УДК 004.912 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2020.3/3046
Поступила в редакцию 20.09.2020
Подписана в печать 30.09.2020

АНАЛИЗ ПОДХОДОВ К АВТОМАТИЧЕСКОМУ ВЫДЕЛЕНИЮ КОНТЕКСТНЫХ СИНОНИМОВ ИЗ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

© 2020 Е. В. Полицына✉, С. А. Полицын, А. С. Поречный, Е. Е. Милованова

Московский авиационный институт (национальный исследовательский университет)
Волоколамское шоссе, 4, 125993 Москва, Российская Федерация

Аннотация. В статье рассматриваются проблемы определения контекстных синонимов, приводятся результаты анализа подходов к их автоматическому выделению из текстов на русском языке. Предлагается несколько подходов на основе применения лингвистических алгоритмов определения в тексте похожих семантико-синтаксических конструкций и их сочетания с методами машинного обучения. Анализируются полученные результаты применения алгоритмов на основе морфологического, синтаксического и семантического анализа текста, фильтрации полученных результатов путем использования ключевых слов и применения различных средств кластеризации. В заключении делаются выводы о применимости реализованных подходов и определяются направления развития сочетания этих подходови.

Ключевые слова: контекстные синонимы, автоматическое выделение синонимов, синонимия, семантико-синтаксический анализ текста.

ВВЕДЕНИЕ

Понимание любой информации, в том числе текста, является динамическим процессом взаимодействия между источником информации, самим знанием, социокультурным контекстом и другими влияющими на восприятие факторами. Этот процесс происходит в сознании человека и включает в себя декодирование лексико-грамматического материала и когнитивную обработку содержания с целью определения его смысла. Аналогичным образом большинство лингвистических программных инструментов основываются на декомпозиции текста с дальнейшим анализом и обработкой полученных данных. Для информационных систем, связанных с определением смысла текста, в процессе логической декомпозиции важно распознавать в качестве входных данных географические названия, личные имена, сокращения, устойчивые сочетания и прочие языковые конструкции.

Многозначность слов и словосочетаний, их зависимость от контекста и настроения, с которым они появляются, личный опыт и мировоззрение автора осложняют процессы анализа сообщений на естественном языке и формирования соответствующего ответа. Например, смысл фразы «Андрей не успел сесть на поезд, потому что застрял в лифте» может быть выражен большим количеством соответствующих синонимичных оборотов: «не успел» равнозначно «опоздал», «не сумел» и т. д.; «сесть» — «совершить посадку» и т. п.; «потому что» — «из-за того», «в связи с тем, что», «по той причине, что» и др. Полученное таким образом количество вариаций настолько большое, что существующие информационные системы (ИС) не способны обработать их все.

Главной задачей компьютерной лингвистики является разработка и создание программного обеспечения, способного поддерживать диалог между человеком и компьюте-

✉ Полицына Екатерина Валерьевна
e-mail: kathrin.beaver@mail.ru

Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.

ром. Из всех направлений развития данной отрасли наиболее сложным является моделирование процесса понимания смысла текстов, т.е. перехода от текста к формализованному представлению его смысла [1].

Одной из сложных проблем автоматизированного анализа текста является обработка многозначных слов. Полисемия нередко встречающееся в естественном языке явление, при котором языковая единица имеет более одного значения. Частным случаем данной проблемы является корректное выделение контекстных синонимов — слов, которые схожи по смыслу лишь в определенном контексте [2].

Контекстные синонимы определяются человеком как продукт индивидуального творческого акта, создающего близость значения слов в рамках определенного контекста. В связи с этим анализ контекстных синонимов имеет особое значение при рассмотрении художественного текста: синонимы возникают из необходимости фиксировать в слове новые оттенки явления, представления или понятия и так определённым образом характеризуют не столько явление, сколько его оценку, отношение к нему. Следствием из этого утверждения является то, что не всегда контекстные синонимы являются синонимами в языковой системе [3].

