходы к семантической разметке коллокаций. Цель работы – исследовать возможности автоматической разметки словосочетания с целью создания размеченного корпуса для русского языка. Раздел 1 посвящён гипотезе «одно значение на коллокацию» - теоретическому основанию создания корпуса размеченных

словосочетаний, в разделе 2 приводятся примеры существующих корпусов словосочетаний, в том числе SyntagNet. В разделе 3 разбираются эксперименты по разметке корпуса коллокаций на материале русского языка, раздел 5 содержит анализ ошибок. Статья завершается краткими выводами и направлениями для будущих исследований.

Корпуса, размеченные с помощью описанных в статье методов, а также файлы с ошибками моделей и «золотой стандарт» размещены в открытом доступе на GitHub по адресу: <https://github.com/Diana-Zaripova/SemanticallyAnnotatedCollocationCorpus/>.

1. Гипотеза «одно значение на коллокацию»

В работе [Yarowsky, 1993] была сформулирована гипотеза «одного значения на коллокацию»: в составе коллокации многозначные слова встречаются в одном конкретном значении. Авторы исследовали распределение значений многозначных слов в рамках коллокаций на материале нескольких типов словосочетаний. Средний процент подтверждения гипотезы «одного значения на коллокацию» составил 95%. Важно отметить, что рассматривались только многозначные слова с бинарной моделью многозначности.

В другой статье [Martinez, Agirre, 2000] проверяют гипотезу на многозначных словах с более сложными, небинарными, моделями многозначности на материале двух корпусов. При таких вводных данных гипотеза подтвердилась только в 70% случаев. Отдельно авторы отмечают, что необходимо принимать во внимание жанры и типы текстов при проведении исследований многозначности на материале более чем одного корпуса.

2. Корпуса с семантической разметкой

2.1. SyntagNet

SyntagNet¹ – это большой по объёму (более 80 тысяч словосочетаний) ресурс, содержащий размеченные по лексическим значениям вручную коллокации. Процесс создания корпуса, а также стоящие за этим идеи подробно описаны в статье [Maru et al., 2019]. Название лексико-семантического ресурса неслучайно: это своего рода аналог известного тезауруса WordNet² (отсюда -Net в названии корпуса), с одной стороны, но элементами являются не отдельные лексемы, а словосочетания, коллокации, или определённого типа синтагмы (отсюда компонент Syntag- названия). Основная цель создания подобных лексико-семантических ресурсов – упрощение и ускорение разметки тренировочных и тестовых коллекций данных для обучения, тестирования и оценки качества работы моделей на основе *машинного обучения* (*Machine Learning, ML*), часто применяемые в задаче автоматического разрешения многозначности. Кроме того знание о значении многозначного слова в пределах известной коллокации может использоваться при решении задачи с помощью *методов на основе знаний* (*Knowledge-based methods*). На текущий момент ресурс размеченных по значениям

¹ <http://syntagnet.org/>

² <https://wordnet.princeton.edu/>

коллокаций доступен для пяти языков: английского, французского, итальянского, немецкого и испанского.

Коллокации для корпуса SyntagNet (авторы называют их *лексическими комбинациями*, *lexical combinations*) были извлечены из английской Wikipedia³ и Британского Национального Корпуса (*British National Corpus*, *BNC*; [Leech, 1992]), а затем размечены вручную с помощью инвентаря значений WordNet версии 3.0. Степень согласованности между аннотаторами, измеренная на выборке из 500 коллокаций, составила 0.71, большинство расхождений в разметке при этом было связано со сложными случаями, где имеет место вариативность тегов, следствие высокой степени детализации значений в WordNet.

