Санкт-Петербургский государственный университет

ПЕТРУШЕНКО Лада Сергеевна

Выпускная квалификационная работа

Параметры лексических конструкций в предсказывающих языковых моделях для русского языка

Уровень образования: магистратура

Направление: 45.04.02 «Лингвистика»

Образовательная программа: ВМ.5805 «Компьютерная и прикладная лингвистика»

Профиль: «Компьютерная лингвистика и интеллектуальные технологии»

Научный руководитель:
кандидат филологических наук, доцент,
Кафедра математической лингвистики,
Митрофанова Ольга Александровна

Рецензент:
кандидат технических наук, доцент,
Высшая школа лингводидактики и перевода,
Гуманитарный институт, СПбПУ
Коган Марина Самуиловна

Санкт-Петербург

2023

Аннотация

В данной работе исследуются оптимальные параметры, влияющие на результаты обучения предсказывающих языковых моделей, и оценивается их способность определять возможные именные словосочетания на основании текстов на русском языке.

В работе рассматриваются теоретические стороны вопроса, такие как классификации и свойства лексических конструкций, а также существующие методы к их выделению. В практической части исследования описаны эксперименты по поиску лучшего набора параметров для различных корпусов, отличающихся по размеру и жанру. Продемонстрировано, как результаты могут быть применены к задачам автоматического выделения именных словосочетаний и определения значений лексических функций.

Ключевые слова: лексические конструкции, предсказывающие языковые модели, параметры предсказывающих моделей, автоматическое выделение конструкций.

This graduation paper addresses the issue of finding optimal parameters that can influence the training results of predictive language models. In the paper, we evaluate the ability of such models to extract noun phrases based on the texts written in Russian.

We analyze the theoretical issues related to the question such as classifications of lexical construction, their properties, and the existing algorithms of construction detection. In the practical part of the paper, we give the description of our experiment which consists of finding the best set of parameters for various corpora of various genres. It is demonstrated how the results can be applied to the tasks of detecting noun phrases and the values of lexical functions.

Оглавление

Введение

С появлением **Word2Vec** в 2013 году дистрибутивные языковые модели стали все чаще применяться исследователями и разработчиками для решения лингвистических задач. Дистрибутивно-семантические модели показали свою эффективность в тех случаях, когда требуется выявить семантические отношениями между лексическими единицами, а также обработать большие массивы данных — **Word2Vec** оказался быстрее, чем счетные модели. Помимо этого, появились предобученные дистрибутивные модели, которые позволили напрямую делать запросы для извлечения нужных данных без предварительной подготовки и обработки больших корпусов самими исследователями, что позволило сэкономить время.

Один из аспектов применения подобных моделей — предсказание парадигматических отношений, на котором акцентировали большое внимание русскоязычные исследователи. Синтагматика традиционно представляла меньший интерес, но в последнее время список задач, в которых требуются данные о лексических конструкциях, стал активно пополняться. Среди них можно выделить перифразирование, суммаризацию, анализ тональностей, упрощение, генерацию текстов (в частотности заголовков к ним) и создание диалоговых систем. Корректность выделения различных типов лексических конструкций может повлиять на результат выполнения задачи, и в связи с этим возникает потребность в изучении потенциала предсказывающих моделей применительно к этой задачи. В данной работе мы акцентируем внимание на именных словосочетаниях, параметрах обучения, которые позволяют выделить их наиболее точно, а также рассматриваем полученные сочетания на предмет устойчивости и композиционности.

Таким образом, **актуальность** работы обусловлена широким спектром задач, при решении которых нужно рассматривать лексические конструкции, и увеличением роли дистрибутивного подхода в них. Потенциал методов, главенствоваших в области ранее — подходов на основе правил и счетных

алгоритмов — изучен достаточно хорошо, их недостатки описаны и существует ряд гибридных моделей, которые их компенсируют. Тем не менее, для предсказывающих моделей, пришедших в область около десяти лет назад и набирающих все большую популярность, подробных исследований не было. Определение их возможностей позволило бы улучшить качество выполнения задач, которые стоят перед лингвистами-исследователями и разработчиками.

Целью нашего исследования является определение оптимальных параметров обучения предсказывающих моделей и оценка их потенциала в выявлении лексических конструкций в текстах на русском языке. Для достижения данной цели требуется выполнение следующих **задач**:

1. Сформировать теоретический фундамент исследования, включающий в себя краткое изложение основных положений лингвистики конструкций и существующих классификаций;
2. Проанализировать существующие методы автоматического выделения конструкций и рассмотреть уже разработанные инструменты;
3. Выбрать языковые модели для исследования, а также определить параметры, с которыми будут обучены модели;
4. Провести эксперименты по обучению и применить полученные модели к задачам предсказания именных словосочетаний и значений лексических функций, а также описать свойства предсказанных конструкций;

Объектом в нашем исследовании выступают именные лексические конструкции, предсказанные дистрибутивными моделями Word2Vec и FastText. **Предметом** исследования являются параметры обучения моделей и способы выделения конструкций. **Материалом** для исследования служат сегменты русскоязычных корпусов «Тайга» и Lib.ru.sec. Нами были отобраны тексты художественного, поэтического, научно-популярного и новостного сегментов.

Теоретическая значимость данной работы заключается в разностороннем исследовании факторов, влияющих на качество обучения моделей и разработке алгоритма, который позволяет выделить оптимальные параметры обучения для модели. **Практическая значимость** работы заключается в возможности применения полученных результатов и сведений для решения задач обработки естественного языка, что было продемонстрировано нами на примере задач предсказания именных конструкций и значений лексических функций при заданном аргументе.

Наше исследование состоит из введения, трех глав, заключения, списка литературы и приложений. В первой главе мы рассматриваем свойства и классификации лексических конструкций. Во второй главе мы изучаем существующие методы извлечения лексических конструкций и работы, посвященные этой задаче. В третьей главе мы описываем эксперимент по поиску лучших условий для обучения моделей, рассматриваем влияние параметров на результаты, а также применяем их в задачах предсказания именных словосочетаний и поиска значения лексических функций.

1. Лексические конструкции и сочетаемость единиц

1.1. Грамматика конструкций

Наиболее известной теорией, центральным элементом которой выступают лексические конструкции, является **Грамматика конструкций** (Construction Grammar). Впервые ее идея была описана американским исследователем Ч. Филлмором в 1980-е гг. применительно к выражениям английского языка [Fillmore 1985]. Согласно Ч. Филлмору, **лексическая конструкция** представляет собой последовательность единиц, среди которых одна или несколько — целевые слова, организующие контекст вокруг себя, а другие — переменные, только заполняющие ее слоты [Fillmore 1988].

Ч. Филлмор приводит основные наблюдения, касающиеся конструкций:

- они несут не только синтаксическую, но и семантическую и прагматическую информацию;
- элементы, входящие в их состав, могут самостоятельно выступать в роли конструкций;
- их значение может быть идиоматично и некомпозиционно.

В дальнейшем теорию Ч. Филлмора развивает, например, А. Голдберг [Goldberg 1995; Goldberg 2006; Goldberg 2013]. В неформальном определении конструкций исследователь отмечает особенности их формы и значения: так, по его наблюдениям, либо форма синтагмы должна быть построена нерегулярно (не по общим правилам языка), либо ее значение должно быть некомпозиционно [Goldberg 1996, с. 68].

В рамках Грамматики конструкций план содержания и план выражения взаимосвязаны и оказывают влияние друг на друга. Конструкция закрепляет сочетаемость целевого слова в конкретном значении, а ее форму фиксируют в первую очередь лексические элементы, входящие в ее состав; помимо этого, можно выделить пропозициональные, грамматические и лексические ограничения сочетаемости. Фразеологизмы могут быть зафиксированы как частично («*jog someone's memory*», где средний элемент конструкции может

быть заменен, например, на местоимение), так и полностью («*let alone*») [Goldberg 2013].

А. Стефанович и С. Грис — исследователи, являющиеся сторонниками корпусного подхода к анализу конструкций. Рассматривая основные положения Грамматики конструкций, они продолжали идею совместного изучения синтаксиса и семантики лексических сочетаний.

В работе «Collostructions: Investigating the Interaction of Words and Constructions» А. Стефанович и С. Грис представляют новый подход к коллокационному анализу [Stefanowitsch, Gries 2003]. Они описывают **коллострукционный анализ** (collostructional analysis), который позволяет обратить большее внимание на связь отдельных лексем и низлежащих грамматических структур. Объектом в коллострукционном анализе выступают лексемы и конструкции. В общем смысле оценке подвергается сила тяготения лексем к слотам в определенной конструкции с точки зрения вероятности заполнения таковых конкретными лексемами. Само понятие коллострукции авторы определяют как комбинацию лексемы, которая тяготеет к определенной конструкции (**коллексема**, англ. *collexeme*), и конструкции, которая связана с определенной лексемой (**коллострукт**, англ. *collostruct*).

Для того, чтобы вычислить силу коллострукции, авторы статьи берут абсолютные частоты вхождения лексемы *L* в конструкцию *C* и в другие конструкции в составе корпуса, а также частоту употреблений конструкции *C* без *L* и частоту употребления всех других конструкций, которые не включают в свой состав *L*. Затем применяется точный тест Фишера [Stefanowitsch, Gries 2003, с. 218]. Выбор этого критерия обусловлен тем, что с его помощью можно игнорировать выбросы в виде служебных слов, а также он не зависит от какого-то определенного объема выборки. Среди недостатков подхода исследователи отмечают высокую сложность вычислений.

1.2. Фразеологические единицы по В.В. Виноградову

В.В. Виноградов выделяет три класса устойчивых лексических конструкций: **фразеологические сращения**, **фразеологические единства** и **фразеологические сочетания** [Виноградов 1977].

Фразеологические сращения обладают такими свойствами, как семантическая неделимость и немотивированность: «*собаку съел (в чем-л.)*», «*бить баклуши*», «*вверх тормашками*», «*попасть впросак*» и т.д. Значение конструкции не связано со значениями ее индивидуальных компонентов и является произвольным. Идиоматичность сращений может быть подчеркнута нарушением грамматической структуры: например, сочетание «*как ни в чем не бывало*» в современном языке воспринимается как наречие, его форма зафиксирована, хотя в текстах XIX в. глагол «*бывать*» должен был быть согласован с подлежащим.

Неделимость значения подобной конструкции может быть связана с устареванием слов и грамматических форм, входящих в его состав: «*во всю Ивановскую*», «*ничтоже сумняшеся*» и т.д. Важную роль может играть также эмоционально-экспрессивный компонент высказывания: «*чего доброго*», «*вот тебе и на*».

В некоторых случаях семантическое единство компонентов приводит к тому, что при сохранении опорных смысловых единиц возможно изменение внешней формы конструкции: замена вторичных элементов или их пропуск без потери смысла (ср. «*держать в ежовых рукавицах*», «*держать в ежовых*», «*держать в ежах*»).

Второй тип сочетаний, который выделяет В.В. Виноградов, — **фразеологические единства**. Конструкции этого типа семантически некомпозиционны, но их отличает слабая мотивированность, обусловленная влиянием компонентов на общее значение — то, что исследователь называет «внутренним стержнем фразы»: «*семь пятниц на неделе*», «*плыть по течению/плыть против течения*», «*держать камень за пазухой*», «*выносить сор из избы*» и т.д.

Фразеологические единства могут по внешней форме совпадать со свободными сочетаниями (ср. «*намылить голову (кому-л.)*» в значении «*ругать, бранить*» и «*намылить голову (кому-л.)*» в прямом значении). При этом сами выражения часто имеют определенную эмоционально-экспрессивную окраску, которая способствует их фиксации в языке. Идиоматичность выражения передается также строго установленным порядком слов: например, при перестановке элементов в конструкции «*белены объелся*» выражение воспринимается скорее буквально, чем в переносном смысле. Таким образом, конструкция зачастую становится неделима синтаксически — в ней не допускается пропуск компонентов или их замена синонимами.

Третий и наиболее многочисленный класс конструкций в русском языке — **фразеологические сочетания**. В таких сочетаниях значения отдельных компонентов выделяются и равноправны, но при этом сами сочетания несвободны. Например, мы можем сказать «*тоска берет*» или «*зависть берет*», но не можем сказать «*радость берет*». По наблюдениям В.В. Виноградова, многие слова в русском языке не имеют свободного значения: например, значение слова «*потупить*» может быть определено лишь из контекста и при сопоставлении устойчивых сочетаний, содержащих это слово, с синонимичными (ср. «*потупить глаза*» и «*опустить глаза*»). В связи с этим исследователь делает вывод, что у доли слов в русском языке ограничена способность формировать семантические отношения. Отчасти из сказанного вытекает и другая особенность подобных конструкций — мы можем заменять элементы с несвободным значением на синонимичные (ср. «*затронуть гордость*» и «*задеть гордость*»).

В классификации В.В. Виноградова в качестве основного критерия деления конструкций выступает их композиционность. Подводя итог исследованию, он подчеркивает, что развитие фразеологических теорий и концепций должно быть связано с развитием теорий, сосредоточенных на синтаксисе.

1.3. Синтагматическое взаимодействие значений по В.Г. Гаку

В седьмой главе монографии «Языковые преобразования» В.Г. Гак изучает проблемы семантической синтагматики и сочетаемости слов [Гак 1998]. Исследователь отмечает, что в отличие от семиотики, где семантика и синтаксис скорее обособлены друг от друга, в естественных языках «синтаксис семантичен, то есть его категории и элементы соотносятся определенным образом со внешними объектами, а семиотика синтаксична, то есть отражает отношения между символами-обозначениями» [Гак 1998, с. 272]. Этим утверждением В.Г. Гак подчеркивает важность изучения одной области в контексте другой.

Исследователь обусловливает закономерности сочетаемости взаимодействием уровней действительности, мышления (плана содержания) и языка (плана выражения). При исследовании сочетаемости В.Г. Гак оперирует понятием **номинации**. Согласно Лингвистическому энциклопедическому словарю, номинация — «образование языковых единиц, характеризующихся номинативной функцией, т. е. служащих для называния и вычленения фрагментов действительности и формирования соответствующих понятий о них в форме слов, сочетаний слов, фразеологизмов и предложений» [Лингвистический энциклопедический словарь 1990].

Когда речь идет о сочетаемости в рамках синтагм, имеет смысл рассматривать номинации в контексте. Автор выделяет три подвида зависимости номинации от окружения:

1. ***синтагматически обусловленная*** номинация подразумевает прямую зависимость номинации от ее окружения;
2. ***ситуационно обусловленная*** номинация зависит от знания участниками диалога ситуации;
3. ***повторная*** номинация связана с предыдущим наименованием элемента ситуации в контексте диалога.

В нашей работе интерес представляет синтагматически обусловленная номинация. Классическим примером семантической сочетаемости можно

назвать конструкцию, в которой у обоих компонентов есть общая категориальная сема — так называемая **классема**. Так, например, сочетание «*простуженная женщина*» допустимо в русском языке, так как в обоих словах присутствует сема одушевленности. Напротив, сочетание «*простуженный стул*» невозможно — второй элемент конструкции не обладает свойством неодушевленности, и у слов нет общих классем. К категориальным семам можно отнести также пол, возраст, материальность и т.д.

В.Г. Гак расширяет понятие связующего семантического компонента — помимо классем в качестве таких компонентов могут выступать любые семы родового значения. Подобные связующие компоненты значения исследователь называет **синтагмемами**.

По наблюдениям исследователя, реализация семантического компонента в синтагме определяет три варианта взаимодействия элементов в ней.

1. ***Семантическое согласование*** — наличие одной семы в обоих элементах конструкции: например, «*птица летит*».
2. ***Семантическое несогласование*** — пропуск общего семантического компонента в одном из элементов: например, «*птица приближается*». Такое возможно только в устойчивых сочетаниях: так, в предложении «*Возле стены находился шкаф*» глагол «*находиться*» синонимичен глаголу «*стоять*», так как шкаф чаще всего занимает вертикальное положение.
3. ***Семантическое рассогласование*** — наличие несовместимых компонентов в синтагме. Для иллюстрации этого явления В.Г. Гак рассматривает компонент значения «*протяженность во времени*». С точки зрения его наличия в существительных, их можно разделить на те, которые обладают им («*путешествие*», «*урок*»), и те, которые им не обладают («*дом*»). Некорректной будет конструкция «*во время дома*». Тем не менее, с предлогом

«*после*» слово «*дом*» может обозначать определенный период времени; таким образом, предлог может влиять на значение элемента конструкции в контексте.

В.Г. Гак подчеркивает, что законы семантического согласования нельзя назвать универсальными. Так, можно привести примеры предложений из русского и французского языков, где в одном и том же предложении наблюдаются разные виды семантического согласования: «*Кошка кормит котят*» и «*La chatte allaiteses petits*». В предложении на французском не повторяется элемент значения «*кошка*» [Апресян 1995, с. 80].

1.4. Законы согласования значений по Ю.Д. Апресяну

Отталкиваясь от понятия некомпозиционности, или неаддитивности, Ю.Д. Апресян исследует законы согласования значений [Апресян 1995]. Ю.Д. Апресян рассматривает описанное Ш. Балли понятие грамматического плеоназма, при котором одно понятие должно быть выражено в рамках синтагмы как минимум два раза [Балли 1955]. Также он приводит пример с расширением этого понятия Ч. Осгудом, М. Мастерман и А. К. Жолковским на семантику [Osgood 1957; Masterman 1957; Жолковский 2004]. Исследователь подчеркивает важность теории В.Г. Гака в изучении семантического плеоназма — в ней подробно описано явление повторения компонента значений в элементах конструкции.