Контекстные синонимы могут возникать между:

- родовыми и видовыми понятиями — гипонимами и гиперонимами;
- словами одной тематической группы;
- близкими по значению словами, не являющимися синонимами в языковой системе — принимают на себя определенные признаки или вследствие акцента на одном или части признаков.

Изучение различных видов синонимии, примеров использования в речи важно как для лингвистики в целом, поскольку синонимы и система синонимических средств составляет богатство языка, обеспечивающее возможность носителя языка выразить одну и ту же мысль разными способами, расставить смысловые акценты или перефразировать свои мысли [3], так и для улучшения существующих и создания новых алгоритмов автоматического смыслового анализа текстов.

1. ПРОБЛЕМА МНОГОЗНАЧНОСТИ СЛОВ

В широком смысле синонимы — это слова, равнозначные или похожие по значению. Синонимы могут быть представлены как одной частью речи, так и разными, но близкими семантически [5]. Синонимы также могут включать одно или несколько слов: идти-шагать; гигантский-огромный; противогаз-резинотехническое изделие номер один. Также выделяются контекстные или контекстуальные синонимы — слова, которые равнозначны другим словам лишь в определенной ситуации или в конкретном контексте. Например: «душная, гнетущая атмосфера»; «пустынная, неприветливая равнина». В русском языке «душный» и «гнетущий», «пустынный» и «неприветливый» — разные понятия, но в определенном случае они близки по значению.

Контекст — это относительно законченная по смыслу часть текста или высказывания. Общий смысл контекста складывается из значений отдельных слов, в тоже время контекст дополняет и помогает прояснить значение каждого слова [6]. Нередко слова, выступающие в роли контекстных синонимов, таковыми не являются без учета контекста [6]. Например: «Все у них было как-то черство, неотесанно, неладно, негоже, нестройно, нехорошо…» (Н. В. Гоголь «Мёртвые души»).

Синонимическое сближение может возникать между родовыми и видовыми понятиями — гипонимами и гиперонимами, между словами одной тематической группы [7]. Например: «Я позвал собаку. Барбос подошел, он был мне рад. Овчарка села рядом.» В данном случае речь идет об одном и том же животном, поэтому слова «Барбос», «собака», «овчарка» синонимичны. Кроме того, в данном случае в качестве контекстных синонимов использованы местоимения и имена собственные.

Перифраз, метафора и другие литературные тропы могут порождать контекстные синонимы. Например, у М. Ю. Лермонтова в перечислении «звучал булат, картечь визжала»

булат — синоним холодного оружия, сабель и штыков.

Возникновение синонимии этого типа обусловлено семантическими процессами, возникающими при взаимодействии значений слов в тексте, т. е. синонимичным сближением. Синонимы возникают из необходимости фиксировать в слове новые оттенки явления, представления или понятия, такие слова могут характеризовать не собственно явление, а своеобразное видение, оценку отношения к нему. В тексте одни и те же объективные свойства явления характеризуются разными словами в зависимости от оценки повествующего. Часто эти слова не являются синонимами в языковой системе, но близки к ним по значениям, потому что в окружении синонимических слов принимают на себя определенные их признаки, и поэтому воспринимаются как синонимичные. В таком случае происходит сближение в тексте значений слов, которые в языке не являются синонимами [6].

По мнению Т. Б. Радбиля, назначение синонимических рядов заключено в разном представлении одного и того же действия путем выделения и акцентирования разных аспектов и элементов содержания ситуации. Слово в определённом значении «выбирает» из ситуации только часть информации, создавая определенный способ концептуализации, осмысления этой ситуации. Иными словами, слово представляет собой семантическую модель ситуации. В своей семантической модели ситуации говорящий может что-то выделить, сделать акцент, подчеркивая в слове какой-либо особый признак описываемого объекта, а в своей модели он может что-то «затемнять» в слове, отодвигая на задний план или даже искажая семантику ситуации. Поэтому для описания ситуации важна также оставшаяся информация, не вошедшая в ассертивную часть исходного, основного значения слова [3].