Для извлечения деревьев зависимостей из всех предложений Wikipedia и BNC авторами использовался пайплайн Stanford CoreNLP [Manning et al., 2014]. Далее для определения релевантных комбинаций авторами определялась сила корреляции между лемматизированными содержательными словами с частеречными метками w_1 , w_2 , встретившимися в одном контексте в пределах скользящего окна в три слова. Каждая пара-кандидат взвешивалась с помощью *коэффициента Дайса* (*Dice's coefficient*), умноженного на логарифм частоты совместной встречаемости слов:

$$\text{score}(w_1, w_2) = \log_2(1 + n_{w_1w_2}) \frac{2n_{w_1w_2}}{n_{w_1} + n_{w_2}},$$

где n_{w_i} ($i \in \{1, 2\}$) – это частотность слова w_i , а $n_{w_1w_2}$ это частота совместной встречаемости двух слов w_1 и w_2 в пределах окна. На следующем этапе к полученному списку пар были применены три фильтра: 1) были убраны пары со стоп-словами английского языка в соответствии с библиотекой NLTK; 2) были отсеяны комбинации двух глаголов; 3) были убраны те пары слов, которые не были связаны одним из пяти наиболее частотных типов зависимостей (сложное слово (*compound*), прямой объект, не прямой объект, субъект-существительное и модификатор-существительное). На заключительном этапе подготовки данных список коллокаций был проранжирован в соответствии со средним геометрическим логарифмических коэффициентов Дайса и счётчика частотности пары в данной комбинации частеречных тегов и с тем же типом отношений. Также авторами были проведены эксперименты по извлечению пар при других условиях: 1) ширина окна – шесть слов; 2) отсутствует ограничение на тип отношения; 3) не учитываются пары, уже попавшие в первый список и 4) отбирались

³ https://en.wikipedia.org/wiki/Main_Page

только единицы, встретившиеся в нескольких словарях английского языка и/или словарях коллокаций.

Затем восемь аннотаторов производили ручную разметку 20 000 первых коллокаций из первого списка и 58 000 пар из второго списка по синсетам WordNet. Аннотаторы пропускали пары с ошибками, вызванными автоматическим парсингом, и пары, в которых хотя бы для одного слова невозможно подобрать ни один из синсетов WordNet, а также идиоматические фразы и многословные именованные сущности.

В общей сложности процесс разметки занял 9 месяцев, в результате получилось 78 000 размеченных лексических комбинаций (пар слов) и 88 019 семантических комбинаций, то есть сочетаний синсетов WordNet.

Авторами были проведены эксперименты с целью оценить качество работы модели WSD на основе знаний, обогащённой информацией из SyntagNet, и сравнить его с качеством той же модели, но дополненной информацией из других лексических баз знаний, а также качеством моделей машинного обучения с учителем. Для экспериментов была выбрана модель на основе персонализированного алгоритма PageRank ([Haveliwala, 2002]), подробно описанная в статье [Agirre et al., 2014], которая применялась к разным лексическим базам знаний, в том числе к SyntagNet. Среднее значение F-1 меры на пяти англоязычных датасетах составило 71.5%; для сравнения модель из [Yuan et al., 2016] на основе рекуррентных нейронных сетей LSTM также продемонстрировала F1-меру, равную 71.5%, в среднем на тех же пяти наборах данных. На материале коллекций данных на разных языках (итальянский, испанский, немецкий, французский) среднее значение F-1 меры модели с информацией из SyntagNet получилось самым высоким среди измеренных (69.3) и превзошло результаты моделей на основе нейронных сетей.

На официальном сайте SyntagNet можно ввести слово или предложение в поисковую строку, выбрать язык запроса (при условии ввода предложения) и получить все коллокации из корпуса для слова или предложения в запросе.

Также на странице проекта можно скачать zip-архив со всеми лексическими комбинациями, размеченными авторами. В архиве содержится файл в формате txt, каждая строка которого соответствует отдельной коллокации в следующем формате:

(1) a. 09827683n 10285313n baby n boy n

б. ID синсета для первого слова ID синсета для второго слова
первое слово часть речи первого слова второе слово часть речи второго слова.