Ю.Д. Апресян отмечает, что в большинстве свободных сочетаний значение строится по аддитивному, или композиционному, принципу (например, «*зеленый лист*»), в то время как фразеологизированные конструкции отличаются неаддитивностью. Ю.Д. Апресян формулирует **правила зачеркивания** — один из способов неаддитивного сложения значений. Им выделяются два вида правил:

1) **зачеркивание в результате несочетамости значений**;
2) **зачеркивание совпадающей части значений**.

Для иллюстрации первого случая Ю.Д. Апресян рассматривает два предложения, которые могут быть представлены в виде структуры «*обучать X-а Y-у*»: «*обучать студентов математике*» и «*обучать студентов водить комбайн*». По результатам первого процесса предполагается, что студенты будут *знать* математику, а по результатам второго — что они будут *уметь* управлять рабочей машиной. Таким образом, сочетание «*обучать X-а Y-у*» можно трактовать как «*каузировать X знать или уметь Y*». Компонент значения «*обучать*» зависит от значения элемента *Y*: он должен либо представлять собой информацию, которую можно *знать*, либо навык или деятельность, которую можно *уметь выполнять*. В контексте эти компоненты значения обычно взаимоисключаемы, и при реализации одна из частей зачеркивается.

Зачеркивание совпадающей части значения происходит, когда у обоих элементов конструкции *AB* есть, помимо своего индивидуального, частного значения, некоторый компонент общего значения. Это можно представить в виде следующей записи: *A = 'TY', B = 'YX', AB = 'PY + YX' = 'PYX'*, где *Y* — общий семантический компонент. Ю.Д. Апресян называет подобное явление семантической гаплологией и приводит три ее разновидности:

- семантический повтор обеспечивает связность текста, поэтому второй элемент синтагмы не может быть удален: «*врач лечит*», «*птица летит*»;
- повторение значения необходимо для его усиления: «*совсем слепой*», «*подниматься выше*», «*отступать назад*»;
- семантическая гаплология в собственном смысле слова: «*процедура анализа*» — «*анализ*», «*чувство уважения*» — «*уважение*», «*гречневая крупа*» — «*греча*», «*лингвистическая наука*» — «*лингвистика*»; также конструкции: «*идти быстрым шагом*» — «*идти быстро*», «*улыбаться ласковой улыбкой*» — «*улыбаться ласково*».

Ю.Д. Апресян приводит ряд исключений, при которых повторяющийся компонент значения не редуцируется — в частности, к нему относится значение высокой степени проявления свойства: ср. «*высокие цены*» и «*очень высокие цены*».

Помимо правил зачеркивания, исследователем отмечаются такие возможности влияния на значение, как дополнение синтагмы семантическим компонентом, его замена другим компонентом (т.е. реализация и правил зачеркивания, и правил добавления) и семантический сдвиг.

Как и другие исследователи, теории которых были рассмотрены выше, Ю.Д. Апресян отмечает взаимосвязь синтагматики и семантики и необходимость их совместного изучения.

1.5. Типы устойчивых конструкций по Л.Н. Иорданской и И.А. Мельчуку

В третьей главе монографии «Смысл и сочетаемость в словаре» Л.Н. Иорданская и И.А. Мельчук определяют основные типы устойчивых сочетаний и рассматривают **лексические функции** — инструмент, разработанный с целью описания одного из подвидов устойчивых сочетаний — **коллокаций** [Мельчук, Иорданская 2007].

Прежде всего авторы выделяют две группы лексико-синтаксических двухкомпонентных выражений, способных существовать в некотором языке *L*, — свободные и фиксированные словосочетания (фраземы). Под свободным сочетанием понимается сочетание, построенное по следующим двум принципам.

1. Означаемое конструкции *'S'* должно быть построено регулярно и без ограничений по формуле *'S' = 'A⊕B'*, где *'A'* и *'B'* — означаемые лексем *A* и *B* языка *L*, а ⊕ служит для обозначения операции языкового объединения.
2. Означающее словосочетания */S/* построено регулярно и без ограничений по формуле */S/ = /A⊕B/*, где *'A⊕B'* — семантическое

представление из означающих */A/* и */B/* лексем *A* и *B* языка *L*, а ⊕ служит для обозначения операции языкового объединения.

При этом условие неограниченности подразумевает под собой то, что конструкция состоит из произвольных элементов, а условие регулярности — соответствие при построении общим правилам языка *L*. Любое свободное сочетание должно удовлетворять обоим условиям одновременно.

В случае с некомпозиционным сочетанием может не выполняться либо первое условие (тогда второе не выполняется по умолчанию и речь идет о **прагматемах**), либо только второе (тогда мы имеем дело с **семантическими фраземами**).

У **прагматем** возможно только одно означаемое, которое, тем не менее, построено по стандартным правилам языка. Есть два типа прагматем: те, у которых означающее построено, хотя и регулярно, но с ограничениями, и те, у которых означающее построено без ограничений. В качестве примера явлений первого класса Л.Н. Иорданская и И.А. Мельчук приводят сочетания «*best before…*» (*англ.*), «*à consommer avant…*» (*франц.*), «*mindestens haltbar bis…*» (*нем.*), которые можно считать переводными эквивалентами для русскоязычного «*срок годности*»: в перечисленных конструкциях и смысловой компонент, и текстовая форма строго зафиксированы и ограничены ситуацией. Мы не можем говорить о замене формы на синонимичную. Примером конструкций второго класса могут послужить такие сочетания, как «*No talking please*» и «*Please do not talk*», где допускается, что один смысл может быть выражен рядом синонимичных выражений, при этом любое выражение из этого ряда будет уместно употребить в конкретной ситуации.

Возможна и альтернативная ситуации, когда означаемое построено нерегулярно, но без ограничений. В таком случае речь идет о **семантических фраземах**. Для более точного определения конструкций подобного типа Л.Н. Иорданская и И.А. Мельчук вводят понятие **семантической доминанты**: так, рассматривая семантическую фразему *AB 'S'*, мы можем

представить ее означаемое как *'S'* = *'P'('A')*, где *'P'* представляет собой предикат, а *'A'* — аргумент предиката, который и является семантической доминантой.

Отталкиваясь от понятия выше, можно выделить три типа фразем.

1. **Полные фраземы** (**идиомы**): «Идиома в языке *L* — это семантическая фразема, такая что в состав ее означаемого *'S'* не входит ни одно из означаемых *'A'*, *'B'* её частей — лексем *A* и *B* — в качестве семантической доминанты» [Иорданская, Мельчук 2007, с. 237]. В качестве примеров могут выступать сочетания «*spill the beans*» (русск. «*проболтаться*»), «*pull someone's leg*» (русск. «*разыгрывать (кого-л.)*»), «*private eye*» (русск. «*частный детектив*») и т.д. Отметим, что в последнем случае означаемое *'B'* входит в состав *'S'* (англ. «*private*» в значении русск. «*частный*»), но не является семантической доминантой, в то время как в первых двух конструкциях ни *'A'*, ни *'B'* не входят в состав *'S'* ни в каком виде.
2. **Полуфраземы** (**коллокации**): «Коллокация *AB 'S'* в языке *L* — это семантическая фразема языка, такая что в состав ее означаемого *'S'* входит означаемое одной из составляющих ее лексем — допустим, означаемое лексемы *A* — в качестве семантической доминанты и некоторое дополнительное означаемое *'C'* [так что *'S'* = *'A*$\oplus$*C'*, где *'A'* — семантическая доминанта], а лексема *B* выражает это *'C'* в зависимости от лексемы *A*» [Иорданская, Мельчук 2007, с. 238-239]. Таким образом, в состав полуфраземы обязательно входит **ключевое слово** — лексическая единица, означаемое которой остается неизменным в *'S'* и которая определяет способ выражения значения *'C'* через единицу *B*. Примером могут послужить сочетания «*land a job*» (русск. «*найти работу*»), «*crack a joke*» (русск. «*отпустить шутку*»), «*deeply moved*» (русск. «*глубоко*

тронут»), «*french window*» (русск. «*французское окно*») и т.д. Ключевыми словами в данных сочетаниях являются единицы «*job*», «*joke*», «*moved*» и «*window*», соответственно.

3. **Квазифраземы** (**квазиидиомы**): «Квазиидиома *AB 'S'* в языке *L* — это семантическая фразема, которая удовлетворяет одновременно двум следующим условиям: 1. Ее означаемое *'S'* включает в себя означаемые *'A'* и *'B'* ее составных частей — лексем *A* и *B*, а также дополнительное означаемое *'C'*, отличное от *'A'* и *'B'*. 2. Ни *'A'*, ни *'B'* не является семантической доминантой в *'S'*» [Иорданская, Мельчук 2007, с. 242]. В качестве примеров авторы монографии приводят сочетания «*bacon and eggs*» (русск. «*яичница с беконом*»), «*shopping center*» (русск. «*торговый центр*») и т.д.

При этом в описании и представлении коллокаций большую роль играют лексические функции, которые будут рассмотрены нами в следующем разделе.

1.6. Понятие лексической функции

Описание **лексических функций** было представлено в работах А.К. Жолковского и И.А. Мельчука [Жолковский, Мельчук 1965; Mel'cuk, Zolkovskij 1970]. В общем определении лексическая функция *f* соотносит с некоторой единицей, ее аргументом *A*, множество лексических коррелятов $\{A_i\}$, таких что они:

- представляют собой значение функции *f* при аргументе *A*;
- определяются в зависимости от *A*;
- несут смысл, предполагаемый *f* — с этим пунктом связана и природа коррелятов: синтагматическая либо парадигматическая [Иорданская, Мельчук 2007].

Можно выделить два основных подтипа лексических функций: параметры и замены. **Параметры** фиксируют окружение, т.е. синтагматику. **Замены** фиксируют парадигматические варианты.

По наблюдениям исследователей существует около 60 стандартных лексических функций. Л.Н. Иорданская и И.А. Мельчук приводят некоторые из наиболее распространенных:

- Функция S_2 несет смысл «*то, что подвергается*» или «*тот, кто подвергается*»: например, S_2(врач) = пациент.
- Функция *Magn* имеет значение интенсивности действия или приведения чего-либо в высшую степень; например, *Magn(осудить) = решительно*.
- Функция *Syn* ставит в соответствие лексической единице ряд синонимов: например, *Syn(лингвистика) = языкознание, языковедение,*
- и т.д.

Применимость или неприменимость лексической функции по отношению к определенной лексеме зависит от семантического класса рассматриваемой единицы. Лексические функции, применимые к большому количеству лексических единиц, называют **нормальными**; лексические функции, применимые только к одной-двум близким по смыслу лексическим единицам, принято называть **вырожденными**.

Стандартная лексическая функция должна удовлетворять трем условиям: гомогенности, максимальности и фразеологичности [Иорданская, Мельчук 2007, с. 251-252].

В нашем исследовании интерес представляют функции-параметры.

Вывод к главе 1

В главе 1 нами были рассмотрены основные положения лингвистики конструкций. Так, мы привели описание конструкции с точки зрения

Грамматики конструкций, а также рассмотрели, как другие лингвисты адаптировали теорию Ч. Филлмора к своим исследованиям.

Помимо этого, нами были рассмотрены классификации устойчивых лексических конструкций по В.В. Виноградову, В.Г. Гаку, Ю.Д. Апресяну, а также Л.Н. Иорданской и И.А. Мельчуку. Отдельное внимание было уделено изучению лексических функций — инструмента для описания коллокаций.

Подводя итог, можно отметить, что в классификациях в качестве критерия деления лексических конструкций выступают композиционность их значения и степень их устойчивости в языке; они же представляют собой основные свойства подобных выражений. Композиционность конструкции подразумевает возможность вывести ее значение из значений составляющих, а устойчивость — степень ее идиоматичности. При этом всеми исследователями подчеркивалась важность совместного изучения лексических конструкций как с точки зрения синтаксиса, так и с точки зрения семантики, поскольку эти два уровня языка оказывают сильное влияние друг на друга.

2. Автоматическое выделение конструкций

Исследователями были разработаны методики, позволяющие автоматически выделять лексические конструкции различных типов. К подзадачам автоматического выделения конструкций можно отнести извлечение коллокаций, именованных сущностей, синтаксических групп и ключевых выражений, тем не менее, подходы для их выделения схожи и их можно объединить в следующие группы:

- метод ***на основе правил*** предполагает предварительное создание шаблонов целевых конструкций для автоматизированного поиска; такой подход реализован, например, в Национальном корпусе русского языка (НКРЯ), где есть функция лексико-грамматического поиска: с ее помощью можно выделить все конструкции с заданными лексико-семантическими и морфологическими тегами (*lex*, *sem*, *gr*).
- ***семантически*** обусловленные модели и меры определяют возможные конструкции путем анализа окружений слов. Можно выделить две их разновидности:
    - ***счетные***, или ***статистические***, алгоритмы позволяют выделить устойчивые конструкции при помощи мер ассоциации и обращения к подсчету различных простейших статистик корпуса;
    - ***дистрибутивно-семантические,*** или ***предсказывающие***, позволяют непосредственно предсказывать возможные конструкции на основании данных о дистрибуции единиц и вычислении степени семантического сходства между ними;
- ***гибридный***, или лингвостатистический, подход сочетает данные о лексико-грамматических шаблонах и корпусных статистиках;

Границу между счетными и дистрибутивно-семантическими моделями можно провести, основываясь на возможности проинтерпретировать координаты их векторов. В счетных векторных моделях каждой лексической единице отведено только одно место в векторе, т.е. мы сразу можем понять, за что отвечает то или иное его значение. В связи с этим подобные векторные представления получаются разреженными, так как для единиц, которые не встречаются рядом с данной, значение координаты может быть равно 0. Дистрибутивно-семантические, или предсказывающие, модели представляют собой модели распределенных векторных вложений. Они решают проблему разреженности, заполняя и улучшая во время обучения значения всех координат вектора, которые, тем не менее, невозможно проинтерпретировать однозначно в связи с тем, что выбор важных параметров для отображения лингвистических свойств с вектором определенной размерности происходит непосредственно во время обучения.

В следующем разделе мы кратко рассмотрим статистический и гибридный подходы, а затем перейдем к дистрибутивно-семантическим моделям, представляющим наибольший интерес для нашего исследования.

2.1. Счетный, или статистический, подход

В рамках **статистического** подхода можно выделить такие меры ассоциации, как **PMI**, **log-likelihood**, **критерий Хи-квадрат**, **t-score**, **logDice** и др. С их помощью можно выделить различные устойчивые сочетания — в основном сочетания коллокационного типа. Для иллюстрации опишем, как происходит подсчет значений для метрик **PMI** и **Хи-квадрат**.

Коэффициент поточечной взаимной информации (PMI) определяет силу синтагматической связи двух слов *a* и *b* через подсчет вероятности их совместной встречаемости по отношению к их независимым вероятностям вхождения в корпус.

Критерий Хи-квадрат опирается на матрицу совместной встречаемости и сравнение наблюдаемых частот в корпусе с ожидаемыми.

Значимость сочетания определяется путем сопоставления того, как часто его элементы встречаются совместно по сравнению с их случайным распределением в документе, или тексте.

Другая модель — **TF-IDF (Term Frequency — Inverse Document Frequency)** — позволяет построить разреженные векторные представления единиц корпуса и выделить слова и сочетания, характерные для конкретного текста или документа. Наибольший вес при использовании TF-IDF получают такие лексические единицы, которые часто встречаются в рассматриваемом тексте корпуса по сравнению с остальными его документами. Достоинством этого метода является то, что с его помощью можно отсеять служебные слова и стоп-слова: они будут присутствовать во всех текстах и получат наименьший вес.

Для компонента TF происходит простой подсчет абсолютной частоты рассматриваемой лексической единицы в конкретном документе. В свою очередь IDF представляет собой обратную частоту, с которой единица встречается во всех документах коллекции. Именно второй компонент отвечают за установление баланса между частотными и низкочастотными единицами при подсчете окончательного значения.

Примерами языковых моделей, которые опираются на статистику, могут послужить **латентный семантический анализ** (**LSA**, **Latent Semantic Analysis**) [Landauer и др. 1998], модель **HAL** (**HyperSpace Analogue to Language**) [Lund, Burgess 1996], а также вдохновленная ими модель **COALS** [Rohde и др. 2005]. Стоит отметить, что в статистических языковых моделях обычные счетные показатели векторов зачастую трансформируются исследователями, чтобы лучше отражать семантику: так, например, в модели **COALS** абсолютные частоты совместной встречаемости слов преобразуются в коэффициенты корреляции, а затем отсеиваются те значения, где коэффициент меньше нуля; от оставшихся значений берется корень.

Что касается проблемы разреженности векторов, хотя модель **LSA** подразумевает использование метода сингулярного разложения матрицы для

ее решения, она все еще актуальна для большинства статистических моделей по причине, которую мы приводили выше — если слова не встречаются совместно в корпусе, их значение в векторе будет равно 0. Как уже было упомянуто ранее, эта проблема решается дистрибутивными моделями, о которых пойдет речь в разделе 2.3.

2.2. Гибридный, или лингвостатистический, подход

К **гибридным** алгоритмам можно отнести такие алгоритмы, как **RAKE** [Rose и др. 2009] и **YAKE** [Campos и др. 2020].