Таким образом, создание алгоритмов автоматического выделения контекстных синонимов может основываться на следующем:

1. Контекстные синонимы:

• семантически выражают одну и ту же мысль автора;

• вызывают одинаковые ассоциации у читателя;

• практически не используются в научных текстах;

• много используются в художественной литературе.

2. Контекстные синонимы применимы:

• к отдельным словам;

• к словосочетаниям или понятиям;

• к персонажам;

• отсылка на контекст других произведений или событий.

3. Контекстные синонимы могут быть схожи:

• по морфологическим признакам, если встречаются в рамках предложения;

• по окружению в синтаксической структуре предложения;

• по смыслу.

В ходе данного исследования для анализа результатов разрабатываемых подходов и алгоритмов была произведена выборка контекстных синонимов из литературных произведений и собрано порядка сотни рядов контекстных синонимов, анализ которых подтверждает предположение о том, что данное средство выражения используется для уточнения характеристики ключевых понятий и редко употребляется для описания объектов, не играющих важной роли в тексте [8].

2. АЛГОРИТМИЧЕСКИЙ ПОДХОД К ВЫДЕЛЕНИЮ КОНТЕКСТНЫХ СИНОНИМОВ

Для автоматического анализа текста очень важно корректно определять контекстные синонимы, чтобы правильно выделять из текста необходимые смысловые единицы, проводить семантическое сравнение текстов, получать рефераты и аннотации, более корректно выделять ключевые слова и словосочетания и т. д.

Был разработан алгоритм выделения контекстных синонимов из текста на русском языке, позволяющих на разных этапах анализа определить «кандидатов» на контекстные синонимы, а затем выбрать из них наиболее подходящие. Схема алгоритма представлена на рис. 1.

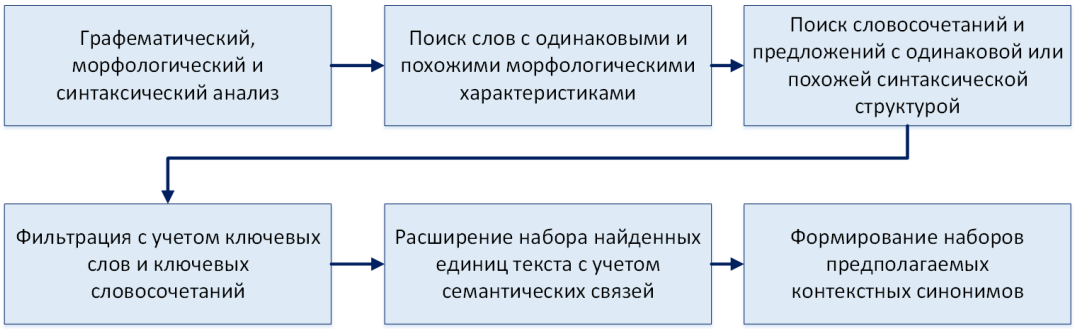

Рис. 1. Схема алгоритма выделения контекстных синонимов
[Fig. 1. Схема алгоритма выделения контекстных синонимов]

Алгоритм состоит из следующих шагов [9]:

1. Проведение морфологического и семантико-синтаксического анализа текста с применением фреймворка TAWT:

• разделение текста на предложения и слова, для слов необходимо определить их морфологические характеристики, там, где их можно определить однозначно;

• применение фильтра для снятия неоднозначностей определения морфологических характеристик [10];

• выделение словосочетаний из текста [10].

2. Формирование списка «шаблонов» словосочетаний с одинаковой структурой, например, сочетания прилагательных и существительных, совпадающие по формам.

3. Формирование списков словосочетаний с одинаковыми шаблонами.

4. Фильтрация с учетом ключевых слов и ключевых словосочетаний.

5. Расширение списка с учетом проверки по семантической сети [11], учитывающей отношения род-вид, антонимия, ассоциативной связи, словарных синонимов.