2.2. Размеченные данные для русского языка

Для русского языка на данный момент существует недостаточно семантически аннотированных корпусов, в принципе, о существовании корпусов коллокаций нам вообще неизвестно. Однако стоит упомянуть недавнюю глубокую работу [Bolshina, Loukachevitch, 2020] по автоматическому созданию корпусов с семантической разметкой на основе однозначных родственных слов. Полученные в ходе экспериментов корпуса, а также исходный код доступны по ссылке: https://github.com/loenmac/russian_wsd_data/tree/master/data. Также в исследовании [Kirillovich et al. 2022] авторами были вручную аннотированы тексты средней длины из коллекции OpenCorpora⁴. Разметка производилась с помощью синсетов RuWordNet.

3. Данные и эксперименты

3.1. Цели, данные и предобработка

Авторами статьи был проведён ряд экспериментов по автоматическому порождению семантически размеченного корпуса коллокаций в формате SyntagNet, но для русского языка. Целью создания корпуса является упрощение процесса разметки данных для тестирования и обучения моделей WSD на основе машинного обучения, а также для использования в моделях автоматического разрешения лексической многозначности, основанных на знаниях и правилах.

Разрешение многозначности происходило с помощью родственных слов, найденных в рамках группировок коллокаций по первому и по второму слову. Основная идея метода заключается в предположении, что слова, встречающиеся с одним и тем же словом на одной позиции, имеют некоторую семантическую близость, которая позволит автоматически разрешать многозначность. Ниже приведены примеры таких группировок.

(1) Пример группировки коллокаций по первому слову:

абонентский терминал

абонентский книжка

⁴ <http://opencorpora.org/>

абонентский номер
абонентский база
абонентский служба
абонентский устройство

(2) Пример группировки коллокаций по второму слову:

сушеный абрикос
урожай абрикос
косточка абрикос
заготовка абрикос.

Источником коллокаций выступил корпус текстов новостей за 2017 год объёмом 8 Гб. На первом этапе были извлечены все коллокации из корпуса, в которых оба слова относятся либо к существительным, либо к прилагательным, например:

- (3) а. *местный житель*: прилагательное + существительное
б. *кубок конфедерация*: существительное + существительное

Словосочетания были упорядочены по мере взаимной информации MI3:

$$MI3(w_1, w_2) = \sum_{i=0}^n \sum_{j=0}^m p(x, y) \log\left(\frac{p^3(x, y)}{p(x)p(y)}\right)$$

Далее использовался тезаурус RuWordNet⁵, лексико-семантический ресурс по типу WordNet для русского языка ([Loukachevitch et al., 2016]) для распределения сначала 100 000, а затем 200 000 пар слов на четыре категории: 1) хотя бы одного слова нет в RuWordNet; 2) словосочетание как отдельная единица присутствует в RuWordNet; 3) оба слова однозначны и 4) хотя бы одно слово многозначно. Примеры пар из каждой категории приведены ниже:

- (4) а. Хотя бы одного слова нет в тезаурусе: **неиммиграционный виза, шаговый доступность, вотум недоверие, фазовый автофокус**;
б. Словосочетание есть в тезаурусе как отдельная единица: *отопительный сезон, прибор учет, куриный яйцо*;
с. Оба слова однозначны: *краеведческий музей, декан факультет, пятизвездочный отель*;

⁵ <https://ruwordnet.ru/ru>

d. Хотя бы одно слово многозначно: *мокрый снег, прием граждан, сигнал светофор*.

Статистика по этому этапу Распределение первых 100 000 и 200 000 пар по этим группам приводится в таблице 1:

Таблица 1

	Словосочетание целиком есть в RuWordNet	Хотя бы одного слова нет в RuWordNet	Оба слова однозначные	Хотя бы одно слово многозначное
100 000 пар	1989 (1.989%)	18660 (18.66%)	18950 (18.95%)	60401 (60.401%)
200 000 пар	2409 (1.2%)	33123 (16.56%)	38352 (19.18%)	126116 (63.06%)

Словосочетания из первой категории сразу помещаются в корпус в формате, схожем с SyntagNet:

слово₁ слово₂ <ID синсета RuWordNet для слова₁> <часть речи для слова₁> <ID синсета RuWordNet для слова₂> <часть речи для слова₂>.