Так, **RAKE — Rapid Automatic Keyword Extraction** — был изначально разработан для текстов на английском языке и используется для выделения различных n-граммных ключевых выражений, которые не включают в себя знаки пунктуации и стоп-слова. Кандидаты в ключевые выражения определяются при помощи специального набора разделителей для конструкций. Все кандидаты равноценно рассматриваются с точки зрения частоты их совместной встречаемости с другими словами. Каждому кандидату присваивается вес, который определяется как сумма трех параметров:

- частота;
- степень, т.е. мера совместной встречаемости;
- отношение частоты к степени.

Затем кандидаты ранжируются в зависимости от полученного веса, и выдается список ключевых выражений, получивших наибольшее значение.

Алгоритм был адаптирован к русскому языку [Moskvina и др. 2017; Moskvina и др. 2018].

Аналогичным способом действует алгоритм **YAKE**, однако вес в нем считается более сложным образом. Так, учитываются:

- нормализованная частота;
- положение кандидата в документе;
- число предложений, в состав которых входит кандидат;

- число употреблений кандидата в корпусе, когда он написан с заглавной буквы;
- ходство кандидата со словами из списка стоп-слов.

Преимуществом гибридных моделей является то, что они позволяют обратить большее внимание на лингвистические аспекты изучаемых единиц. Тем не менее, зачастую такие модели не являются универсальными, поскольку сосредотачиваются лишь на одном или нескольких родственных языках.

2.3. Предсказывающие модели

Применение **дистрибутивно-семантических моделей** (ДСМ) — наиболее актуальная для нашего исследования методика извлечения лексических конструкций, так как они напрямую решают задачу предсказания контекстов без использования различных эвристик, которые используются для счетных моделей.

Для объяснения основного принципа предсказывающего типа языковых моделей можно привести классическую цитату Дж. Фёрса: «You shall know a word by the company it keeps» [Firth 1957]. В их основе лежит гипотеза о дистрибуции значения слова по контексту, или возможности выведения значения из контекста. Ее можно сформулировать следующим образом: «Степень семантического сходства между двумя любыми языковыми выражениями *A* и *B* есть функция сходства языковых контекстов, в которых *A* и *B* встречаются» [Lenci 2008].

При обучении информация об окружениях токена накапливается, и значения представляются векторами в многомерном пространстве. Среди преимуществ предсказывающих моделей можно выделить относительно небольшую вычислительную стоимость, а также высокую точность в отображении различных свойств слов — морфологических, синтаксических и семантических — по сравнению со счетными, или статистическими, моделями.

В качестве примеров предсказывающих моделей можно привести неконтекстуализированные, такие как **Word2Vec** [Mikolov и др. 2013a; Mikolov и др. 2013b] и **FastText** [Joulin и др. 2018; Bojanowski и др. 2017], а также контекстуализированные, такие как **BERT** [Devlin и др. 2018] и **ELMo** [Peters и др. 2018].

2.3.1. Модели семейства Word2Vec

Рассмотрим основные характеристики неконтексуализированных моделей на примере **Word2Vec**. В основе **Word2Vec** лежат два алгоритма обучения — Continuous Bag-of-Words (CBOW) и Skip-gram (SG). По результатам обучения обеих моделей мы получаем векторные представления всех токенов корпуса, выведенные на основе изучения контекста. Тем не менее, их отличает подход к обучению: CBOW предсказывает целевое слово, исходя из векторов его контекста, в то время как SG берет вектор целевого слова, чтобы предсказать его окружение. Ни в CBOW, ни в SG не учитывается порядок слов в заданном контекстном окне.

Пусть последовательность w_1, w_2, …, w_T представляет собой последовательность всех токенов в корпусе. В таком случае целевая функция для алгоритма Continuous Bag-of-Words может быть определена следующим образом:

$$\frac{1}{T}\Sigma_{t=1}^{T} log\ p(w_t|w_{t-c}\dots w_{t+c}) \quad (1)$$

В формуле t — целевое слово, а c — ширина контекста. В то же время целевая функция для Skip-gram может быть представлена в виде:

$$\frac{1}{T}\Sigma_{t=1}^{T}\Sigma_{-c<j<c;j\neq 0} log\ p(w_{t+j}|w_t) \quad (2)$$

В работе исследователей М. Барони, Дж. Дину и Дж. Кружевски проводятся эксперименты с обучением и оценкой 48 моделей **Word2Vec** (в частности CBOW) на различных задачах обработки естественного языка [Baroni и др. 2014]. Результаты сопоставляются с результатами 36 стандартных счетных моделей — **PMI** и **Local Mutual Information** (схожий с

Log-likelihood) — векторы которых различными способами трансформируются. Среди классических задач для предсказывающих языковых моделей, которые рассматривают исследователи, есть следующие:

- **оценка степени семантического сходства** (semantic relatedness): информанты получают пары слов, для каждой из которых нужно оценить степень сходства по числовой шкале; подсчитывается корреляция между усредненной оценкой информантов и косинусом между векторами слов, полученными моделью;
- **выявление синонимов** (synonym detection): модели на вход поступает ряд слов, для каждого из которых она должна выявить, какой из представленных кандидатов является истинным синонимом (т.е. семантически сочетается с ним);
- **категоризация понятий** (concept categorization): на вход модели поступают понятия-существительные, которые модель должна сгруппировать вместе (например, «*helicopters*» и «*motorcycles*» должны быть ближе к классу «*vehicle*», чем к какому-либо другому); итоговые кластеры оцениваются с точки зрения того, сколько экземпляров на самом деле не подходят в те группы, куда они попали;
- **избирательное предпочтение** (selectional preferences): информанты размечают ряд существительных, которые могут выступать либо в качестве субъектов, либо в качестве объектов по отношению к некоторому глаголу; затем векторы для обоих актантных позиций усредняются, и целевое слово сравнивается с полученным усредненным вектором; таким образом, проверяется, может ли оно занимать слот конструкции;
- **поиск слова по аналогии** (analogy): данную задачу легче всего проиллюстрировать на конкретном примере: предположим, у нас есть пара слов «*брат*» и «*сестра*» и слово «*внук*». Модель

должна найти такое слово, которое будет относится к слову «*внук*» как «*сестра*» к «*брату*» — «*внучка*».

Перечисленные подзадачи часто становятся компонентами более сложных лингвистических задач. Задача и возможность поиска по аналогии появилась вместе с предсказывающими языковыми моделями [Mikolov и др. 2013a]. По результатам исследования, М. Барони, Дж. Дину и Дж. Кружевски заключают, что счетные и предсказывающие языковые модели выдают сопоставимые результаты только для задачи избирательного предпочтения, а в остальных задачах дистрибутивные модели намного превосходят статистические по качеству предсказаний.

Word2Vec также используется в задачах определения лексических функций. Например, О. Колесникова и А. Гельбух используют **Word2Vec** для предсказания лексических функций, связывающих элементы коллокаций, состоящих из глагола и существительного [Kolesnikova, Gelbukh 2020]. Схожим образом, отталкиваясь от изучения лексических функций, в работе «Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases» исследователи во главе с M. Hartung сосредотачиваются на предсказании конструкций с атрибутивным значением [Hartung и др. 2017].

2.3.2. Модели типа Трансформер

В отличие от предыдущего типа моделей, модели типа Трансформер являются контекстуализированными, или контекстно-зависимыми. Впервые идея моделей Трансформер была описана в 2017 году в работе «Attention is All You Need» [Vaswani и др. 2017]. Прежде чем построить вектор для конкретного слова, модель анализирует весь контекст (предложение), в котором оно находится. Трансформеры применяют механизм внимания (англ. *self-attention*), чтобы каждое слово могло отображать информацию о своем окружении.

Среди моделей Трансформеров получили особую известность **BERT** (Bidirectional Encoder Representations from Transformers) [Devlin и др. 2018], его разновидности, такие как **ALBERT** [Lan и др. 2019] и **RoBERTa** [Liu и др. 2019]; **ELMo** [Peters и др. 2018] и т.д.

Некоторыми исследователями был проведен анализ того, насколько контекстуализированные представления отображают различные языковые явления и закономерности. Например, были изучены морфологические свойства векторов [Edmiston 2020] и был сделан вывод, что контекстуализированность Трансформеров позволяет им разрешать случаи морфологической неоднозначности в большинстве случаев. В свою очередь, исследователем А. Тенней совместно с коллегами был проведен анализ того, насколько контекстуализированные представления отображают морфологию, синтаксис и семантику [Tenney и др. 2019]. В работе рассматриваются такие задачи, как частеречная разметка, разметка по именованным сущностям и семантическим ролям, кореференция, аннотирование с точки зрения синтаксических зависимостей, и т.д. Исследователи приходят к выводу, что Трансформеры могут быть более эффективны в задачах, которые завязаны на анализе синтаксиса, но не результативнее неконтекстуализированных моделей в семантически-ориентированных задачах.

Понимание того, в какой степени языковые модели, покрывают информацию, находящуюся в тексте, важно для оценки того, как они способны предсказывать различные типы устойчивых сочетаний. Так, в работе Л. Эспиноза-Анке и коллег рассматривается способность языковых моделей предсказывать конструкции и лексические функции [Espinosa-Anke и др. 2021]. Авторы исследования составляют датасет употреблений коллокаций в корпусе, при этом употребления разделены на 16 категорий в зависимости от лексической функции, которой можно описать отношения между коллокатом и целевым словом. В работе исследователи решают две задачи:

- **извлечение коллокаций** (collocation retrieval): оценивается, как хорошо модель может восстановить коллокат при наличии целевого слова (например, «*rain*»), лексической функции («*intense*») и последовательности потенциальных коллокатов: «*heavy*», «*torrential*», «*violent*» и т.д.; используется маскированное языковое моделирование (англ. *masked language modelling, MLM*);
- **присвоение тегов классов коллокациям в контексте** (in-content collocation categorization): оценивается способность языковой модели определить лексическую функцию коллокации при наличии предложения, в котором она находится.

По результатам исследования Л. Эспиноза-Анке и коллеги делают вывод, что некоторые типы глагольных лексических функций, такие как *Oper1()*, *Real1()* и *Real2()*, хорошо предсказываются моделями типа **BERT**. В то же время функции, которые описывают связь между прилагательным и существительным, например, *Magn()*, *Ver()* и *Bon()*, получили одни из наиболее низких значений при оценке качества обучения.

Как мы уже отметили ранее, основным отличием моделей типа Трансформер от предыдущего поколения предсказывающих моделей является их контекстуализированность. Еще одним важным отличием является обращение со словами, которых нет в словаре корпуса (англ. *out-of-vocabulary words*, *OOV*). Неконтекстуализированные модели в большинстве случаев способны составлять векторные представления только для тех слов, которые есть в исходном датасете; в то же время такие модели, как **BERT**, могут вычислять векторы для несловарных слов. Большинство подобных моделей используют особый алгоритм токенизации, который заключается в кодировании пар байт (англ. «*byte-pair encoding*») [Gage 1994]. Данный подход к компрессии данных позволяет группировать отдельные символы и обозначать их другим символом, тем самым разбивая слово на ряд таких последовательностей.

2.4. Способы оценки результатов обучения векторных моделей

Для оценки качества обучения модели может быть применена процедура **псевдодизамбигуации** (pseudo-disambiguation) [Gale и др. 1992; Dagan и др. 1993; Pereira и др. 1993]. Тестовые данные при таком подходе представляют собой множество комбинаций из трех токенов:

- целевое слово;
- слово-кандидат для формирования лексической конструкции, с которым целевое слово встречается в корпусе или может употребляться с ним;
- слово-кандидат для формирования лексической конструкции, которое не употребляется с целевой единицей в данном корпусе.

Перед моделью ставится задача определить, какое из слов-кандидатов способно составить целевую конструкцию с целевым словом. Это происходит путем оценки силы семантической связи, т.е. сопоставлением векторных представлений.

В зависимости от задачи процедура псевдодизамбигуации может быть дополнена различными ограничениями — например, в отношении частот вхождения целевого слова в корпус и биграмм с корректным словом-кандидатом; в отношении частоты некорректного слова-кандидата, а также лексического состава тестового множества. Например, в задачах разрешения неоднозначности слова-кандидаты могут быть семантически связаны, и псевдодизамбигуация позволяет проверить, способна ли модель выявить по контексту особые лексические значения.

Для вычисления сходства векторов используются **метрики**, такие как косинусная мера, евклидово расстояние, квадрат евклидова расстояния, расстояние городских кварталов, корреляция векторов, расстояние Чебышева и т.д. В нашем исследовании была поставлена задача рассмотреть несколько метрик — были выбраны две, в основе которых лежит евклидово расстояние, и две, которые на нем не базируются. Рассмотрим формулы для них:

- **Евклидово расстояние**:

$$Dist(a,b) = \sqrt{\Sigma_{i=1}^{N}(a_i - b_i)^2}, \text{ где} \quad (3)$$

a и b — векторные представления слов; а a_i и b_i — их значения по измерению i.

- **Квадрат евклидова расстояния**:

$$Dist(a,b) = \Sigma_{i=1}^{N}(a_i - b_i)^2, \text{ где} \quad (4)$$

a и b — векторные представления слов; а a_i и b_i — их значения по измерению i.

- **Мера косинусной близости**:

$$Sim(a,b) = cos(a,b) = \frac{\Sigma_{i=1}^{N} a_i \times b_i}{\sqrt{\Sigma_{i=1}^{N} a_i^2} \times \sqrt{\Sigma_{i=1}^{N} b_i^2}}, \text{ где} \quad (5)$$

a и b — векторные представления слов; a_i и b_i — их значение по измерению i; $a \times b$ — их скалярное произведение.

- **Корреляция векторов**:

$$Corr(a,b) = cos(a,b) = \frac{\Sigma_{i=1}^{N}(a_i - \bar{a}) \times (b_i - b)}{\sqrt{\Sigma_{i=1}^{N}(a_i - \bar{a})^2} \times \sqrt{\Sigma_{i=1}^{N}(b_i - \bar{b})^2}}, \text{ где} \quad (6)$$

a и b — векторные представления слов; a_i и b_i — их значение по измерению i; $\bar{a}$ и $\bar{b}$ — средние значения по a и b, соответственно; × служит для обозначения скалярного произведения.

Традиционно при оценке сходства векторов используется косинусная мера; как отмечалось некоторыми исследователями, метрики, основывающиеся на евклидовом расстоянии, не способны точно отображать семантические взаимоотношения [Rohde и др. 2005, с. 8]. В то время как евклидова метрика вычисляет *расстояние*, косинусная мера вычисляет *угол* между векторами, и чем выше сходство, тем меньше угол.

Касательно корреляции векторов, эта метрика и мера косинусной близости вычисляются аналогичным образом за исключением компонента подсчета среднего значения. Так, например, авторы методики COALS

используют корреляцию в качестве меры определения сходства и отмечают, что корреляция дает более точные результаты в тех случаях, когда векторы находятся в квадранте положительных абсцисс и ординат (квадранте I) [Rohde и др. 2005].

Вывод к главе 2

В данной главе нами была рассмотрена основная классификация алгоритмов выделения лексических конструкций различных типов — подходы с использованием правил, статистический (счетный), гибридный и предсказывающий (дистрибутивно-семантический).

Были приведены примеры широко применяемых мер ассоциации в статистическом подходе (**PMI**, **критерий Хи-квадрат** и т.д.), а также языковых моделей, которые основываются на счетной информации из корпуса — **LSA** и **HAL**. Помимо этого, мы рассмотрели, что влияет на принцип выбора и вес конструкции в двух гибридных подходах: **RAKE** и **YAKE**.

Большая часть главы посвящена изучению предсказывающих языковых моделей, таких как **Word2Vec**, **FastText**, **BERT**, **ELMo** и т.д. Мы рассмотрели, в чем отличие контекстуализированных моделей от неконтекстуализированных, какие стандартные задачи можно решить посредством дистрибутивно-семантических моделей, а также привели примеры исследований, посвященных анализу векторов с точки зрения сохранения языковой информации. Наконец, мы рассмотрели работы, целью которых является адаптация как контекстуализированных, так и неконтекстуализированных моделей к задаче выявления лексических конструкций и лексических функций.

3. Эксперименты по обучению моделей и анализ результатов

3.1. Лингвистические данные для разрешения задачи псевдодизамбигуации

Для обучения моделей мы использовали фрагменты русскоязычного веб-корпуса «Тайга» [Shavrina, Shapovalova 2017], а также корпус lib.ru.sec [Panchenko и др. 2017].

Исходный корпус «Тайга» включает в себя 6 сегментов: новостной (92 млн токенов), подкорпус художественных текстов (4605 млн токенов), подкорпус социальных сетей (90 млн токенов), подкорпус субтитров (101 млн токенов), поэтический (1130 млн токенов), а также сегмент датасетов, размеченных для конкретных задач, (например, определение удобочитаемости текста, куда входят научно-популярные тексты) (2.5 млн токенов). Для нашего исследования мы взяли три фрагмента:

1. ***Fontanka*** (новостной): 73,140,388 токенов, 3,885,119 предложений;
2. ***Nplus1*** (научно-популярный): 1,667,938 токенов, 72,002 предложения;
3. ***Stihi_ru*** (поэтический): 5,986,693 токена, 421,956 предложений (первые 50,000 текстов);

Помимо этого, была использована часть художественных текстов из ***lib.ru.sec***: 9,669,140 токенов, 677,134 предложения (первые 100 текстов).