6. Формирование для каждого слова списков слов, которые в одинаковых шаблонах встречаются на том же месте с группировкой по частям речи.

Автоматически полученные списки слов со схожим морфологическим и синтаксическим «поведением» в конкретном тексте содержат достаточно много общеязыковых словоупотреблений и распространенных конструкций, не относящихся к тематике данного текста и не являющиеся «контекстными синонимами». Для их фильтрации необходимо проведение более глубокого анализа, как на шаге 5 описанного алгоритма, так и проведение дополнительного анализа путем определения семантической близости выделенных «похожих» фрагментов текста.

Для реализации описанного алгоритма сначала получаются морфологические характеристики для всех слов каждого предложения, например:

«Старику хотелось важных, серьезных мыслей…» (А. П. Чехов)

важный: (Часть речи — **18, прил.**) важных: морф. характеристики — **4272**

старик: (Часть речи — **17, сущ.**) старику: морф. характеристики — 230

хотелось: (Часть речи — **20, глаг. форма**) хотелось: морф. характеристики — 670764

серьёзный: (Часть речи — **18, прил.**) серьёзных: морф. характеристики — **4272**

мысль: (Часть речи — **17, сущ.**) мыслей: морф. характеристики — 187

Затем производится группировка единиц текста с одинаковыми синтаксическими структурами:

важный: (Часть речи — **18, прил.**) важных: морф. характеристики — **4272**

серьёзный: (Часть речи — **18, прил.**) серьёзных: морф. характеристики — **4272**.

Для дальнейшего поиска кандидатов в контекстные синонимы был выбран подход на основе поиска предложений со схожими синтаксическими структурами (рис. 2). Например:

«Мой новый друг снисходительно улыбнулся.»

*«Но как же я удивился, когда мой строгий **судья** вдруг просиял.» (А. де Сент-Экзюпери)*

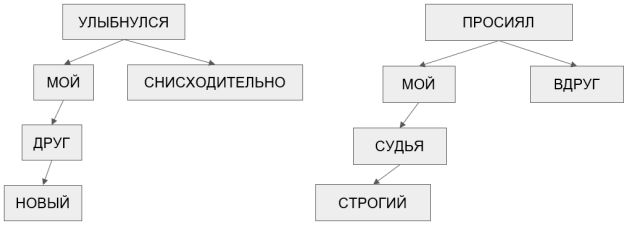

Рис. 2. Примеры схожих синтаксических деревьев [Fig. 2. Примеры схожих синтаксических деревьев]

После уточнения сформированных списков слов со схожими морфологическими характеристиками за счет исследования синтаксических конструкций, в которых они используются, осуществляется дополнительная фильтрация списка путем сравнения с выделенными ключевыми словами.

Дальнейшими шагами алгоритма автоматического выделения контекстных синонимов является определение семантической близости полученных списков слов. Это возможно благодаря тому, что в основе базы знаний открытой системы автоматизированного анализа текстов лежит двухуровневое семантическое представление (СП), которое отражает взаимосвязи окружающего мира или отдельных его компонент в виде модели текста или предметной области и позволяет обеспечить фиксацию множественности связей между разными способами выражения смысла понятий [11].

Нижний уровень СП основывается на оперировании словами и словосочетаниями в том виде, как они используются в текстах для выражения смысла. Верхний уровень СП, базируясь на словарных понятиях, позволяет отражать взаимосвязи окружающего мира, абстрагируясь от их конкретного выражения и опираясь на понятия как главные структурные единицы, т. е. верхний уровень предназначен для представления знаний. Нижний уровень СП основан на использовании существующего в системе типа структур данных — семантической сети, которая позволяет хранить информацию о понятиях и различных видах связей между ними. Такая структура была взята за основу для реализации верхнего уровня СП и расширена для хранения описаний понятий, источников этих понятий. Для начала реализации верхнего уровня СП необходимо было заложить в систему фундамент базовых знаний, в основу которого была положена информация, полученная из толковых словарей. Это позволяет создать список основных понятий в разных областях, добавить информацию о них и установить некоторые связи между ними.