Например:

(5) петербургский метро 2642-A Adj 192-N N.

Коллокации из категории 4 (хотя бы одно слово многозначно, и оба слова есть в RuWordNet) являются целевыми, то есть нуждаются в разрешении многозначности для занесения в корпус.

3.2. Использование родственных слов для разрешения многозначности

Во всех следующих этапах принимали участие только пары слов из четвёртой категории. Сначала они были сгруппированы двумя разными способами: 1) по первой лексеме и 2) по второй лексеме.

Затем группы коллокаций по первому и второму слову были дополнены теми парами слов, которые вошли в категорию коллокаций, содержащихся в RuWordNet как отдельная единица, а также коллокациями из RuWordNet, которые также размечены по значениям в самом тезаурусе и первое или второе слово в которых совпадают с тем словом, по которому организована соответствующая группа.

Например, для коллокации *абрикосовый цвет*, в которой каждое слово по отдельности соотнесено с тремя синсетами RuWordNet, в тезаурусе уже есть разметка по значениям:

(6) абрикосовый цвет 109498-A 106944-N.

На следующем этапе для разрешения многозначности слов в рамках коллокаций производился поиск близких и дальних родственников слов в рамках осуществлённых ранее группировок по тезаурусу RuWordNet: синонимов, гиперонимов, гипонимов, так называемых «дальних родственников» (более подробно описывается ниже). Для каждого слова каждой пары слов в рамках группы сохранялся список синонимов, гипонимов и гиперонимов, а также массив значений – синсетов RuWordNet, за которые «голосуют» эти родственники. Так, для отношения синонимии такими синсетами признаются общие для синонимов синсеты, для гиперонимии и гипонимии – те значения, которые связаны соответствующим отношением с гиперонимом и гипонимом соответственно. Подход к решению задачи автоматического порождения корпусов с семантической разметкой на основе слов с родственными значениями описан в статье [Bolshina, Loukachevitch, 2020]: авторы используют однозначные родственные слова для автоматического разрешения многозначности в процессе автоматической разметки по значениям.

Например, в группе коллокаций по первому слову *абстрактный* есть пара (*абстрактный, живопись*), и в данной группе для второго слова – *живопись* – нашлись следующие синонимы: *полотно* и *картина*, причём оба синонима голосуют за одно конкретное значение целевой лексемы – 6001-N ('произведение живописи'), а именно данный синсет входит в пересечение наборов синсетов для всех трёх лексем. В другой группе коллокаций, сформированной уже вокруг второго слова, у пары (*район, авария*) для первой лексемы *район* был обнаружен гипероним *место*, голосующий за значение 106611-N ('место, местность').

В качестве «дальних родственников» извлекались синсеты, отстоящие от целевого слова на два отношения по дереву тезауруса, рассматривались следующие типы таких цепочек отношений:

- 1) $s_1 \xrightarrow{\text{гипероним}} s_2 \xrightarrow{\text{гипероним}} s_3$;
- 2) $s_1 \xrightarrow{\text{гипоним}} s_2 \xrightarrow{\text{гипоним}} s_3$;
- 3) $s_1 \xrightarrow{\text{гипероним}} s_2 \xrightarrow{\text{гипоним}} s_3$;
- 4) $s_1 \xrightarrow{\text{гипоним}} s_2 \xrightarrow{\text{гипероним}} s_3$.

3.3. Методы взвешивания

После того как в каждой группе по всем парам была собрана информация о родственных связях входящих в их состав многозначных слов, а также о том, за какие

значения они голосуют, встала задача разрешить многозначность слов, а именно выбрать только одно значение и, соответственно, синсет RuWordNet. Авторами были протестированы разные подходы к выбору значения.