Количество токенов во всех случаях указано до проведения предобработки текста.

Новостной и научно-популярный фрагменты были скачаны нами с ресурса, предоставляемого исследователями-создателями корпуса[1]. Оттуда

[1]URL: https://tatianashavrina.github.io/taiga_site/downloads

уже были удалены дубликаты; помимо этого, он размечен морфологически и синтаксически при помощи парсера *UDpipe*[2]. Каждый подкорпус доступен в табличном формате CoNLL-U [Buchholz, Marsi 2006].

В связи с вычислительными ограничениями у нас не было возможности обработать поэтический[3] и художественный[4] корпусы целиком. Для удобства данные были скачаны из открытого доступа на Hugging Face. Hugging Face предоставляет специальные инструменты для языка программирования Python, благодаря которым мы можем выделить только часть желаемого датасета и работать в дальнейшем только с ним.

Поэтический и художественный датасеты были представлены в виде необработанных текстов, поэтому мы разметили их посредством библиотеки *spacy_udpipe*[5], которая является «оберткой» для написанной на языке C *ufal.udpipe*[6]. Итоговые данные были представлены в табличном формате CoNLL-U аналогично разметке новостного и научно-популярного датасетов.

Наличие четырех корпусов разных жанров позволяет нам проверить работу нашего алгоритма на текстах с разными характерными особенностями и выявить лучший набор параметров для каждого из них. Мы допускаем, что выбранные нами корпусы могут быть попарно похожи с точки зрения языка и лексического состава:

- новостной и научно-популярный;
- поэтический и художественный.

Помимо этого, данные отличаются друг от друга по объему, что позволяет нам проверить зависимость качества обучаемых моделей от количества данных.

---

[2]URL: https://ufal.mff.cuni.cz/udpipe

[3]URL: https://huggingface.co/datasets/IlyaGusev/stihi_ru

[4]URL: https://huggingface.co/datasets/IlyaGusev/librusec

[5]URL: https://spacy.io/universe/project/spacy-udpipe

[6]URL: https://ufal.mff.cuni.cz/udpipe

3.2. Предобработка данных для обучения моделей

Основные шаги в предобработке каждого корпуса — это:

1. извлечение разметки отдельных предложений в корпусе или разметка корпуса (в случае работы с непредобработанными текстовыми коллекциями);
2. удаление токенов, которые есть в словаре стоп-слов;
3. сохранение каждого предложения в виде списка токенов с частями речи, например, «*дом_NOUN*»;
4. сохранение обработанного корпуса в виде списка списков в формат pickle для дальнейшей работы.

Для предобработки новостного и научно-популярного корпусов нам было необходимо открыть директорию с текстовыми файлами и подать каждый из них поочередно на вход парсеру библиотеки *conllu*[7]. В результате подобной обработки мы получаем conllu-объект, к разметке которого можно обратиться через специальные атрибуты.

Для обращения к поэтическому и художественному корпусам мы использовали библиотеку *datasets*[8], которая позволяет работать с датасетами с ресурса Hugging Face. Каждый текст в корпусе обрабатывается при помощи модели проекта Universal Dependencies[9] для русского языка, .

В качестве стоп-слов мы используем списки токенов из *nltk*[10] (151 токен) и Yandex Wordstat[11] (263 токена). Помимо этого, мы удаляем из корпуса все символы с помощью *string.punctuation* и отдельно включаем как стоп-токен специальный символ '--'.

---

[7]URL: https://pypi.org/project/conllu/

[8]URL: https://huggingface.co/docs/datasets/index

[9]URL: https://universaldependencies.org/treebanks/ru_taiga/index.html

[10]URL: https://www.nltk.org/

[11]URL: https://yandex.ru/support/direct/keywords/keywords.html

При комбинировании всех токенов, которые нужно удалить, и после удаления из списка стоп-слов дубликатов мы получаем словарь объемом 363 токена.

Для каждого предложения мы создаем пустой словарь, который затем поэтапно заполняем токенами, которые не были удалены на предыдущем этапе. Для нашего исследования нам важны данные о лемматизированной форме каждого слова и его части речи. Эту информацию мы можем получить, обратившись к атрибутам 'lemma' и 'upos' (для научно-популярного и новостного корпусов) и lemma_ и pos_ (для поэтического и художественного корпусов), соответственно.

В результате мы получаем корпус в виде списка списков — предложений, входящих в его состав. Такой формат нужен нам для этапа обучения моделей. Поскольку преобработка корпусов может занимать большое количество времени, мы предусматриваем возможность обратиться к данным и обучить модель на них сразу и выносим все, что касается предобработки, в отдельный модуль. Для сохранения данных нами был выбран формат pickle. Библиотека *pickle*[12] позволяет провести сериализацию объекта посредством преобразования его в поток байтов, а также предусматривает его десериализацию. Благодаря этому, мы можем восстановить исходный тип данных (в данном случае список), нужный нам для обучения.

3.3. Отбор коллокатов для псевдодизамбигуации

Для оценки обученных моделей мы прибегаем к процедуре псевдодизамбигуации [Букия и др. 2015]. В нашем исследовании на вход моделям поступает набор комбинаций, каждая из которых состоит из слова целевой части речи («ADJ» — прилагательное или «NOUN» —

---

[12]URL: https://docs.python.org/3/library/pickle.html

существительное) в сочетании с корректным и некорректным коллокатами с точки зрения лексической сочетаемости на основании данных из корпуса. Модель должна проанализировать комбинации и оценить, какой из коллокатов ближе для каждого целевого слова. Одна из проблем подхода — это то, что тот коллокат, который является некорректным и не встречается с целевым словом в данных, все равно может быть сочетаем с ним лексически. В нашем тестовом множестве мы стремимся сократить количество таких пар, заранее просматривая и удаляя неподходящие варианты.

Опишем ситуацию, когда мы ищем коллокаты для имени существительного. На первом этапе происходит их отбор. Поскольку в нашей задаче мы используем несколько корпусов, имеет смысл подобрать такие комбинации, которые есть во всех корпусах. Для начала мы фильтруем каждый корпус по той части речи, слова которой будут выступать в роли коллокатов — для существительных это будут, соответственно, прилагательные. Затем мы подсчитываем относительную частоту вхождения в корпус для каждого прилагательного и сохраняем результат в виде отсортированного словаря — эти данные понадобятся нам для сопоставления частот найденных лексем во всех корпусах.

На следующем шаге алгоритм итерируется по словарю, сохраняя биграммы, где целевое слово — существительное, а коллокат — прилагательное. После этого мы подсчитываем частотность употребления всех биграмм при помощи *collections.Counter* и отбираем только те, частота которых больше или равна 8. Выбор данного порога связан с тем, что он позволяет отсеять большое количество спорных комбинаций, когда и корректный, и некорректный коллокаты могли бы сочетаться с целевым словом.

Как только мы получаем список частотных биграмм для каждого корпуса, мы ищем общие биграммы, которые есть во всех четырех текстовых коллекциях, и оставляем только их. Для поиска некорректных коллокатов мы обращаемся к данным об относительных частотах, полученных нами ранее.

Для каждого корректного коллоката мы сравниваем относительные частоты в корпусах и берем максимальную относительную частоту как порог, выше которого не должна быть относительная частота потенциального некорректного коллоката. По этому параметру мы фильтруем созданный ранее словарь с лексемами-кандидатами, а затем из получившегося списка случайным образом выбираем одно прилагательное.

Аналогичным образом происходит поиск коллокатов-существительных для прилагательного. Результаты сохраняются в текстовые файлы *collocates_for_NOUN.txt* и *collocates_for_ADJ.txt*, соответственно; в качестве разделителя выступает знак табуляции. В результате мы получаем 87 комбинаций для имен существительных и 68 комбинаций для имен прилагательных (155 комбинаций всего). Ниже мы приводим примеры коллокатов для имен прилагательных:

век_NOUN, классик_NOUN, прошлый_ADJ
ветер_NOUN, отвал_NOUN, сильный_ADJ
рука_NOUN, сокол_NOUN, левый_ADJ
расстояние_NOUN, самолечение_NOUN, большой_ADJ
человек_NOUN, лужица_NOUN, взрослый_ADJ

Также следует привести примеры коллокатов для имен существительных:

молодой_ADJ, взаимный_ADJ, человек_NOUN
тихий_ADJ, кудрявый_ADJ, океан_NOUN
сильный_ADJ, четвероногий_ADJ, ветер_NOUN
главный_ADJ, фешенебельный_ADJ, герой_NOUN
сегодняшний_ADJ, психический_ADJ, день_NOUN

В каждом случае правильным будет выбор первой лексемы. Перед обучением из корпусов удаляются все предложения, в состав которых входят корректные биграммы.

С полным набором данных, которые мы использовали при дизамбигуации, можно ознакомиться в Приложении 1.

3.4. Отбор параметров для обучения моделей

Нами был проведен ряд экспериментов с параметрами для обучения неконтекстуализированных моделей **Word2Vec** и **FastText**. Эксперименты с моделями из семейства **BERT** не были предусмотрены в рамках нашего исследования в силу имеющихся данных о сложности контроля над их предсказаниями, проявляющейся в сходных задачах [Belyi и др. 2023]. Для обучения моделей была выбрана библиотека *gensim*[13], которая включает в себя имплементацию как первой, так и второй моделей. Мы брали во внимание следующие параметры:

- метрика для оценки степени сходства векторов: евклидово расстояние, квадрат евклидова расстояния, косинусная мера и корреляция векторов, формулы для которых и пояснения к ним были приведены нами в предыдущем разделе: {euclidean, sqeuclidean, cosine, correlation}; в нашем исследовании для вычисления этих мер мы используем инструменты библиотек для языка программирования Python *numpy*[14] и *scipy*[15];
- **size** — размерность вектора; возможные значения: {100, 150, 200, 250, 300}.
- **window** — ширина контекстного окна, т.е. максимальное расстояние между текущим и предсказанным словом в тексте; возможные значения: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}.
- **min_count** — фильтрация слов по принципу их частотности, так, в разных экспериментах мы не учитывали слова, которые встретились в корпусе меньше указанного значения: {5, 10, 15}.

---

[13]URL: https://radimrehurek.com/gensim/

[14]URL: https://numpy.org/

[15]URL: https://scipy.org/

- **cbow_mean** — способ подсчета вектора контекстных признаков при обучении алгоритма CBOW: {0, 1}. Если указано 1, будет произведена операция усреднения контекстных векторов; если 0 — суммирование.
- **sorted_vocab** отвечает за ранжирование словаря корпуса по частоте: {0, 1}. При указании значения 1 (используется по умолчанию) словарь будет отсортирован по убыванию частоты перед назначением индексов слов.
- **max_vocab_size** ограничивает использование оперативной памяти во время построения словаря: {None, 30000, 60000}; если число уникальных слов больше, чем указанный порог, то отсеиваются нечастотные токены.

Для первоначального обучения мы используем все возможные комбинации первых 6 параметров: их количество составило 2400 для каждого корпуса и для каждой модели (итого 57600 комбинаций на все корпусы и модели). Затем для корпусов, которые показали низкие результаты, мы дополнительно настраиваем при обучении параметр **max_vocab_size**.

3.5. Оценка обученных моделей

Оценка алгоритма производится при помощи следующих метрик: **точность** (precision), **полнота** (recall) и **F-мера** (F-score).

Точность отвечает за долю корректных предсказаний по отношению ко всем предсказаниям и вычисляется по формуле:

$$precision = \frac{tp}{tp + fp}, \text{ где} \quad (7)$$

tp — количество раз, когда модель выдала корректный вариант, т.е. указала на принадлежность объекта положительному классу, а *fp* — количество раз, когда модель выдала некорректный вариант. При суммировании двух метрик можно получить общее количество предсказаний на датасет.

Полнота отражает, какую долю из всех возможных единиц, принадлежащих классу, мы предсказали верно:

$$recall = \frac{tp}{tp + fn}, \text{ где} \quad (8)$$

tp — количество раз, когда модель выдала корректный вариант, а *fn* — количество раз, когда модель должна была предсказать положительный класс, но не сделала этого.

F-мера представляет собой среднее гармоническое между метриками точности и полноты и вычисляется по формуле:

$$f - score = \frac{2(precision \times recall)}{precision + recall} \quad (9)$$

В случае с неконтекстуализированными моделями эти метрики адаптируются следующим образом:

- *tp* — количество раз, когда определенная конструкция представляет собой синтагму, в которой единицы сочетаются друг с другом семантически;
- *tn* — количество раз, когда предсказанная конструкция не употребляется в языке, ее элементы не сочетаются друг с другом семантически;
- *fn* — количество раз, когда модель не предсказала конструкцию и не смогла сделать выбор между некорректным и корректным коллокатами: мера косинусной близости между обеими парами схожа с точностью до двух знаков после запятой.

Для нашего исследования наиболее релевантной является метрика точности, и мы отталкиваемся от нее при анализе результатов.

3.5.1. Анализ результатов разрешения псевдодизамбигуации

Как было упомянуто ранее, на каждом корпусе были обучены две неконтекстуализированные модели по 2400 параметров. Мы взяли 500 лучших результатов для каждой получившейся модели, отсортированных по точности предсказаний, и рассмотрели, какими были условия их обучения.

Результаты представлены в Приложении 2. В нем можно увидеть 8 таблиц: по одной для каждой комбинации определенного корпуса и алгоритма обучения. Результаты отсортированы по точности (ее значение представлено в колонке **pr. (%)**). Затем следуют столбцы **metric**, **size**, **min_count** и **sorted_vocab**, в которых указаны соответствующие параметры: оставшиеся параметры, значение которых варьируется более сильно, указаны в колонке **other**. В **n_comb** приведено общее количество комбинаций условий обучения для определенной зоны, выделенной на основании устойчивости параметров;

При анализе результатов нами были сделаны следующие наблюдения.

- ***О параметрах, оказывающих наибольшее влияние на результаты предсказаний***

На основании данных из таблиц в Приложении 2 можно сделать вывод о том, что точность предсказаний моделей «прикреплена» к некоторым устойчивым зонам комбинаций параметров. Эти комбинации составляют:

1. мера измерения сходства между векторами;
2. размерность векторов;
3. минимальная частота слова для принятия его во внимание моделью;
4. ранжирование словаря корпуса;

В рамках этих зон определяющими и наиболее устойчивыми являются мера измерения сходства и размерность векторов; значения параметра минимальной частоты слова и параметра, отвечающего за ранжирование словаря, могут показывать большую вариативность. Например, для модели **Word2Vec**, обученной на научно-популярном корпусе ***Nplus1*** с параметрами: мера измерения — корреляция, размер вектора — 150, с сортировкой словаря, — как минимальный порог частоты, равный 10, так и порог, равный 15, приведут к одинаковым значениям точности. Аналогичные примеры есть и со значениями параметра ранжирования словаря, тем не менее, в большинстве случаев их значения устойчивы в рамках упомянутых нами зон.

Интересным является тот факт, что параметры ширины контекстного окна и способа подсчета вектора могли принимать любые значения в рамках выделенных зон, но никак не влияли на итоговый результат. Это отображено нами в таблицах в колонке **other**, где параметр **window** может быть равен любому значению от 1 до 10, а параметр **cbow_mean** — любому из пары 0 и 1. Описанное явление универсально и для моделей на основе **Word2Vec**, и для моделей на основе **FastText**.

Что касается значений элементов комбинаций, которые лежат в основе зон, — хотя можно сказать, что предпочтительнее было бы использовать определенный диапазон значений при обучении (например, как мы покажем дальше, предпочтительнее использовать косинусную меру, а не квадрат евклидова расстояния), нельзя сказать, что одно определенное значение или конкретная комбинация параметров дадут лучшие результаты при предсказании конструкции в определенном корпусе. Одним из факторов, влияющих на это, является объем текстов. Ниже мы приводим пример с различными по объему срезами поэтического корпуса и оптимальными условиями обучения для них:

Таблица 1. Зависимость предсказаний и оптимальных параметров от объема корпуса

Объем корпуса	**Точность**	**Оптимальные параметры**
419,872 предложений	87.1%	**metric**={'correlation'}; **size**={100}; **min_count**={15}; **sorted_vocab**={1}; **window**={1…10}; **cbow_mean**={0,1}
200,000 предложений	90.97%	**metric**={'correlation'}; **size**={200}; **min_count**={15}; **sorted_vocab**={0}; **window**={1…10}; cbow_mean={0,1}
70,000 предложений	96.77%	**metric**={'correlation'}; **size**={300}; **min_count**={15};

		sorted_vocab={1}; **window**={1…10}; **cbow_mean**={0,1}

Еще один вопрос, который мы хотели рассмотреть — какая размерность векторов лучше всего фиксирует синтагматические отношения. По данным, представленным в Приложении 2, было выявлено, что наиболее частотным значением является 200 (встретилось 33 раз); за ним следуют 150 (31) и 100 (31) и — с небольшим отрывом — 300 (25) и 250 (21). Судя по статистике, большой размер вектора менее предпочтителен, тем не менее, разница не настолько выразительная, чтобы можно было судить о том, что более крупный вектор хуже справляется с отображением синтагматики.