Таким образом, следующими шагами алгоритма являются:

1. Построение на основе выделенных семантико-синтаксических представлений похожих фрагментов текста семантических сетей нижнего уровня и получение их отображений на СП верхнего уровня.

2. Сравнение полученных семантических представлений на основе использования алгоритмов определения изоморфности графов, вычисления длины маршрута или площади семантического поля с учетом ранее выделенных понятий и др.

Для уменьшения размерности задачи целесообразно применять сегментацию общей семантической сети по предметным областям средствами классификатора текстов на русском языке, предоставляющего программный веб-интерфейс для взаимодействия с ним [12].

Определение семантической близости фрагментов текстов, имеющих схожую семантико-синтаксическую структуру, позволит отфильтровать распространенные языковые конструкции, не являющиеся близкими по смыслу. Оставшиеся слова с похожим синтаксическим «поведением» в этом тексте, вероятнее всего, являются контекстными синонимами.

3. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫДЕЛЕНИЯ КОНТЕКСТНЫХ СИНОНИМОВ

Другим подходом к выделению контекстных синонимов является применение методов машинного обучения на основе векторизации текстов с учетом морфологических и синтаксических характеристик единиц текста.

Методы машинного обучения делятся на обучение с учителем и без учителя. Обучение с учителем предполагает наличие подготовленных экспертом данных, на которых происходит обучение системы. В то время как обучение без учителя предполагает, что система будет сама обучаться в процессе работы. Яркими примерами обучения с учителем и без учителя является классификация и кластеризация соответственно.

Классификация — это отнесение каждого элемента множества объектов $X = \{x_1, x_2, \ldots, x_n\}$ к определенному заранее заданному классу. Примером классификации может является определение стиля текста или определение эмоциональной окраски текста, для этого необходимо подготовить данные для каждого класса, после чего обучить систему.

Кластеризация или кластерный анализ — это группировка множества объектов $X = \{x_1, x_2, \ldots, x_n\}$ таким образом, чтобы элементы одной группы были более «похожи» друг на другу, а по отношению к элементам других групп как можно «более отличны» [16].

Возможность группировать множество объектов без предварительного определения классов, позволяет решать такие задачи, как:

1) Упрощение обработки данных и принятия решений за счёт того, что к каждому кластеру можно применить метод анализа, который подходит к конкретному кластеру, но не подходит к исходному множеству.

2) Сжатие данных за счет того, что один кластер схожих элементов можно заменить на одного представителя такого кластера.

3) Выявление аномалий или выявление новых объектов.

Кластеризация с помощью методов машинного обучения позволяет разделить множество объектов на несколько групп в соответствии с какими-либо признаками. Количество групп и признаков в общем случае произвольно и может быть выбрано в соответствии с решаемой задачей. В отличие от задачи классификации, в котором каждый из объектов относится к одному из заранее определённых классов, в задаче кластеризации происходит отнесение объекта к одному из заранее неопределённых классов. Данный процесс позволяет сгруппировать сходные данные, сформировав списки возможных контекстных синонимов.

Все методы машинного обучения используют представление сложного объекта в виде вектора признаков. Векторизация текста является отдельной трудоемкой задачей, в данном случае она была осуществлена как численное представление каждого слова текста в виде упорядоченного набора числовых морфологических характеристик (рис. 3).