Самый простой способ – это отобрать те пары, у которых для каждого слова в итоге осталось по одному значению. Так происходит, когда все родственные слова голосуют за одно-единственное значение слова либо когда слово изначально было однозначным. Однако число таких пар оказалось невелико: для выборки из 200 000 пар всего 22 073. Поэтому были применены разные алгоритмы взвешивания голосов для значений, полученных от родственников.

3.3.1. Простой алгоритм взвешивания

Простой алгоритм взвешивания заключался в выборе того значения, которое получило максимальное число голосов от значений-родственников. Для тех многозначных слов, где такое значение только одно, алгоритм возвращает это значение, иначе возвращает None.

3.3.2. Взвешивание на основе Shortest is-a-Path

Также в процессе экспериментов применялся алгоритм взвешивания на основе расстояний между синсетами по тезаурусу RuWordNet, а именно пути от целевого значения до значения родственника, голосующего за это значение [Liu et al., 2007]. Для каждой пары значение-родственник подсчитывался путь по следующей формуле:

$$\text{path}(a, b) = \frac{1}{\text{shortest is-a path}(a,b)}.$$

Далее значения счётчиков по каждому значению взвешиваются с помощью полученных значений путей по тезаурусу до родственника.

3.3.3. Взвешивание на основе предобученных векторов Word2Vec

Для данного метода взвешивания использовались вектора сокращённой размерности Word2Vec ([Mikolov et al., 2013]) для русского языка с ресурса RusVectōrēs⁶ ([Kutuzov, Kuzmenko, 2017]), обученные на материале Национального Корпуса Русского Языка (НКРЯ) и русского сегмента Wikipedia за ноябрь 2021 года (размер корпуса 1.2 миллиарда слов). Размерность векторов равнялась 300, для получения векторов применялся алгоритм Continuous Bag-of-Words (CBOW). На основе данной модели для каждой пары значение многозначного слова – голосующее за него

⁶ <https://rusvectors.org/ru/>

родственное значение подсчитывалась **косинусная мера близости** (*cosine similarity*) их векторов Word2Vec, если таковые имеются в модели по формуле:

$$\text{cosine similarity} = \sum_{i=1}^n A_i B_i / \sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}.$$

Счётчики голосов за то или иное значение взвешивались с помощью значения косинусной меры близости для значения и голосующего за него родственника.

3.3.4 Взвешивание на основе предобученных векторов FastText

Также был протестирован алгоритм взвешивания на основе косинусной меры близости предобученных векторов сокращённой размерности FastText⁷ ([Bojanowski et al., 2017]). При подсчёте использовалась та же формула, что в подпункте 3.3.3.

3.3.5. Взвешивание на основе векторов BERT

Кроме того, в экспериментах использовались вектора BERT ([Devlin et al., 2018]): контекстно-свободные от DeepPavlov через библиотеку transformers для языка Python и вектора для целевого многозначного слова в контексте коллокации. Далее также подсчитывается косинусная мера двух векторов: для многозначного слова в контексте коллокации и для слова с родственным значением также в контексте его коллокации.

3.4. Подготовка тестовой выборки и оценка качества разных методов

Для сравнения разных методов взвешивания была произведена ручная разметка выборки объёмом 1735 пар, которая принималась за «золотой стандарт» для сравнения с результатом работы описанных выше алгоритмов взвешивания и извлечения коллокаций для корпуса. При этом для всех моделей на первом шаге работы отбирались те пары, в которых у обоих слов многозначность снята полностью, то есть в результате подсчёта и взвешивания голосов по всем значениям слова в итоге можно выбрать одно с наибольшим весом, оно и идёт в результат работы алгоритма. Для оценки качества моделей подсчитывались стандартные метрики: полнота, точность и F1-мера. Результаты приведены в таблице 2:

Таблица 2

	Число пар с полностью снятой	Качество разметки (accuracy)	Precision	Recall	F1-мера

⁷ <https://github.com/facebookresearch/fastText>

	многозначностью				
<i>До взвешивания</i>	3 287	10.37 %	0.89	0.1	0.19
<i>Простой алгоритм взвешивания</i>	23 679	68.59 %	0.82	0.69	0.75
<i>До взвешивания + простой алгоритм</i>	26 966	78.96 %	0.83	0.79	0.81
<i>Взвешивание на основе is-a Path</i>	30 997	80.06 %	0.82	0.8	0.81
<i>Взвешивание на основе косинусной близости векторов Word2Vec</i>	32 117	77.06 %	0.8	0.77	0.79
<i>Взвешивание на основе косинусной близости векторов FastText</i>	32 780	80.17 %	0.81	0.8	0.81
<i>is-a Path + Word2Vec</i>	33 114	77.35 %	0.8	0.77	0.79
<i>is-a Path + FastText</i>	33 803	80.63 %	0.81	0.81	0.81
<i>Взвешивание на основе косинусной близости контекстных векторов BERT</i>	32 808	80.35 %	0.81	0.8	0.81

Как можно увидеть из таблицы многим методам взвешивания удалось достичь F1-меры 80% - на основе контекстных векторов BERT, векторов сокращённой размерности FastText, кратчайшего пути по дереву тезауруса. Эти же методы позволяют добавить порядка 30 000 новых размеченных пар в корпус.

4. Анализ ошибок

Ошибки, то есть коллокации со снятой многозначностью, автоматическая семантическая разметка которых не совпадает с «золотым стандартом», распределены следующим образом между семью приведёнными в таблице моделями:

Таблица 3

Сколько моделей допустили ошибку	Количество коллокаций
1	6
2	35
3	18
4	9
5	37
6	15
7	257
	377

Как видно из таблицы, в 257 коллокациях (68.2%) ошиблось семь моделей из семи, то есть в большинстве случаев коллокации вызывают затруднения у всех моделей.

В данном разделе приводится анализ только таких коллокаций, в разметке которых ошиблись все модели. Всего их 257 (68.2% от всех пар с ошибками).

К категории пар, при разметке которых ошиблись все изученные методы, этой категории относятся в том числе те, для лексем в составе которых в рамках группировок не нашлось подходящих слов-родственников, которые бы помогли при снятии многозначности. Например, при разметке коллокации (*'праздничный', 'концерт'*) все семь методов допустили ошибку в определении значения второго слова, определив его как 'скандал, ссора' вместо правильного 'концерт, концертная программа', причём в группировку по первому слову *праздничный* вошло 80 пар. Другой пример: все модели приписали неверный тег значения второй лексеме в коллокации (*'солдатский', 'каша'*) – 'месиво (полужидкая смесь)', потому что в рамках группировки коллокаций по первому слову подобрались семантически разные лексемы: *котелок, слава, форма, письмо, привал, шинель, казарма, могила, вдова, орден*.

В рамках коллокации (*'всероссийский', 'олимпиада'*) всеми моделями вторая лексема была размечена как 'олимпийские игры', однако в новостных текстах, из

которых был составлен корпус-источник коллокаций данная словосочетание чаще употребляется в контексте школьных олимпиад. В группе пар слов по первому слову *всероссийский* оказалось несколько пар слов, обозначающих спортивные мероприятия всероссийского уровня: *тренировка, марафон, автопробег, состязание, регата, чемпионат, гонка, турнир*. Попали в группу также пары со вторыми словами, тематически относящимися к школьной предметной области – школа и урок, но их было всего две и они не связаны родственными отношениями с искомой лексемой.

Некоторые ошибки в разметке были вызваны ошибками на этапе лемматизации, например в коллокации ('*должное*', '*образ*') первая лексема была определена как существительное *должное*, образованное с помощью конверсии прилагательного, но на самом деле в данной коллокации первая лексема должна быть *должный* (прилагательное).