- ***О выборе меры измерения сходства векторов***

По данным, основанным на результатах обучения, мы можем судить, что семантические отношения между элементами конструкций лучше всего определяются путем подсчета коэффициента корреляции векторов или вычисления косинусной меры. Так, в список 500 лучших результатов каждой модели не попали те, в которой мерой измерения сходства послужили евликдово расстояние и квадрат евклидова расстояния. Это подтверждает гипотезы других исследователей о том, что данные метрики намного хуже отображают семантические свойства по сравнению, например, с косинусной мерой, также о том, что коэффицент корреляции векторов и косинусная мера дают схожие результаты [Rohde и др. 2005].

Рассмотрим пример модели **Word2Vec** на основе научно-популярного корпуса, где был выбран квадрат евклидова расстояния. Как правило, у подобных моделей низкий показатель не только точности, но и полноты. Ниже указаны значения точности, полноты и F-меры при параметрах обучения: размер вектора — 300; ширина контекста — 3; минимальная частота слова — 5; способ подсчета вектора — усреднение; предварительная сортировка словаря отсутствует:

- *F-score: 57.62%*

- *Precision: 56.13%*
- *Recall: 59.18%*

Что касается выбора между косинусной мерой и вычислением корреляции, стоит отметить, что данные метрики равноценны, и в некоторых случаях вычисление коэффициента корреляции между векторами помогает достичь более высокой точности. На основании данных в Приложении 2 мы приводим сводку по тому, какие метрики лежат в основе трех лучших результатов моделей — отдельно для моделей **Word2Vec** и отдельно для моделей **FastText**. С колонках **Cosine similarity** и **Correlation** указано количество комбинаций параметров, в которых задействована данная метрика и при этом они привели к той же точности.

Таблица 2. Меры сходства векторов в трех лучших результатах моделей

Модель	Место по точности	Cosine similarity (кол-во моделей)	Correlation (кол-во моделей)
Word2Vec	I	60	60
	II	100	60
	III	120	100
FastText	I	80	0
	II	140	0
	III	160	20

Стоит отметить, что для моделей, обученных с **Word2Vec**, частоты использования коэффициента корреляции сопоставимы с косинусной мерой; тем не менее, он почти не фигурирует в предсказаниях моделей **FastText**. Поскольку, как будет показано нами позже, **FastText** скорее ориентирован на предсказание парадигматики, чем синтагматики, этот показатель нельзя считать надежным — в дальнейших экспериментах мы будем использовать как коэффициент корреляции, так и косинусную меру.

- ***О влиянии параметра max_vocab_size на обучение моделей:***

Как уже было упомянуто ранее, параметр **max_vocab_size** убирает из словаря нечастотные токены с целью уменьшения загрузки оперативной памяти при его построении. В ряде экспериментов нами было выявлено, что данный параметр способен оказать критическое влияние на качество обучения моделей **Word2Vec**, в частности — при работе с большими массивами текстов. Ниже мы приводим пример с полным новостным корпусом ***Fontanka***: наиболее оптимальный набор параметров дал точность в 62,58% (**metric**='cosine', **size**=250, **min_count**=15, **sorted_vocab**=0). Тем не менее, с указанием параметра **max_vocab_size** мы можем намного более высокого показателя:

Таблица 3. Качество предсказаний на примере модели **Word2Vec** (***Fontanka***)с разными значениями **max_vocab_size**

Параметр max_vocab_size	**Pr. (%)**
max_vocab_size=None	62,58%
max_vocab=30000	85.16%
max_vocab=60000	76.13%

Как можно увидеть, при указании значения 30000 можно увеличить точность предсказания конструкций почти на 23%. Рассмотрим пример с поэтическим корпусом, точность предсказаний для которого изначально высокая (87.1%) при условиях **metric**='correlation', **size**=100, **min_count**=15, **sorted_vocab**=1:

Таблица 4. Качество предсказаний на примере модели **Word2Vec** (***Stihi_ru***) с разными значениями **max_vocab_size**

Параметр max_vocab_size	**Pr. (%)**
max_vocab_size=None	87.1%
max_vocab_size=30000	92.9%
max_vocab_size=60000	84.52%

Ограничение словаря в 30,000 тоже оказывается наиболее удачным вариантом при обучении поэтического корпуса.

Стоит отметить, что ограничение словаря значительно улучшает результаты, когда в качестве меры измерения сходства между векторами выступают евклидово расстояние или квадрат евклидова расстояния. В примере ниже мы обучаем поэтический корпус с такими же параметрами, как было указано в абзаце выше, но меняем метрику на 'euclidean':

Таблица 5. Качество предсказаний на примере модели **Word2Vec** (***Stihi_ru***) с метрикой евклидова расстояния

Параметр max_vocab_size	**Pr. (%)**
max_vocab_size=None	72.9%
max_vocab_size=30000	81.94%
max_vocab_size=60000	70.97%

Наконец, приведем пример с моделью **FastText**, обученной на корпусе художественной литературы с оптимальными параметрами: **metric**='cosine', **size**=200; **min_count**=10; **sorted_vocab**=0:

Таблица 6. Качество предсказаний на примере модели **FastText** (***Lib.ru.sec***) с разными значениями **max_vocab_size**

Параметр max_vocab_size	**Pr. (%)**
max_vocab_size=None	57.42%
max_vocab_size=30000	50.97%
max_vocab_size=60000	53.55%

Как можно отметить, в данной модели **FastText max_vocab_size** не улучшает, а ухудшает качество предсказаний конструкций. Нами были проанализированы еще несколько примеров, и был сделан вывод, что подобное явление можно наблюдать во всех моделях, обученных с помощью **FastText**: точность при использовании **max_vocab_size** либо падает, либо

остается на таком же уровне. Отчасти это может быть обусловлены парадигматической природой предсказаний модели, о которой мы поговорим в следующем разделе.

- ***О влиянии разметки на результаты предсказаний***

Как было выявлено нами при анализе предсказанных конструкций, исходная разметка подкорпусов «Тайги» содержит погрешности, которые повлияли на качество предсказаний — в частности, при обучении модели **FastText**. Несмотря на то, что части речи были по большей части размечены корректно, в некоторых случаях была некорректно определена лемма. Ярче всего это можно увидеть на примере слова «*год*» в следующем пункте этого раздела: так, например, *«2016год_NOUN*» должно быть разделено на две разные леммы, а *«вод_NOUN»* представляет собой пример некорректного определения исходной формы. Хотя в большинстве случаев были приведены корректные леммы, появление шума уменьшило вероятность предсказания корректных значений лексических функций и именных словосочетаний. В связи с этим для предсказания лексических функций и именных конструкций мы в дальнейшем используем собственную разметку, осуществленную при помощи инструмента *pymorphy2*[16].

- ***О синтмагматической природе предсказаний моделей FastText.***

После анализа результатов, приведенных в Приложении 2, нами было сделано предположение, что низкая точность предсказания конструкций моделями на основе **FastText** связана с тем, что данная модель больше подходит для предсказания парадигматических отношений, чем для моделирования синтагматики.

Рассмотрим два слова: «*год*» и «*изучение*». Первое является одним из наиболее частотных токенов во всех корпусах, а второе взято случайным

---

[16]URL: https://pymorphy2.readthedocs.io/en/stable/

образом и содержит достаточно большое количество синонимов. Обучим модель научно-популярного корпуса с набором оптимальных параметров: **metric**='cosine', **size**=200, **window**=5, **min_count**=10, **cbow_mean**=1, **sorted_vocab**=1. Результаты выдачи для указанных слов будут следующими:

- «*год_NOUN*»: *«код_NOUN», «полгод_NOUN», «погод_NOUN», «нод_NOUN», «2016год_NOUN», «плод_NOUN», «род_NOUN», «2015год_NOUN», «ход_NOUN», «вод_NOUN»*;
- «*изучение_NOUN*»: *«учение_NOUN», «обучение_NOUN», «научение_NOUN», «излучение_NOUN», «вручение_NOUN», «получение_NOUN», «облучение_NOUN», «заключение_NOUN», «лечение_NOUN», «понижение_NOUN»*.

В обоих случаях результаты предсказаний нельзя назвать точными: для леммы «*год_NOUN*» модель приводит ряд несинонимичных существительных, появление части которых является результатом ошибок при лемматизации и токенизации. Для слова «*изучение_NOUN*» семантически связанными с ним в результатах выдачи являются только леммы *«учение_NOUN», «обучение_NOUN», «научение_NOUN»*, а другие похожи только таким же финалем, объединяющим словообразовательный суффикс и флексию (*«-ение»*). Стоит отметить, что в результатах выдачи нет ни одного прилагательного или слова другой части речи кроме существительного.

Поскольку в нашем исследовании акцент сделан на именные словосочетания, мы рассмотрели, насколько частотны леммы с окончанием «*_ADJ»* в 1000 лучших предсказаний для обоих слов: в результате оказалось, что предсказаниях нет существительных в целом.

Для сравнения мы обучили модель **Word2Vec** с таким же набором параметров: в результатах выдачи оказалось 192 и 199 прилагательных для слов «*год_NOUN*» и «*изучение_NOUN*», соответственно.

Несмотря на проблемы, связанные с неточной разметкой, о которых мы писали в прошлом пункте, данное явление действительно для модели **FastText** и на более «чистых» текстах. Таким образом, мы заключаем, что

модель **FastText** не годится для предсказания синтагматики в нашем исследовании, и поэтому не рассматриваем ее на следующем этапе исследования, когда анализируем лингвистические особенности и свойства конструкции, которые предсказывают модели.

3.6. Подготовка данных для предсказания конструкций и значений лексических функций

Используя данные и алгоритм, описанные нами в предыдущих разделах, в этом разделе мы проводим следующие операции:

- увеличение объема некоторых подкорпусов (посредством извлечения большего объема данных);
- предобработка корпуса инструментом *pymorphy2*;
- поиск оптимальных параметров для обучения модели **Word2Vec** для данных, как на предыдущем шаге;

Помимо шагов предобработки, которые были описаны нами ранее, — непосредственно разметки и ее сохранения, удаления стоп-слов, выгрузки токенизированных предложений в отдельный файл — мы добавляем следующие ограничения:

- длина токена должна быть больше 2;
- длина токенизированного предложения должна быть больше 2;
- классический набор тегов в *pymorphy2* преобразуется в универсальные теги: в частности, «ADJF» (полное прилагательное) и «ADJS» (краткое прилагательное) переходят в «ADJ».

В некоторых случаях подобная предобработка привела к незначительному уменьшению корпусов. Данные, которые мы использовали, для исследования, следующие:

- ***Fontanka*** (новостной): 41,234,011 токенов, 3,611,338 предложений;

- ***Nplus1*** (научно-популярный): 1,328,657 токенов, 90,313 предложения;
- ***Stihi_ru*** (поэтический): 5,961,406 токена, 703,358 предложений (первые 100,000 текстов);
- ***lib.ru.sec*** (художественный): 31,591,065 токенов, 3,791,616 предложения (первые 1000 текстов).

Нами был вычислен следующий оптимальный набор параметров для обучения каждой модели:

- ***Fontanka*** (новостной): **metric**='cosine', **size**=200, **min_count**=15, **sorted_vocab**=0, **window**=any (default: 5), **cbow_mean**=any (default: 1);
- ***Nplus1*** (научно-популярный): **metric**='cosine', **size**=150, **min_count**=15, **sorted_vocab**=0, **window**=any (default: 5), **cbow_mean**=any (default: 1);
- ***Stihi_ru*** (поэтический): **metric**='cosine', **size**=150, **min_count**=10, **sorted_vocab**=1; **window**=any (default: 5), **cbow_mean**=any (default: 1);
- ***lib.ru.sec*** (художественный): **metric**='cosine', **size**=100, **min_count**=15, **sorted_vocab**=0; **window**=any (default: 5), **cbow_mean**=any (default: 1);

3.7. Лингвистические особенности предсказанных конструкций

В данном разделе мы исследуем лингвистические свойства конструкций, которые предсказывают модели, и оцениваем, насколько предсказания соотносятся с тем, какие конструкции действительно могут иметь место. Также мы сравниваем их с конструкциями, представленными проектом **Портрет слова** НКРЯ[17], и результатами двух предсказывающих

[17]URL: https://ruscorpora.ru/word/main

моделей **Дистрибутивно-семантического калькулятора** (ДСМ) [Bukia и др. 2016]: моделью, обученной на выгрузке русскоязычной ***Википедии*** в 2017 году (обозначим ее как **DSM-Wiki**), а также моделью, обученной на корпусе ***Lib.ru*** в 2017 году (размер контекстного окна равен 5) (обозначим ее как **DSM-Lib**);

Для начала рассмотрим способность моделей предсказывать стандартные именные словосочетания. За основу мы берем алгоритм, реализованный в проекте ДСМ. Калькулятор, способный предсказывать конструкции, представлен в **Word2Vec-калькуляторе**[18]. Предполагается, что ему на вход поступает целевое слово и его часть речи, затем уточняется, какая модель будет использоваться, и количество результатов в выдаче.

В Приложении 3 представлены результаты выдачи для ряда слов русского языка, выбранных случайным образом: «*сайт*», «*человек*», «*научно-исследовательский*», «*красивый*», «*день*» и «*система*». Внизу таблицы приведены результаты выдачи моделей, упомянутых нами выше. Также в ней действует следующая разметка:

- зеленым цветом обозначены предсказания, совпавшие с моделью ***DSM-Wiki***;
- лиловым цветом обозначены конструкции, совпавшие с моделью ***DSM-Lib***;
- синим цветом обозначены конструкции, совпавшие с моделью ***НКРЯ***;
- также подчеркнуты конструкции, встретившиеся сразу в нескольких списках;

В целях проведения сравнительного анализа предсказаний обученных нами моделей разных типов мы провели два эксперимента.

---

[18]URL: https://dsm-calculator.ru/solutions/W2V/calc

В первом эксперименте была проведена оценка согласованности исследовательских моделей. Для этого проводилось попарное сравнение результатов ***Nplus1***, ***Fontanka***, ***Lib.ru.sec*** и ***Stihi_ru***.

Коэффициент согласованности предсказаний моделей *A* рассчитывался как отношение числа повторов к общему числу предсказанных коллокатов. В Таблице 7 приведены результаты анализа данных. Наименьшая согласованность зарегистрирована для коллокатов целевого слова «*научно-исследовательский*» (0,033), наибольшая согласованность у коллокатов целевого слова «*день*» (0,15).

Обращает на себя внимание несоответствие в значениях косинусов для большинства совпадающих коллокатов. При единичных совпадениях типа «*верующий* (*человек*)», значения cosine = 0,43, однако разница между значениями косинусной меры может быть двукратной: например, «(*электронный*) *сайт*», cosine = 0,31 и cosine = 0,61. Высокий разброс в значениях косинусной меры указывает на то, что между моделями существуют различия по силе связей внутри совпадающих коллокаций. Малое число совпадений между предсказаниями моделей объясняется как стилистическими, так и тематическими различиями между корпусами, на которых были обучены рассмотренные нами модели.

Таблица 7. Оценка согласованности моделей.

Целевое слово	**Повторяющиеся коллокаты**	**Согласованность предсказаний моделей *A***
сайт	*новостной, электронный, подробный*	0,075 (3 повтора на 40 коллокатов)
человек	*верующий, больной, чужой, нищий*	0,1 (4 повтора на 40 коллокатов)
красивая	*блондинка, прелесть*	0,05 (2 повтора на 40 коллокатов)

система	*дистанционная, автоматическая*	0,05 (2 повтора на 40 коллокатов)
день	*выходной, летний, июньский, десятый, сегодняшний, бессонный*	0,15 (6 повторов на 40 коллокатов)
научно-исследовательский	*машиностроение*	0,033 (1 повтор на 30 коллокатов)

Во втором эксперименте мы сравниваем результаты предсказаний исследовательских моделей с эталонными и выделяем общие конструкций. Коэффициент *A* рассчитывается исходя из суммарного числа предсказаний моделей ***Nplus1***, ***Fontanka*** и ***Lib.ru.sec*** для 6 целевых слов, равного 60 коллокатам, и модели ***Stihi_ru***, которая предсказала 50 коллокатов:

- **Nplus1**: совпадает с **НКРЯ** по 3 предсказаниям (*A* = 0,05); с **DSM-Wiki** — по 5 (*A* = 0,083); с **DSM-lib** — по 0 (*A* = 0).
- **Fontanka**: совпадает с **НКРЯ** по 3 предсказаниям (A = 0,05); с **DSM-Wiki** — по 7 (*A* = 0,116); с **DSM-lib** — по 7 (*A* = 0,116).
- **Lib.ru.sec**: совпадает с **НКРЯ** по 2 предсказаниям (*A* = 0,033); с **DSM-Wiki** — по 2 (*A* = 0,033); с **DSM-lib** — по 1 (*A* = 0,016).
- **Stihi_ru**: совпадает с **НКРЯ** по 5 предсказаниям (*A* = 0,1); с **DSM-Wiki** — по 1 (*A* = 0,02); с **DSM-lib** — по 0 (*A* = 0).

У всех обученных нами моделей низкое число пересечений с эталонными, при этом есть большое число корректных индивидуальных предсказаний. Это служит еще одним доказательством, подтверждающим, что предсказания моделей по типу и содержанию обусловлены тематикой и стилем корпусов. Результаты данного эксперимента указывают на бóльшую близость предсказаний моделей ***Fontanka*** и моделей ***ДСМ-калькулятора***, обученных на корпусе ***Википедии*** и на художественном корпусе, что можно объяснить разнообразием тем, которые освещаются в издании "Фонтанка.ru". Несколько неожиданным кажется малое число совпадений между

предсказаниями модели ***Lib.ru.sec*** и ***DSM-lib***, которые обучены на стереотипных данных (художественные тексты). Различия между моделями могут быть объяснены тем, что они были обучены на равнообъемных данных (свыше 9 млн токенов и свыше 146 млн токенов), тем самым, единообразие источников по важности уступает объему исходных текстов. Модель **Stihi_ru** ожидаемо ближе к результатам профилирования целевых слов по корпусу ***НКРЯ***, содержащему стихотворные тексты, нежели чем к моделям ДСМ-калькулятора.