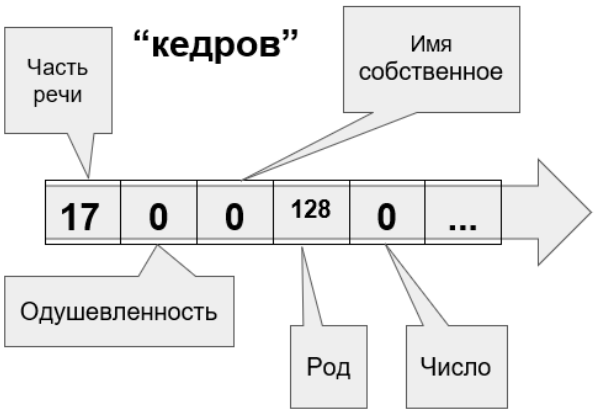

Рис. 3. Представление морфологических характеристик слова в виде вектора признаков

[Fig. 3. Представление морфологических характеристик слова в виде вектора признаков]

Кластеризация была выполнена с помощью библиотеки Weka 3.9.4 с применением метода кластеризации HierarchicalClusterer (Иерархический кластеризатор) с настройкой измерения расстояния EuclideanDistance.

На рис. 4 и 5 приведены результаты кластеризации для следующих текстов:

«Сколько раз пытался я ускорить
Время, что несло меня вперед.
Подхлестнуть, вспугнуть его, пришпорить,
Чтобы слышать, как оно идет.»
(С.Я. Маршак)

«Мой новый друг снисходительно улыбнулся. Но как же я удивился, когда мой строгий судья вдруг просиял.»
(А. де Сент-Экзюпери)

Другой подход состоит в применении нейронных сетей. Нейронные сети Кохонена — это класс нейронных сетей, основным элементом которых является слой Кохонена.

Рис. 4. Результаты кластеризации слов из примера из С. Я. Маршака
[Fig. 4. Результаты кластеризации слов из примера из С. Я. Маршака]

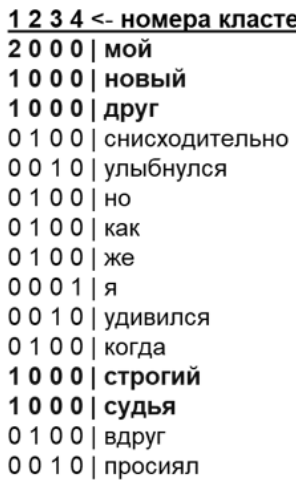

Рис. 5. Результаты кластеризации слов из примера из А. де Сент-Экзюпери
[Fig. 5. Результаты кластеризации слов из примера из А. де Сент-Экзюпери]

Он состоит из адаптивных линейных сумматоров («линейных формальных нейронов»), в которых выходные сигналы обрабатываются по правилу «Победитель получает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль. Сети Кохонена относятся к самоорганизующимся нейронным сетям, которые позволяют выделять кластеры из набора входных векторов, обладающих некоторыми общими свойствами. [10, 11].

Контекстные синонимы как выразительное средство применяются для описания понятий, предметов, людей с целью уточнения их характеристики в данный момент времени конкретным лицом (автором, литературным персонажем и проч.) или при конкретных обстоятельствах. Основываясь на данном утверждении, можно сделать вывод, что наибольшее количество контекстных синонимов принадлежит к некоторой главной мысли, основным предметам, персонажам, темам — к ключевым понятиям, о которых идёт речь в конкретном тексте. Поэтому для сохранения контекста необходимо использовать алгоритмы не обычного вычисления нейронов-победителей, а выделять ключевые понятия, на основе которых и будет выполняться дальнейшая организация нейронов.

Таким образом, алгоритм выделения контекстных синонимов с использованием нейронной сети включает в себя:

1. Предобработку текста, в результате которой исходный текст преобразуется к набору векторов на основе выделения ключевых понятий, проведения графематического, морфологического, синтаксического и семантико-синтаксического анализ для формирования зависимостей.

2. Кластеризацию по ключевым понятиям.

Процесс векторизации текста включает в себя следующие этапы:

1) Графематический анализ и формирование списка слов для дальнейшего исследования.

2) Морфологический анализ. На данном этапе происходит фильтрация полученного на предыдущем этапе списка по частям речи с целью выборки только имен существительных, выбранных в качестве объекта исследования.