5. Заключение

В статье был рассмотрен подход к автоматическому порождению корпуса коллокаций в формате SyntagNet с семантической аннотацией. Такие корпуса могут значительно облегчить процесс подготовки тренировочных и тестовых коллекций для моделей машинного обучения, а также применяться в качестве источника знаний в моделях на основе правил и знаний. Создание размеченных корпусов коллокаций подразумевает принятие гипотезы «Одно значение на коллокацию». Для автоматического разрешения лексической многозначности в рамках коллокаций производились группировки по первой и второй лексеме и поиск родственных слов в рамках данных группировок – синонимов, гиперонимов, гипонимов, более дальних родственников, отстоящих от целевой лексемы на два шага по дереву тезауруса. Значения родственных слов голосуют за тот или иной синсет многозначного слова в рамках коллокации.

Те пары, у которых для обоих слов после подбора родственных значений осталось по одному синсету, добавлялись в корпус. К остальным коллокациям применялись различные методы взвешивания голосов, полученных от родственников: косинусная мера близости векторов Word2Vec, FastText, контекстных векторов BERT, кратчайшего пути по тезаурусу. Большинство методов достигают F1-меры в 80% и позволяют расширить корпус словосочетаний на 30 000 пар.

В результате анализа ошибок была выявлена следующая закономерность: 68.2% коллокаций, в которых хотя бы один метод допустил ошибку, оказались проблемой для

всех моделей. Среди причин неточностей в семантической разметке ошибки, допущенные на этапе лемматизации, состав группировок по первому или второму слову – слова из разных предметных областей, преобладание определённой области.

В качестве направлений будущих исследований можно изучить зависимость качества разметки от объёма группы, оценить вклад разных семантических отношений, а также оценить качество работы методов на более объёмной коллекции с ручной разметкой.

Список литературы

1. Agirre E., López de Lacalle O., Soroa A. Random walks for knowledge-based word sense disambiguation // *Computational Linguistics*. 2014. V. 40. №. 1. P. 57-84.
2. Blloshmi R. et al. IR like a SIR: Sense-enhanced information retrieval for multiple languages // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021. P. 1030-1041.
3. Bojanowski P. et al. Enriching word vectors with subword information // *Transactions of the Association for Computational Linguistics*. 2017. V. 5. P. 135-146.
4. Bolshina A., Loukachevitch N. Monosemous Relatives Approach to Automatic Data Labelling for Word Sense Disambiguation in Russian // *Linguistic Forum 2020: Language and Artificial Intelligence*. 2020. P. 12-13.
5. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // *arXiv preprint arXiv:1810.04805*. 2018.
6. Gale W. A., Church K., Yarowsky D. One sense per discourse // *Speech and Natural Language: Proceedings of a Workshop*. 1992. P. 233-237.
7. Haveliwala T. H. Topic-sensitive PageRank // *Proceedings of the 11th international conference on World Wide Web*. 2002. P. 517-526.
8. Kirillovich A. et al. Sense-Annotated Corpus for Russian // *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*. 2022. P. 130-136.
9. Kutuzov A., Kuzmenko E. WebVectors: a toolkit for building web interfaces for vector semantic models // *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Revised Selected Papers 5*. Springer International Publishing. 2017. P. 155-161.
10. Leech G. N. 100 million words of English: the British National Corpus (BNC) // *Language Research*. 1992.
11. Liu X. Y., Zhou Y. M., Zheng R. S. Measuring semantic similarity in WordNet // *2007 International Conference on Machine Learning and Cybernetics*. 2007. V. 6. P. 3431-3435.
12. Loukachevitch N. et al. Creating Russian WordNet by conversion // *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*. 2016. P. 405-415.