Также можно отметить, что даже несмотря на корректность предсказания большей части конструкций с точки зрения лексической сочетаемости, косинусная мера обычно выше всего при вычислении предсказаний на научно-популярном корпусе; при этом данный корпус меньше всех остальных по объему. Данное явление может говорить о высокой плотности векторов модели ***Nplus1***. Потенциально это может быть связано с размером корпуса, тем не менее, для подтверждения данного предположения у нас еще недостаточно данных.

В предсказаниях есть как устойчивые сочетания («*официальный сайт*», «*выходной день*», «*социальная система*» и т.д.), так и сочетания, которые нельзя назвать устойчивыми и универсальными, но между элементами которых есть семантическое согласование, например, «*общественная система*»; при этом некоторые конструкции представляют собой термины или части терминов: «*социальная система*» , «*бортовая система*» и т.д.

Предсказанные конструкции по большей части композиционны и не фразеологизированные; интересным является тот факт, что в результатах выдачи встречается конструкция «*разумный человек*», при перестановке слов в которой мы можем получить конструкцию «*человек разумный*». Алгоритм, который мы используем, всегда ставит на первое место прилагательное. Тем не менее, приведенный случай может относится к тому, что В.В. Виноградов называл фразеологическими единствами (см. «*белены объелся*»), где

перестановка элементов меняет смысл конструкции и влияет на его идиоматичность [Виноградов 1977].

Научно-популярная и новостная модели хуже справляются с предсказанием конструкций для прилагательного «*красивый*», чем художественная и поэтическая; это может быть обусловлено тем, что данное прилагательное содержит оценочное суждение, которое реже встречается в первых двух типах текстов по сравнению с текстами художественной направленности.

Встречаются конструкции, в которых значения элементов не согласованы: например, «*красивое очарование*», «*красивая география*», и «*безлунный день*». Анализ аномальных коллокаций выходит за рамки данного исследования, более подробно этот вопрос рассмотрен в [Паничева 2019].

В предсказаниях для прилагательного «*научно-исследовательский*» есть ряд существительных, которые уже содержат этот компонент в сокращенном виде — НИ. Несмотря на то, что нами было введено ограничение на совпадение частей целевого слова и коллокатов (так как в выдаче могли встретиться конструкции такого типа, как «*человечный человек*»), данное явление достаточно тяжело обработать. Таким образом, при дальнейшей работе с конструкциями можно попробовать ввести некоторые правила для обработки именованных сущностей и их сокращений.

Также нами была проведена процедура предсказания значения лексических функций при заданных параметрах. Были рассмотрены несколько функций, которые описывают связь между прилагательным и существительным: *Bon()*, *Magn()*, *Antimagn()*. Ниже приведены их значения:

- *Bon()* — положительная оценка X-а;
- *Magn()* — интенсивность или большая степень проявления свойства X-а;
- *Antimagn()* — слабая интенсивность или степень проявления свойства X-а;

Калькулятор, с помощью которого можно предсказать коллокаты, соответствующие функциям, представлен в **W2VSemRel**[19].

В основе предсказания значений функций лежит процедура линейного преобразования, впервые описанная применительно к синтагматике в работе С. Родригез-Фернандез и ее коллег [Rodriguez-Fernandez 2016], а также примененная к данным русского языка в работе Е.В. Еникеевой и О.А. Митрофановой [Enikeeva, Mitrofanova 2017].

При анализе лексических функций у нас есть два векторных пространства — *A* — для аргументов, *C* — для коллокатов. Предположим, что *T* — данные, используемые для исследования, в которых представлен некоторый набор лексических функций. Строки матриц аргументов и коллокатов представляет векторные вложения конкретных слов. Тогда задачу можно определить как задачу нахождения такого преобразования, что для вектор некоторого аргумента *a* будет преобразован в вектор *b*, который отображает значение лексической функции при заданном параметре.

Рассмотрим несколько примеров. Мы используем модель, обученную на ***Lib.ru.sec***, как векторное пространство аргументов и модель, обученную на ***Stihi_ru***, — как пространство коллокатов:

Так, для аргумента «*человек*» при лексической функции *Bon()* мы можем получить следующие результаты:

'хороший', 0.6564709544181824,
'плохой', 0.5212271809577942,
'талантливый', 0.5040051341056824,
'замечательный', 0.4963124990463257,
'примерный', 0.4942880570888519 3,
'порядочный', 0.46803560853004456,
'незаменимый', 0.46504777669906616,
'уважаемый', 0.45950859785079956,

---

[19]URL: https://dsm-calculator.ru/solutions/W2VSemRel/calc

'дорогой', 0.45677119493484497,

'успешный', 0.449128657579422

Из всех вариантов некорректным оказывается только слово «*плохой*», другие же образуют семантические сочетаемые и композиционные конструкции с целевым словом.

Приведем также пример предсказания конструкций со значением *ANTIBON()*:

'слабый', 0.8093779683113098,

'плохой', 0.67351895570755,

'сильный', 0.6350550651550293,

'скверный', 0.6150011420249939,

'бессильный', 0.595156311988,,

'несправедливый', 0.5754484534263611,

'уязвимый', 0.575110137462616,

'ничтожный', 0.5734682679176331,

'ревнивый', 0.570009708404541,

'неразумный', 0.5606335401535034

Как и в случае с прилагательными, которые несут положительную характеристику, среди отрицательных прилагательных нам встретилось только одно некорректно выделенное слово — «*сильный*».

Стоит отметить также, что корпус и его лексические состав могут повлиять на результаты. Например, как мы упомянули ранее, для новостных текстов не характерно оценочное суждение. В связи с этим для для аргумента «*человек*» при лексической функции *Bon()* мы можем получить такие результаты, как:

'принадлежащий', 0.25943243503570557,

незарегистрированный', 0.2503795027732849

'причастный', 0.24600529670715332

'приразломный', 0.21205782890319824,

'преднамеренный', 0.19666612148284912

'справедливый', 0.1845957487821579

'беспартийный', 0.18193380534648895

Выводы к главе 3

В третьей главе мы переходим к экспериментальной части нашего исследования, заключающейся в обучении нескольких подкорпусов разных жанров и поиске лучших наборов параметров для них. На данном этапе нами были выбраны модели для обучения, а также рассмотрены значения таких параметров, как мера измерения степени сходства между векторами, ширина контекстного окна, минимальный порог частоты слова, сортировка корпуса по частотности, ограничение на размер словаря при обучении, а также способ подсчета целевого вектора на основании векторов контекста.

Для оценки результатов была проведена процедура псевдодизамбигуации, в которой модель должна выбрать, какое из двух слов-кандидатов является семантически близким для целевого. Также был подобран набор данных для псевдодизамбигуации на основании четырех корпусов. Результаты были проанализированы и подробно описаны.

Наконец, мы применили полученные результаты к двум лингвистическим задачам:

- Предсказание значений лексических функций *Bon()*, *Magn()* и *Antimagn()* при заданных аргументах;
- Предсказание именных конструкций с заданным целевым словом.

Итоговые конструкции были также рассмотрены нами и проанализированы.

Заключение

В результате исследования нами была достигнута поставленная цель: выявлены закономерности, влияющие на параметры обучения предсказывающих языковых моделей, а также оценено то, насколько они могут быть полезны в задачах, требующих выявления лексических конструкций в русскоязычных текстах.

В рамках теоретической части данного исследования нами были проанализированы различные классификации конструкций и их особенности. Большинством исследователей подчеркивается то, что синтаксис и семантика взаимосвязаны, и нельзя обойтись без обращения к какой-либо из этих двух областей при анализе сочетаний и явления сочетаемости как такового. Так, Грамматика конструкций стала основной теорией, которая отталкивается от понятия конструкции как базового компонента языка; при этом еще ее создатель Ч. Филлмор отмечает, что конструкции обладают и синтаксическим, и семантическим элементами значения [Fillmore 1985; Fillmore 1988]. В.В. Виноградовым описываются устойчивые сочетания с точки зрения композиционности и степени идиоматичности [Виноградов 1977]. Помимо исследования непосредственно семантики конструкций, исследователь показывает, для каких сочетаний их синтаксическая форма важна, и не может быть изменена, а также рассматривает ряд факторов, которые влияют на фиксацию выражений в речи, например, эмоционально-экспрессивный фактор. В.Г. Гак рассматривает законы сочетаемости и проводит зависимость между сочетаемостью и наличием или отсутствием определенных компонентов семантического значения в обоих элементах конструкции [Гак 1998]. Его идеи развивает Ю.Д. Апресян [Апресян 1995]. Л.Н. Иорданская и И.А. Мельчук в свою очередь рассматривают лексико-синтаксические конструкции с точки зрения регулярности их образования и произвольности компонентов в них, а также описывают лексические функции [Мельчук, Иорданская 2007].

Помимо этого, нами были проанализированы подходы к автоматическому выделению конструкций: с использованием правил, счетные, гибридные и предсказывающие. Так, например, функция лексико-семантического поиска реализована во многих корпусах (НКРЯ и т.д.). К счетным алгоритмам можно отнести определение конструкций при помощи мер ассоциации (**PMI**, **log-likelihood**, **критерий Хи-квадрат** и т.д.), а также языковые модели, использующие данные о совместной встречаемости слов: например, **LSA** [Landauer и др. 1998] и **HAL** [Lund, Burgess 1996]. Данные о семантике используют также предсказывающие языковые модели: **Word2Vec** [Mikolov и др. 2013a; Mikolov и др. 2013b], **FastText** [Joulin и др. 2018; Bojanowski и др. 2017], **BERT** [Devlin и др. 2018] и их разновидности. Различие между дистрибутивными счетных моделей и дистрибутивными предсказывающими моделей связано с интерпретируемостью векторов с точки зрения соответствия их координат контекстным элементам,, — у предсказывающих моделей с высокой обобщающей способностью эта способность выражена слабо.. Наконец, можно выделить лингвостатистические алгоритмы, такие как **RAKE** [Rose и др. 2009] и **YAKE** [Campos и др. 2020].

В экспериментальной части исследования мы выбрали архитектуры и данные для обучения языковых моделей и рассмотрели ряд параметров, таких как мера измерения сходства векторов, контекстное окно, ограничение на размер словаря, порог частоты слова, сортировка корпуса, а также способ подсчета целевого вектора. Были обучены модели **Word2Vec** и **FastText**, а материалом для исследования послужили корпусы «Тайга» [Shavrina, Shapovalova 2017] и lib.ru.sec [Panchenko и др. 2017]. Для проверки качества обучения применялась процедура псевдодизамбигуации, и был выявлен ряд следующих наблюдений, например:

- **FastText** лучше предсказывает парадигматические отношения в корпусе и намного хуже — синтагматические;

- среди метрик, оказывающих наибольшее влияние на предсказания, можно выделить меру измерения сходства векторов, их размерность, минимальную частоту слов для их учета моделью, а также способ сортировки словаря; помимо этого при обучении относительно больших корпусов (от 5 млн токенов) влияние может оказывать параметр, отвечающий за ограничение количество единиц в словаре модели;
- для вычисления степени сходства между векторами предпочтительнее использовать косинусную меру или вычислять коэффициент их корреляции;
- незначительные погрешности при разметке корпуса могут сильно снизить качество обучения, поэтому важно проверять разметку корпуса.

Наконец, мы применили результаты работы для предсказания именных словосочетаний и значений лексических функций. Для оценки качества предсказаний именных словосочетаний мы провели два эксперимента: оценка согласованности предсказаний моделей и сравнение моделей с эталонными данными. Результаты экспериментов отражают стилистические различия между корпусами, на которых были обучены рассматриваемые модели.

Эмпирические данные, полученные в рамках представленного проекта, будут применены в совершенствовании ДСМ-калькулятора и в разработке дистрибутивно-семантических моделей для лингводидактических целей.

Список литературы

1. Апресян Ю.Д. Избранные труды. В 2 т. Том I. Лексическая семантика: синонимические средства языка // Ю. Д. Апресян. – М.: Языки славянских культур, 1995. – 480 с.
2. Балли Ш. Общая лингвистика и вопросы французского языка // Ш. Балли. – М.: Издательство иностранной литературы, 1955. – 416 с.
3. Букия Г.Т., Протопопова Е.В., Митрофанова О.А. Корпусная оценка степени близости единиц в лексических конструкциях. // Структурная и прикладная лингвистика. 2015. № 11. – С. 252−270. – [Электронный реусурс] URL: https://elibrary.ru/item.asp?id=25849112&ysclid=l9v2yxqhkw846091732 (дата обращения: 31.05.2023).
4. Виноградов В.В. Избранные труды. Лексикология и лексикография. // М.: Наука, 1977. – 322 с.
5. Гак В.Г. К проблеме семантической синтагматики // Языковые преобразования. – М.: Языки русской культуры, 1998. – С. 272–297.
6. Жолковский А.К. О принципиальном использовании смысла при машинном переводе / А.К. Жолковский, Н.Н. Леонтьева, Ю.С. Мартемьянов // Мартемьянов Ю.С. Логика ситуаций. Строение текста. Терминологичность слов. – М.: Языки славянской культуры, 2004. – С. 84–99.
7. Иорданская Л.Н., Мельчук И.А. Смысл и сочетаемость в словаре. // Л.Н. Иорданская, И.А. Мельчук – М.: Языки славянских культур, 2007 – 665 с.
8. Лингвистический энциклопедический словарь. // Институт языкознания АН СССР. [под редакцией В.Н. Ярцевой и др.] - М.: Советская энциклопедия, 1990.
9. Паничева П.В. Анализ параметров семантической связности с помощью дистрибутивных семантических моделей: на материале русского языка: диссертация кандидата филологических наук: // Паничева Полина Вадимовна; [Место защиты: Рос. гос. пед. ун-т им. А.И. Герцена]. – Санкт-Петербург, 2019.
10. Baroni M., Dinu G., Kruszewski G. Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Baltimore, Maryland: Association for Computational Linguistics, 2014. – C. 238–247. – [Электронный ресурс] URL: https://aclanthology.org/P14-1023/ (дата обращения: 28.05.2023).
11. Belyi A.V. , Mitrofanova O.A., Dubinina N.A. Distributive Semantic Models in Language Learning: Automatic Generation of Lexical-Grammatical Tests for Russian as

a Foreign Language. // Corpus Linguistics, 2023 Proceedings. – Санкт-Петербург, 2023. – In press.

12. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. // Transactions of the Association for Computational Linguistics, Volume 5. — Cambridge, MA: MIT Press, 2017. – С. 135–146. – [Электронный ресурс] URL: https://aclanthology.org/Q17-1010/ (дата обращения: 28.05.2023).
13. Buchholz S., Marsi E. CoNLL-X shared task on Multilingual Dependency Parsing. // Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). – New York City: Association for Computational Linguistics, 2006. – С. 149-164. – [Электронный ресурс] URL: https://aclanthology.org/W06-2920/ (дата обращения: 28.05.2023).
14. Bukia G., Protopopova E., Panicheva P., Mitrofanova O. Estimating Syntagmatic Association Strength Using Distributional Word Representations. // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference «Dialogue» (2016). – Issue 15 – Москва: РГГУ, 2016 – С. 112–122.
15. Campos R., Mangaravite V., Pasquali A., Jatowt A., Jorge A., Nunes C., Jatowt A., YAKE! Keyword Extraction from Single Documents using Multiple Local Features. // Information Sciences Journal, 2020. – С. 257-289.
16. Dagan I., Marcus S., Markovitch S. Contextual Word Similarity and Estimation from Sparse Data. // Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. – Columbus, Ohio, USA: Association for Computational Linguistics, 1993. – С. 164–171. – [Электронный ресурс] URL: https://aclanthology.org/P93-1022/ (дата обращения: 28.05.2023).
17. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) – Minneapolis, Minnesota: Association for Computational Linguistics, 2018. – С. 4171–4186. – [Электронный ресурс] URL: https://aclanthology.org/N19-1423/ (дата обращения: 28.05.2023).
18. Edmiston D. A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages, 2020. – [Электронный ресурс] URL: https://arxiv.org/abs/2004.03032 (дата обращения: 28.05.2023).
19. Enikeeva E.V., Mitrofanova O.A. Russian Collocation Extraction based on Word Embeddings. // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue». – Moscow, 2017. – С. 52–64.