Для каждого слова определяется:

- является ли слово именем нарицательным или аббревиатурой;
- число;
- род;
- одушевлённость;
- падеж.

3) Статистический анализ, выполняющий вычисление частоты употребления слова в тексте.

4) Семантико-синтаксический анализ, в рамках которого исследуется роль слова в конкретных предложениях и вычисляется:

• количестве раз, когда слово было главным в предложении;

• количестве зависимых слов;

• частоте употребления;

Для получения ключевых понятий из текстов использовался «Сервис автоматического реферирования текста» [13], поддерживающий поиск ключевых понятий.

Для уточнения результатов формируется список синтаксических зависимостей между словами, для этого использовано два метода:

• Выборка по дереву зависимостей. При помощи инструментов фреймворка TAWT [12] было произведено выделение слов, зависимых от ключевых понятий. Однако, в ходе исследования было выявлено, что на текстах малых размеров (от одной страницы до десяти) в результате выделяется не более двух зависимостей от ключевых понятий. Этих данных недостаточно для корректировки результатов поиска контекстных синонимов.

• Выборка по соседним словам. Формируется список слов, находящихся непосредственно по соседству с рассматриваемыми именами существительными. Полученные результаты для малых текстов являются более информативными, чем при анализе по дереву зависимостей, однако при выборке из текстов больших размеров результаты получаются слишком объемными, что негативно влияет на дальнейший анализ.

Анализ полученных результатов показал, что критерий количества учитываемых соседних слов на данном этапе невозможно сделать фиксированным. Качество выборок варьировалось в зависимости от входных данных, и на одном тексте полученный ряд слов был ближе к контекстным синонимам в случае выборки с одним соседом с обеих сторон, а в другом – в случае выборки с двумя соседями. На основе данного наблюдения была сделана гипотеза, что параметр количества соседей слова зависим от авторского стиля и размера текста. Примеры полученных результатов приведены в табл. 1.

Так, в произведении «Каштанка» было верно сопоставлено слово «хозяин» и понятие «незнакомец», так как речь идёт об одном и том же человеке, который подобрал Каштанку на улице и стал её новым хозяином. При анализе рассказа «Гимназистки. Сфинкс» были правильно определены контекстные синонимы к слову «гимназистки» — «Саша», «Нина» и «девочка», которые действительно являются гимназистками в рамках данного произведения, а «сфинкс» — прозвище одной из учениц. Анализ полученных результатов по рассказу «Живая шляпа» показал возможность правильного определения взаимодействия между объектами. Например, по ходу сюжета мальчишки кидают картошку в живую шляпу, не зная, что под ней сидит кот Васька, однако алгоритм смог выделить взаимосвязь между данными понятиями.

4. АНАЛИЗ РЕЗУЛЬТАТОВ ПРИМЕНЕНИЯ РАЗЛИЧНЫХ ПОДХОДОВ К ВЫДЕЛЕНИЮ КОНТЕКСТНЫХ СИНОНИМОВ

Анализ полученных результатов показал, что независимо от применения алгоритмического подхода или подхода, основанного на использовании методов машинного обучения, необходим учет ключевых понятий, которые используются в тексте, т. к. контекстные синонимы чаще используются применительно к ним.

Применение алгоритмического подхода позволяет сформировать списки слов и словосочетаний, которые схожим образом используются в тексте. Последующее применение семантико-синтаксического и семантического анализа позволяет выделить из них наиболее схожие по смыслу с учетом связанных слов, которые в некотором приближении определяют контекст.

Однако, применение алгоритмического подхода существенно усложняется необходимостью разработки большого количества синтаксических и семантических правил фильтрации получаемых списков слов и словосочетаний. Подход на основе кластеризации позволяет устранить эту проблему.