13. Manning C. D. et al. The Stanford CoreNLP Natural Language Processing Toolkit // Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations. 2014. P. 55-60.
14. Martinez D., Agirre E. One Sense per Collocation and Genre/Topic Variations // 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000. P. 207-215.
15. Maru M. et al. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. P. 3534-3540.
16. Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. 2013. V. 26.
17. Pu X. et al. Integrating weakly supervised word sense disambiguation into neural machine translation // Transactions of the Association for Computational Linguistics. 2018. V. 6. P. 635-649.
18. Seifollahi S., Shajari M. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction // Journal of Intelligent Information Systems. 2019. V. 52. P. 57-83.
19. Yarowsky D. One sense per collocation // Proceedings of the workshop on Human Language Technology. 1993. P. 266-271.
20. Yuan D. et al. Semi-supervised word sense disambiguation with neural models // Proceedings of COLING. 2016. P. 1374-1385.

References

1. Agirre E., López de Lacalle O., Soroa A. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 2014, v. 40, №. 1, pp. 57-84.
2. Blloshmi R. et al. IR like a SIR: Sense-enhanced information retrieval for multiple languages. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1030-1041.
3. Bojanowski P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017, v. 5, pp. 135-146.

4. Bolshina A., Loukachevitch N. Monosemous Relatives Approach to Automatic Data Labelling for Word Sense Disambiguation in Russian. *Linguistic Forum 2020: Language and Artificial Intelligence*, 2020, pp. 12-13.
5. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.
6. Gale W. A., Church K., Yarowsky D. One sense per discourse. *Speech and Natural Language: Proceedings of a Workshop*, 1992, pp. 233-237.
7. Haveliwala T. H. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 517-526.
8. Kirillovich A. et al. Sense-Annotated Corpus for Russian. *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, 2022, pp. 130-136.
9. Kutuzov A., Kuzmenko E. WebVectors: a toolkit for building web interfaces for vector semantic models. *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Revised Selected Papers 5. Springer International Publishing*, 2017, pp. 155-161.
10. Leech G. N. 100 million words of English: the British National Corpus (BNC). *Language Research*, 1992.
11. Liu X. Y., Zhou Y. M., Zheng R. S. Measuring semantic similarity in WordNet. *2007 International Conference on Machine Learning and Cybernetics*, 2007, v. 6, pp. 3431-3435.
12. Loukachevitch N. et al. Creating Russian WordNet by conversion. *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, 2016, pp. 405-415.
13. Manning C. D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations*, 2014, pp. 55-60.
14. Martinez D., Agirre E. One Sense per Collocation and Genre/Topic Variations. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 207-215.
15. Maru M. et al. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3534-3540.
16. Mikolov T. et al. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013, v. 26.

17. Pu X. et al. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 2018, v. 6, pp. 635-649.
18. Seifollahi S., Shajari M. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *Journal of Intelligent Information Systems*, 2019, v. 52, pp. 57-83.
19. Yarowsky D. One sense per collocation. *Proceedings of the workshop on Human Language Technology*, 1993, pp. 266-271.
20. Yuan D. et al. Semi-supervised word sense disambiguation with neural models. *Proceedings of COLING*, 2016, pp. 1374-1385.

ОБ АВТОРАХ:

Зарипова Диана Александровна – аспирант Кафедры теоретической и прикладной лингвистики филологического факультета МГУ (e-mail: diana.ser.sar96@gmail.com),
+7 929 572-03-17

Лукашевич Наталья Валентиновна – доктор технических наук, кандидат физико-математических наук, профессор Кафедры теоретической и прикладной лингвистики филологического факультета МГУ, ведущий научный сотрудник НИВЦ МГУ (e-mail: louk_nat@mail.ru)

ABOUT THE AUTHORS:

Diana Zaripova – PhD student, Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University (e-mail: diana.ser.sar96@gmail.com),
+7 929 572-03-17

Natalia Loukachevitch – Prof. Dr., Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University (e-mail: louk_nat@mail.ru)