20. Espinosa-Anke L., Codina-Filba J., Wanner L., Evaluating language models for the retrieval and categorization of lexical collocations. // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. – Association for Computational Linguistics, 2021. – С. 1406–1417.
21. Fillmore C.J. Syntactic Intrusion and the Notion of Grammatical Construction /// Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society. – Berkeley, 1985. – С. 73–86.
22. Fillmore C.J., Kay P., O'Connor M. C. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. // Linguistic Society of America (Vol. 64, No. 3), 1988.
23. Firth J.R. A Synopsis of Linguistic Theory, 1930-55. // Special Volume of the Philological Society. – Oxford: Blackwell, 1957. – С. 1–32.
24. Gage Ph. A New Algorithm for Data Compression. // The C Users Journal, 1994. – С.23–38.
25. Gale W.A., Church K.W., Yarowsky D. Work on Statistical Methods for Word Sense Disambiguation. // AAAI Fall Symposium on Probabilistic Approaches to Natural Language: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics. – Berkeley, 1992. – [Электронный ресурс] URL: https://studylib.net/doc/13790396/work-on--statistical-methods-for--word-sense--disambiguation (дата обращения: 28.05.2023).
26. Goldberg A.E. Constructions: A Construction Grammar Approach to Argument Structure. // University of Chicago Press, 1995. – 271 с.
27. Goldberg, A.E. Construction Grammar. // Concise Encyclopedia of Syntactic Theories, – Oxford: Pergamon, 1996.
28. Goldberg A.E. Constructions at Work: The Nature of Generalization in Language. // Oxford University Press, 2006. – 290 с.
29. Goldberg A.E. Constructionist Approaches // The Oxford Handbook of Construction Grammar. – Oxford University Press, 2013. – С. 15–32.
30. Hartung M., Kaupmann F., Jebbara S., Cimiano Ph. Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. // 15th Meeting of the European Chapter of the Association for Computational Linguistics (EACL). – Valencia, Spain: 2017. – 11 с. – [Электронный ресурс] URL: https://aclanthology.org/E17-1006/ (дата обращения: 28.05.2023).
31. Joulin A., Bojanowski P., Mikolov T., Jegou H., Grave E. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. // In Proceedings of the 2018

Conference on Empirical Methods in Natural Language Processing. – Berlin, Belgium: Association for Computational Linguistics, 2018. – С. 2979–2984. – [Электронный ресурс] URL: https://aclanthology.org/D18-1330/ (дата обращения: 28.05.2023).

32. Kolesnikova O., Gelbukh A. A Study of Lexical Function Detection with Word2Vec and Supervised Machine Learning. // Journal of Intelligent and Fuzzy Systems, 2020. – С. 1-8.
33. Lan Zh., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. – 2019. – [Электронный ресурс] URL: https://arxiv.org/abs/1909.11942 (даьа обращения: 28.05.2023).
34. Landauer T.K., Foltz P.W., Laham D. An Introduction to Latent Semantic Analysis. // Discourse Processes (Vol. 25), 1998. – С. 259–284.
35. Lenci A. Distributional Semantics in Linguistic and Cognitive research. // The Italian Journal of Linguistics, 2008. – [Электронный ресурс] URL: https://www.italian-journal-linguistics.com/app/uploads/2021/05/1_Lenci.pdf (дата обращения: 28.05.2023).
36. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. – 2019. – [Электронный ресурс] URL: https://arxiv.org/abs/1907.11692 (дата обращения: 28.05.2023).
37. Lund K., Burgess C. Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence // Behavior Research Methods, Instruments, and Computers (Vol. 28), 1996. – С. 203–208.
38. Masterman M. The Thesaurus in Syntax and Semantics / M. Masterman // Mechanical Translation (Vol. 4), 1957. – С. 35-43.
39. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. – 2013a. – [Электронный ресурс] URL: https://arxiv.org/abs/1301.3781 (дата обращения: 28.05.2023).
40. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations. // Proceedings of NAACL-HLT 2013. – Atlanta, Georgia, 2013b. – С. 746–751. – [Электронный ресурс] URL: https://aclanthology.org/N13-1090/ (дата обращения: 28.05.2023).
41. Moskvina A.D., Yerofeyeva A.R., Mitrofanova O.A., Kharabet Ya.K. Automatic Selection of Keywords and Phrases from the Russian-language Corpus of Texts Using the RAKE Algorithm. // Proceedings of the International Conference "Corpus

Linguistics-2017" (St. Petersburg, June 27-30, 2017). – Publishing house of St Petersburg State University, 2017. – C. 268–275.

42. Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase Extraction from the Russian Corpus on Linguistics by means of KEA and RAKE Algorithm. // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9-12, 2018, Moscow, Russia): Conference Proceedings. – 2018. – [Электронный ресурс] URL: https://www.elibrary.ru/item.asp?id=41112843 (дата обращения: 28.05.2023).
43. Osgood C.E. The Measurement of Meaning / C. E. Osgood, G. Suci, P. Tannenbaum. // Urbana: University of Illinois Press, 1957. – 342 с.
44. Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N., Biemann Ch. Human and Machine Judgements for Russian Semantic Relatedness. // Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers. – Springer International Publishing, 2017.
45. Pereira F., Tishby N., Lee L. Distributional Clustering of English Words // Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. – Columbus, Ohio: Association for Computational Linguistics, 1993. – C. 183–190.
46. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark Ch., Lee K., Zettlemoyer L. Deep Contextualized Word Representations. // Proceedings of the 2018 Conference of the North American Chapter of the Association of the Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). – New Orleans, Louisiana, 2018. – C. 2227–2237. – [Электронный ресурс] URL: https://aclanthology.org/N18-1202/ (дата обращения: 28.05.2023).
47. Rodriguez-Fernandez S., Espinosa-Anke L., Carlini R. Wanner L., Semantics-Driven Recognitin of Collocations Using Word Embeddings. // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – Berlin, Germany: Association for Computational Linguistics, 2016. – C. 499–505. – [Электронный ресурс] URL: https://aclanthology.org/P16-2081 (дата обращения: 30.05.2023).
48. Rohde D.L.T., Gonnerman L.M., Plaut D.C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. – 2005. – [Электронный ресурс] URL: https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=07FDBBF67A990E3272E5F53FE4F2B195?doi=10.1.1.131.9401&rep=rep1&type=pdf (дата обращения: 28.05.2023).

49. Rose S.J, Cowley W.E., Crow V.L., Cramer N.O. Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. – 2009. – [Электронный ресурс] URL: https://www.researchgate.net/publication/254994054_Rapid_automatic_keyword_extraction_for_information_retrieval_and_analysis (дата обращения: 28.05.2023)
50. Shavrina T., Shapovalova O. To the Methodology of Corpus Construction for Machine Learning: «TAIGA» Syntax Tree Corpus and Parser. // CORPORA2017, International Conference, Saint-Petersburg, 2017. – [Электронный ресурс] URL: https://publications.hse.ru/pubs/share/direct/228708458.pdf (дата обращения: 28.05.2023).
51. Stefanowitsch A., Gries S.Th. Collostructions: Investigating the Interaction of Words and Constructions. // International Journal of Corpus Linguistics (Vol. 8), 2003. – С. 209–243. – [Электронный ресурс] URL: https://www.researchgate.net/publication/37929828_Collostructions_Investigating_the_interaction_of_words_and_constructions (дата обращения: 28.05.2023).
52. Tenney I., Xia P., Chen B., Wang A., Poliak A., McCoy R. T., Kim N., Van Durme B., Bowman S. R., Das D., Pavlick E. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. – 2019. – [Электронный ресурс] URL: https://arxiv.org/abs/1905.06316 (дата обращения: 28.05.2023).
53. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones Ll., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is All You Need. // In Proceedings of NIPS, 2017. – [Электронный ресурс] URL: https://arxiv.org/abs/1706.03762 (дата обращения: 28.05.2023).

Приложение 1. Список ассоциатов для процедуры псевдодизамбигуации

Корректный коллокат	Некорректный коллокат	Целевое слово
внешний_ADJ	80-й_ADJ	вид_NOUN
подводный_ADJ	коммунальный_ADJ	лодка_NOUN
последний_ADJ	заразительный_ADJ	время_NOUN
молодой_ADJ	взаимный_ADJ	человек_NOUN
другой_ADJ	мясной_ADJ	слово_NOUN
тихий_ADJ	кудрявый_ADJ	океан_NOUN
будущий_ADJ	туринский_ADJ	год_NOUN
левый_ADJ	привычный_ADJ	рука_NOUN
разный_ADJ	пернатый_ADJ	страна_NOUN
сильный_ADJ	четвероногий_ADJ	ветер_NOUN
главный_ADJ	фешенебельный_ADJ	герой_NOUN
взрослый_ADJ	платиновый_ADJ	человек_NOUN
сегодняшний_ADJ	психический_ADJ	день_NOUN
железный_ADJ	требовательный_ADJ	дорога_NOUN
первый_ADJ	любопытный_ADJ	шаг_NOUN
обычный_ADJ	спартанский_ADJ	человек_NOUN
другой_ADJ	угловой_ADJ	планета_NOUN
первый_ADJ	истребительный_ADJ	место_NOUN
прошлый_ADJ	сигнальный_ADJ	век_NOUN
первый_ADJ	неподатливый_ADJ	день_NOUN
мировой_ADJ	плотоядный_ADJ	война_NOUN
нервный_ADJ	повседневный_ADJ	система_NOUN
настоящий_ADJ	однородный_ADJ	момент_NOUN
последний_ADJ	невидимый_ADJ	год_NOUN

белый_ADJ	психологический_ADJ	свет_NOUN
целый_ADJ	принудительный_ADJ	ряд_NOUN
следующий_ADJ	минимальный_ADJ	день_NOUN
первый_ADJ	решительный_ADJ	год_NOUN
солнечный_ADJ	личный_ADJ	свет_NOUN
реальный_ADJ	телеграфный_ADJ	мир_NOUN
крайний_ADJ	дисковый_ADJ	мера_NOUN
черный_ADJ	ржавый_ADJ	дыра_NOUN
больший_ADJ	дискретный_ADJ	часть_NOUN
другой_ADJ	пересохший_ADJ	страна_NOUN
синий_ADJ	развлекательный_ADJ	цвет_NOUN
красный_ADJ	несуществующий_ADJ	цвет_NOUN
прошлый_ADJ	слабонервный_ADJ	год_NOUN
будущий_ADJ	неорганический_ADJ	год_NOUN
подводный_ADJ	святый_ADJ	лодка_NOUN
молодой_ADJ	горючий_ADJ	человек_NOUN
синий_ADJ	неприхотливый_ADJ	цвет_NOUN
железный_ADJ	губный_ADJ	дорога_NOUN
взрослый_ADJ	численный_ADJ	человек_NOUN
новый_ADJ	балтийский_ADJ	поколение_NOUN
правильный_ADJ	пекинский_ADJ	ответ_NOUN
реальный_ADJ	бамбуковый_ADJ	мир_NOUN
солнечный_ADJ	диагональный_ADJ	свет_NOUN
последний_ADJ	спектральный_ADJ	время_NOUN
полный_ADJ	шестнадцатый_ADJ	темнота_NOUN
настоящий_ADJ	ботанический_ADJ	момент_NOUN
морской_ADJ	стиральный_ADJ	вода_NOUN
другой_ADJ	неудивительный_ADJ	слово_NOUN
яркий_ADJ	мочевой_ADJ	свет_NOUN

крайний_ADJ	уединенный_ADJ	мера_NOUN
левый_ADJ	китовый_ADJ	рука_NOUN
конечный_ADJ	праведный_ADJ	итог_NOUN
большой_ADJ	сапожный_ADJ	мир_NOUN
главный_ADJ	экскурсионный_ADJ	роль_NOUN
солнечный_ADJ	посылочный_ADJ	луч_NOUN
равный_ADJ	мореходный_ADJ	нуль_NOUN
живой_ADJ	грибной_ADJ	существо_NOUN
ночной_ADJ	верблюжий_ADJ	небо_NOUN
первый_ADJ	редакторский_ADJ	лицо_NOUN
мировой_ADJ	аскетический_ADJ	война_NOUN
солнечный_ADJ	рукотворный_ADJ	система_NOUN
черный_ADJ	проходной_ADJ	дыра_NOUN
естественный_ADJ	шутливый_ADJ	отбор_NOUN
воздушный_ADJ	сиамский_ADJ	шар_NOUN
красный_ADJ	двухдневный_ADJ	цвет_NOUN
главный_ADJ	полномасштабный_ADJ	герой_NOUN
следующий_ADJ	первозданный_ADJ	день_NOUN
живой_ADJ	велосипедный_ADJ	природа_NOUN
другой_ADJ	лесной_ADJ	страна_NOUN
спусковой_ADJ	лучевой_ADJ	крючок_NOUN
прошлый_ADJ	нательный_ADJ	год_NOUN
яркий_ADJ	волшебный_ADJ	звезда_NOUN
тихий_ADJ	шутливый_ADJ	океан_NOUN
другой_ADJ	доверительный_ADJ	сторона_NOUN
первый_ADJ	понятый_ADJ	год_NOUN
новый_ADJ	обширный_ADJ	место_NOUN
первый_ADJ	шаровидный_ADJ	место_NOUN
большой_ADJ	лошадиный_ADJ	расстояние_NOUN

целый_ADJ	мюнхенский_ADJ	ряд_NOUN
разный_ADJ	идиотский_ADJ	страна_NOUN
сильный_ADJ	хвостовой_ADJ	ветер_NOUN
внешний_ADJ	иронический_ADJ	вид_NOUN
средиземный_ADJ	неоткрытый_ADJ	море_NOUN
век_NOUN	классик_NOUN	прошлый_ADJ
сторона_NOUN	отключение_NOUN	разный_ADJ
ветер_NOUN	отвал_NOUN	сильный_ADJ
рука_NOUN	сокол_NOUN	левый_ADJ
расстояние_NOUN	самолечение_NOUN	большой_ADJ
человек_NOUN	лужица_NOUN	взрослый_ADJ
сторона_NOUN	насмешка_NOUN	обратный_ADJ
цвет_NOUN	бомба_NOUN	синий_ADJ
год_NOUN	сбруя_NOUN	прошлый_ADJ
дорога_NOUN	неделя_NOUN	железный_ADJ
планета_NOUN	баланс_NOUN	другой_ADJ
лицо_NOUN	омоложение_NOUN	первый_ADJ
слово_NOUN	подкрепление_NOUN	другой_ADJ
время_NOUN	эталон_NOUN	последний_ADJ
система_NOUN	снеговик_NOUN	последний_ADJ
герой_NOUN	сострадание_NOUN	главный_ADJ
вид_NOUN	снасть_NOUN	внешний_ADJ
день_NOUN	трапеция_NOUN	первый_ADJ
свет_NOUN	переводчик_NOUN	солнечный_ADJ
момент_NOUN	написание_NOUN	настоящий_ADJ
мир_NOUN	токсикоз_NOUN	реальный_ADJ
мера_NOUN	идеология_NOUN	крайний_ADJ
цвет_NOUN	асфальт_NOUN	красный_ADJ
год_NOUN	скат_NOUN	будущий_ADJ

свет_NOUN	перс_NOUN	белый_ADJ
океан_NOUN	дуб_NOUN	тихий_ADJ
война_NOUN	помесь_NOUN	мировой_ADJ
время_NOUN	координата_NOUN	долгий_ADJ
океан_NOUN	драгоценность_NOUN	тихий_ADJ
цвет_NOUN	большинство_NOUN	синий_ADJ
лодка_NOUN	удвоение_NOUN	подводный_ADJ
цвет_NOUN	гримаса_NOUN	красный_ADJ
система_NOUN	ящик_NOUN	солнечный_ADJ
луч_NOUN	оппонент_NOUN	солнечный_ADJ
свет_NOUN	неандерталец_NOUN	солнечный_ADJ
год_NOUN	верховье_NOUN	будущий_ADJ
шар_NOUN	слушатель_NOUN	воздушный_ADJ
вид_NOUN	ненападение_NOUN	внешний_ADJ
рука_NOUN	агрессивность_NOUN	левый_ADJ
вода_NOUN	набросок_NOUN	морской_ADJ
герой_NOUN	вылет_NOUN	главный_ADJ
ветер_NOUN	биржа_NOUN	сильный_ADJ
человек_NOUN	энтропия_NOUN	разный_ADJ
свет_NOUN	убеждение_NOUN	белый_ADJ
планета_NOUN	осязание_NOUN	другой_ADJ
рука_NOUN	ассоциация_NOUN	правый_ADJ
день_NOUN	волк_NOUN	следующий_ADJ
время_NOUN	сгиб_NOUN	долгий_ADJ
день_NOUN	ржавчина_NOUN	сегодняшний_ADJ
небо_NOUN	серебро_NOUN	ночной_ADJ
отбор_NOUN	стаканчик_NOUN	естественный_ADJ
дорога_NOUN	дешевизна_NOUN	железный_ADJ
ряд_NOUN	компенсация_NOUN	целый_ADJ

человек_NOUN	раздражитель_NOUN	молодой_ADJ
крючок_NOUN	стопа_NOUN	спусковой_ADJ
сторона_NOUN	ритм_NOUN	обратный_ADJ
море_NOUN	заповедник_NOUN	средиземный_ADJ
дыра_NOUN	посвящение_NOUN	черный_ADJ
день_NOUN	обсерватория_NOUN	первый_ADJ
война_NOUN	курсант_NOUN	мировой_ADJ
звезда_NOUN	корица_NOUN	яркий_ADJ
страна_NOUN	смысл_NOUN	другой_ADJ
человек_NOUN	лисичка_NOUN	обычный_ADJ
сторона_NOUN	заря_NOUN	разный_ADJ
человек_NOUN	колодец_NOUN	другой_ADJ
темнота_NOUN	двойка_NOUN	полный_ADJ
система_NOUN	смола_NOUN	нервный_ADJ
мир_NOUN	картография_NOUN	реальный_ADJ

Приложение 2. Параметры обучения при наиболее высоких показателях точности

Таблица 8. Параметры обучения модели **Word2Vec** при наиболее высоких показателях точности, научно-популярный корпус ***Nplus1***.

pr. (%)	metric	size	min_count	sorted_vocab	n_comb	other
89.03%	cosine	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	150	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
88.39%	correlation	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
87.74%	correlation	150	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15			
	cosine	200	10	0	40	**window**={1…10}; **cbow_mean**={0,1}
			15			
	cosine	300	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15	0		
	correlation	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
87.1%	cosine	200	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15			
	correlatin	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
86.45%	cosine	150	15	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		
	correlation	200	15	0	40	**window**={1…10}; **cbow_mean**={0,1}

				1		
	correlation	250	10	0	20	**window**={1…10}; cbow_mean={0,1}
	cosine	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
85.81%	cosine	250	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
85.16%	cosine	100	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
84.52%	correlation	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 9. Параметры обучения модели **Word2Vec** при наиболее высоких. показателях точности, поэтический корпус ***Stihi_ru***

pr. (%)	**metric**	**size**	**min_count**	**sorted_vocab**	**n_comb**	**other**
87.1%	correlation	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
86.45%	correlation	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
85.81%	correlation	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
85.16%	correlation	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
84.52%	correlation	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
83.87%	cosine	200	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
83.23%	correlation	200	15	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		

	cosine	300	15	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		
	cosine	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
81.94%	correlation	200	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
81.29%	cosine	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
80.65%	cosine	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	100	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	200	10	0	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 10. Параметры обучения модели **Word2Vec** при наиболее высоких показателях точности, художественный корпус ***Lib.ru.sec***.

pr. (%)	**metric**	**size**	**min_count**	**sorted_vocab**	**n_comb**	**other**
70.32%	correlation	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
69.68%	cosine	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
67.74%	correlation	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}

	cosine	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
67.1%	correlation	200	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
66.45%	correlation	200	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	250	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
65.81%	cosine	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
65.16%	cosine	250	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
64.52%	cosine	300	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
63.87%	correlation	100	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
63.23%	cosine	100	10	0	40	**window**={1…10}; **cbow_mean**={0,1}
			15	1		
	correlation	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	10	1	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 11. Параметры обучения модели **Word2Vec** при наиболее высоких показателях точности, новостной корпус ***Fontanka***.

pr. (%)	**metric**	**size**	**min_count**	**sorted_vocab**	**n_comb**	**other**
62.58%	cosine	250	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
61.94%	cosine	300	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
60.65%	cosine	100	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
60.0%	correlation	100	5	0	40	**window**={1…10}; **cbow_mean**={0,1}
			10	1		
59.35%	correlation	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
58.71%	cosine	100	15	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		
57.42%	cosine	200	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
56.77%	correlation	250	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	150	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	200	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
56.13%	cosine	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	150	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
55.48%	correlation	250	15	1	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 12. Параметры обучения модели **FastText** при наиболее высоких показателях точности, научно-популярный корпус ***Nplus1***.

pr. (%)	metric	size	min_count	sorted_vocab	n_comb	other
57.42%	cosine	200	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
55.48%	cosine	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
54.84%	cosine	150	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
54.19%	cosine	100	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
53.55%	cosine	150	10	0	40	**window**={1…10}; **cbow_mean**={0,1}
			15	1		
52.9%	cosine	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
52.26%	correlation	300	5	1	60	**window**={1…10}; **cbow_mean**={0,1}
			10			
			15	0		
	cosine	200	5	0	40	**window**={1…10}; **cbow_mean**={0,1}
			10	1		
	cosine	100	10	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		
51.61%	cosine	150	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	10	0	20	**window**={1…10}; **cbow_mean**={0,1}

50.97%	cosine	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
50.32%	cosine	100	5	0	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 13. Параметры обучения модели **FastText** при наиболее высоких показателях точности, поэтический корпус ***Stihi_ru***.

pr. (%)	metric	size	min_count	sorted_vocab	n_comb	other
55.48%	cosine	200	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
54.84%	cosine	200	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
54.19%	correlation	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
53.55%	correlation	300	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
52.9%	cosine	200	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
52.26%	cosine	150	5	1	40	**window**={1…10}; **cbow_mean**={0,1}
			10	0		
	cosine	200	10	1	20	**window**={1…10}; **cbow_mean**={0,1}

	cosine	100	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
51.61%	correlation	300	5	1	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
50.97%	cosine	100	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15	0		
	correlation	300	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	150	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	250	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
50.32%	correlation	250	10	0	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 14. Параметры обучения модели **FastText** при наиболее высоких показателях точности, художественный корпус ***Lib.ru.sec***.

pr. (%)	**metric**	**size**	**min_count**	**sorted_vocab**	**n_comb**	**other**
57.42%	cosine	200	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
56.13%	cosine	150	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
55.48%	cosine	200	5	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15	0		
54.84%	cosine	150	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
54.19%	cosine	200	5	0	40	**window**={1…10}; **cbow_mean**={0,1}
			15	1		
	cosine	150	15	0	20	**window**={1…10}; **cbow_mean**={0,1}

53.55%	correlation	300	5	0	40	**window**={1…10}; **cbow_mean**={0,1}
				1		
52.9%	cosine	100	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15	0		
	cosine	150	10	1	40	**window**={1…10}; **cbow_mean**={0,1}
			5			
	correlation	300	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
52.26%	correlation	300	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	100	10	0	20	**window**={1…10}; **cbow_mean**={0,1}
50.97%	cosine	150	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
50.32%	correlation	300	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	250	5	1	20	**window**={1…10}; **cbow_mean**={0,1}

Таблица 15. Параметры обучения модели **FastText** при наиболее высоких показателях точности, новостной корпус ***Fontanka***.

pr. (%)	**metric**	**size**	**min_count**	**sorted_vocab**	**n_comb**	**other**
56.77%	cosine	200	15	0	20	**window**={1…10}; **cbow_mean**={0,1}
54.19%	cosine	100	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
	cosine	200	5	0	20	**window**={1…10}; **cbow_mean**={0,1}
53.55%	cosine	150	5	1	40	**window**={1…10}; **cbow_mean**={0,1}
			15	0		
52.9%	correlation	300	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
52.26%	cosine	100	10	0	40	**window**={1…10}; **cbow_mean**={0,1}
			15			

	cosine	200	15	1	20	**window**={1…10}; **cbow_mean**={0,1}
	correlation	300	5	1	60	**window**={1…10}; **cbow_mean**={0,1}
			10			
			15	0		
51.61%	cosine	150	10	1	20	**window**={1…10}; **cbow_mean**={0,1}
			15			
	correlation	300	5	0	20	**window**={1…10}; **cbow_mean**={0,1}

Приложение 3. Результаты предсказания именных словосочетаний

В таблицах подчеркиванием отмечены попарные соответствия между предсказаниями исследовательских моделей, зеленым цветом отмечены предсказания, совпавшие с моделью ***DSM-Wiki***; лиловым цветом — конструкции, совпавшие с моделью ***DSM-Lib***; синим цветом — конструкции, совпавшие с моделью ***НКРЯ***.

Таблица 16. Результаты предсказания именных словосочетаний со словом *«сайт»*

Слово: «сайт»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
официальный сайт: 0.779	текстовый сайт: 0.4	справочный сайт: 0.68	авторский сайт: 0.689
открытый сайт: 0.738	**новостной сайт**: 0.344	документальный сайт: 0.669	шуточный сайт: 0.689
краткий сайт: 0.726	хакерский сайт: 0.335	**газетный сайт**: 0.659	подробный сайт: 0.672
демонстрационный сайт: 0.705	общедоступный сайт: 0.332	биографический сайт: 0.65	заглавный сайт: 0.665
подробный сайт: 0.668	**информационный сайт**: 0.322	археологический сайт: 0.632	замечательный сайт: 0.655
рекламный сайт:	электронный сайт:	архивный сайт:	увлекательный сайт:

0.654	0.318	0.629	0.652
новостной сайт: 0.638	пиратский сайт: 0.313	рукописный сайт: 0.613	читаемый сайт: 0.645
панорамный сайт: 0.623	файлообменный сайт: 0.298	электронный сайт: 0.611	начинающий сайт: 0.638
картографический сайт: 0.608	**пользовательский сайт**: 0.288	юмористический сайт: 0.609	компьютерный сайт: 0.635
интерактивный сайт: 0.596	справочный сайт: 0.288	анонимный сайт: 0.599	записной сайт: 0.626
Коллокации НКРЯ: антигенный, интернетовский, официальный, новостной, русскоязычный, специализированный, корпоративный, рецепторный, иммунодоминантный, пранкерский			
Коллокации DSM-Wiki: новостный, новостной, пользовательский, рекламный, интерактивный, русскоязычный, промо, эксклюзивный, сетевой			
Коллокации DSM-Lib: рекламный, печатный, газетный, изданий, литературный, банковский, информационный, пишущий, парижский, коммерческий			

Таблица 17. Результаты предсказания именных словосочетаний со словом *«человек»*

Слово: «человек»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
склонный человек: 0.652	приезжий человек: 0.448	трудолюбивый человек: 0.464	разумный человек: 0.494
домашний человек: 0.624	верующий человек: 0.434	верующий человек: 0.431	бездушный человек: 0.475
социальный человек: 0.61	больной человек: 0.387	животный человек: 0.415	многий человек: 0.461
пожилой человек: 0.608	вич-инфицированный человек: 0.358	многоликий человек: 0.415	равный человек: 0.413
чужой человек: 0.607	прохожий человек: 0.357	поднебесный человек: 0.415	ничтожный человек: 0.4
здоровый человек: 0.594	нищий человек: 0.328	мыслящий человек: 0.412	равнодушный человек: 0.383
больной человек: 0.59	**бездомный человек**: 0.317	неполноценный человек: 0.408	различный человек: 0.378
эмоциональный	лайковый человек:	разумный человек:	чуждый человек:

человек: 0.586	0.314	0.404	0.366
незнакомый человек: 0.582	остальной человек: 0.31	нищий человек: 0.4	материальный человек: 0.363
психический человек: 0.577	мирный человек: 0.308	одарённый человек: 0.398	ближний человек: 0.361
Коллокаты НКРЯ: молодой, добрый, русский, умный, хороший, близкий, честный, простой, живой, советский			
Коллокаты DSM-Wiki: трудоспособный, общий, натуральный, человеческий, рогатый, бездомный, здоровый			
Коллокаты DSM-Lib: краснокожий, который, другой, один, тот, молодая, парный, взрослый, весь, такой			

Таблица 18. Результаты предсказания именных словосочетаний со словом *«красивый»*

Слово: «красивый»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
красивая радость: 0.944	**красивая блондинка**: 0.579	красивая брюнетка: 0.601	**красивая девушка**: 0.58
красивое произношение: 0.937	красивое очарование: 0.568	красивый щёголь: 0.589	красивая принцесса: 0.55
красивый герой: 0.924	**красивая причёска**: 0.559	**красивая блондинка**: 0.577	красивая леди: 0.52
красивая география: 0.92	красивый блеск: 0.545	красивая наружность: 0.57	красивая девчонка: 0.506
красивый орангутан: 0.92	красивый скворечник: 0.536	красивая аристократка: 0.562	**красивая внешность**: 0.505
красивый разговор: 0.92	красивый маникюр: 0.536	красивая дурнушка: 0.553	красивая прелесть: 0.501
красивый хэштег: 0.912	красивый цветочек: 0.531	красивая прелесть: 0.536	красивый романтик: 0.501
красивый питомец: 0.909	красивая натура: 0.53	красивая танцовщица: 0.526	**красивая женщина**: 0.498
красивое суждение: 0.908	красивое сердечко: 0.526	красивая шатенка: 0.52	красивая кокетка: 0.498
красивый поцелуй:	красивая экзотика:	красивая	красивая девочка:

0.907	0.525	француженка: 0.514	0.495
Коллокаты НКРЯ: женщина, девушка, лицо, мужчина, юноша, дама, парень, глаз, здание, почерк			
Коллокаты DSM-Wiki: красота, девушка, красавица, блондинка, внешность, зелень, улыбка, платье, зрелище, пейзаж			
Коллокаты DSM-Lib: красавица, молодой, блондинка, молодая, причёска, красота, маленькая, внешность, женщина, цвет			

Таблица 19. Результаты предсказания именных словосочетаний со словом *«система»*

Слово: «система»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
дистанционная система: 0.619	автоматизированная система: 0.515	**функциональная система**: 0.683	моральная система: 0.769
электродистанционная система: 0.535	**дистанционная система**: 0.505	генная система: 0.654	глобальная система: 0.724
радиолокационная система: 0.522	**универсальная система**: 0.453	технологическая система: 0.652	научная система: 0.723
спутниковая система: 0.512	**эффективная система**: 0.44	планетарная система: 0.646	социальная система: 0.677
автоматическая система: 0.494	динамическая система: 0.432	структурная система: 0.641	общественная система: 0.675
бортовая система: 0.486	квантовая система: 0.402	автономная система: 0.64	**политическая система**: 0.673
многофункциональная система: 0.484	**автоматическая система**: 0.402	динамическая система: 0.637	основная система: 0.659
высокоточная система: 0.476	глобальная система: 0.366	экологическая система: 0.632	личностная система: 0.656
сетевая система: 0.468	своевременная система: 0.365	интегральная система: 0.626	религиозная система: 0.655
радиоэлектронная система: 0.468	бесконтактная система: 0.364	дифференциальная система: 0.624	непреложная система: 0.652
Коллокаты НКРЯ: нервный, солнечный, информационный, сложный, единый, банковский, судебный, политический, философский, иммунный			
Коллокаты DSM-Wiki: автоматический, универсальный, дистанционный, модульный, статический, функциональный, стандартный, пассивный, эффективный, базовый			

Коллокаты DSM-Lib: энергетический, компьютерный, электронный, ракетный, силовой, гравитационный, автоматический, транспортный, химический, информационный

Таблица 20. Результаты предсказания именных словосочетаний со словом *«день»*

Слово: «день»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
ледниковый день: 0.64	**выходной день**: 0.404	всенощный день: 0.471	бессонный день: 0.5
летний день: 0.557	десятый день: 0.327	час-другой день: 0.47	безлунный день: 0.414
продолжительный день: 0.521	христов день: 0.307	сентябрьский день: 0.425	летний день: 0.387
сегодняшний день: 0.513	**пасхальный день**: 0.284	**выходной день**: 0.402	дождливый день: 0.383
очередной день: 0.513	**праздничный день**: 0.282	**новогодний день**: 0.395	**сегодняшний день**: 0.369
мировой день: 0.51	**календарный день**: 0.281	бессонный день: 0.382	погожий день: 0.35
зимний день: 0.506	нерабочий день: 0.28	июньский день: 0.374	пасмурный день: 0.346
десятый день: 0.505	выпускной день: 0.279	утренний день: 0.366	**следующий день**: 0.346
четвёртый день: 0.501	юбилейный день: 0.276	пятиминутный день: 0.365	июньский день: 0.345
влажный день: 0.493	майский день: 0.274	августовский день: 0.364	ненастный день: 0.342
Коллокаты НКРЯ: следующий, целый, сегодняшний, последний, завтрашний, рабочий, вчерашний, солнечный, прекрасный, выходной			
Коллокаты DSM-Wiki: прошлый, последний, последующий, следующий, сегодняшний, долгий, завтрашний, поздний, прошедшее, воскресный			
Коллокаты DSM-Lib: дневной, недельный, трехдневный, месячный, праздничный, новогодний, календарный, пасхальный, юлианский, високосный			

Таблица 21. Результаты предсказания именных словосочетаний со словом *«научно-исследовательский»*

Слово: «научно-исследовательский»			
Nplus1	***Fontanka***	***Lib.ru.sec***	***Stihi_ru***
научно-исследовательский судоремонт: 0.653	научно-исследовательская комиссия: 0.912	научно-исследовательское здравоохранение: 0.798	(*не представлено в словаре*)
научно-исследовательский цнии: 0.646	**научно-исследовательское бюро**: 0.882	научно-исследовательский нии: 0.775	
научно-исследовательское растениеводство: 0.64	**научно-исследовательское подразделение**: 0.879	научно-исследовательский филиал: 0.764	
научно-исследовательское нпо: 0.609	научно-исследовательский совет: 0.876	**научно-исследовательский институт:** 0.752	
научно-исследовательское приборостроение: 0.605	научно-исследовательский андрей: 0.871	научно-исследовательская секция: 0.743	
научно-исследовательская геология: 0.603	научно-исследовательский директор: 0.853	научно-исследовательское машиностроение: 0.739	
научно-исследовательское судостроение: 0.603	научно-исследовательский начальник: 0.846	научно-исследовательское мвд: 0.736	
научно-исследовательское машиностроение: 0.601	**научно-исследовательская академия**: 0.828	научно-исследовательский мгу: 0.734	
научно-исследовательская робототехника: 0.586	научно-исследовательский фонд: 0.826	научно-исследовательское проектирование: 0.727	
научно-	научно-	научно-	

исследовательский стройпроект: 0.586	исследовательское ведомство: 0.824	исследовательское внедрение: 0.719	
Коллокаты НКРЯ: институт, лаборатория, учреждение, работа, судно, ыентр, сектор, разработка, подразделение, полигон			
Коллокаты DSM-Wiki: <слова нет в модели>			
Коллокаты DSM-Lib: институт, промышленность, экономика, разработка, обслуживание, академия, медицина, факультет, ассоциация, бюро			