При применении подхода на основе кластеризации при количестве классов на 1–2 меньшем, чем количество частей речи в тексте, в один класс попадают слова с похо-

Таблица 1. Ряды выделенных контекстных синонимов
[Table 1. Ряды выделенных контекстных синонимов]

Произведение	Ключевые слова	Ряды контекстных синонимов (по одному соседу)	Ряд контекстных синонимов (по два соседа)
«Каштанка» А.П. Чехов	хозяин	–	–
	незнакомец	хозяин	хозяин, гусь
«Гимназистки. Сфинкс» Л.А. Чарская	англичанка	–	Саша, сфинкс, девочка, класс
	Саша	–	–
	гимназистка	сфинкс, пион, лицо, мать, сон, красная, дом	Саша, сфинкс, Нина, девочка, костюм, уж, дочь, класс, мать, лицо, слово, почтение, дом, отец
	счастливица	Саша, сфинкс, Нина, пион, лицо, сочинение, мать, сон, слово, красная, дом	сфинкс, Нина, девочка, лицо, сочинение, мать, сон, слово, дом, князь
«Живая шляпа» Н.Н. Носов	Вовка	Вадик	шляпа, пол, Вадик, пол, Васька
	комната	–	шляпа, пол, Вадик, Вовка
	ребята	–	Шляпа
	Васька	шляпа, Вадик, Вовка	шляпа, пол, Вадик, Вовка
	шляпа	–	Вовка
	картошка	–	Васька
	Вадик	шляпа, Вадик, Вовка	шляпа, Вовка
	клюшка	–	шляпа, Вадик, Вовка

жими наборами морфологических характеристик. При количестве классов большем, чем количество различных частей речи в тексте, слова с разными значениями морфологических характеристик распределяются в разные классы, а в одни — с близкими их значениями.

Результаты, полученные в результате применения подхода с использованием нейронной сети, содержат понятия, связанные с ключевыми словами в тексте, но среди них присутствуют не только контекстные синонимы, но и слова с другими семантическими связями. Например, в ряду контекстных синонимов, полученных к слову «счастливица» по тексту произведения «Гимназистки. Сфинкс» Л.А. Чарской, присутствуют слова «мать» или «сочинение».

Для решения этой проблемы необходимо расширение количества характеристик для векторов слов путем учета:

- весов слов в соответствии с эмоциональной и стилистической окраской;
- распределения частоты употребления с учётом слов-указателей (например, местоимений);
- ассоциативных связей между понятиями.

Однако при этом будет существенно возрастать объем обрабатываемой выборки данных. Таким образом, анализ подходов к автоматическому выделению контекстных синонимов показал, что:

- методы машинного обучения могут применяться для группировки различных единиц текста со схожими наборами морфологических, синтаксических и/или семанти-

ко-синтаксических характеристик и позволяют решить проблему создания большого количества правил группировки;

• для сокращения размерности выборки методы машинного обучения показывают лучшие результаты в сочетании с алгоритмами лингвистического анализа.

• результаты кластеризации дают возможность получения исходных данных для создания выборки для обучения классификатора.

ЗАКЛЮЧЕНИЕ

Разработанные алгоритмы выделения контекстных синонимов на основе различных подходов к определению схожести слов и понятий в тексте позволили сделать выводы о поведении контекстных синонимов в тексте и получить данные для дальнейших исследований в области выделения и изучения контекстных синонимов и многозначности слов в русском языке.

Совместное использование алгоритмического подхода на основе использования средств компьютерной лингвистики и методов машинного обучения для кластеризации близких единиц текста позволит получать более точные результаты автоматического выделения контекстных синонимов, что во многом улучшит качество автоматического анализа текстов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Bender, E. M.* Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. / E. M. Bender. – Synthesis Lectures on Human Language Technologies. – London: Morgan & Claypool, 2019 – 268 p. DOI:10.2200/ S00935ED1V02Y-201907HLT043

2. *Реформатский, А. А.* Введение в языковедение / под ред. В.А. Виноградова. – Москва: Аспект-Пресс, 1999. – 536 с.

3. *Путятина, Е. И.* Контекстуальная синонимия в тексте и его дискурсе / Е. И. Путятина // Вестник КГУ им. Н.А. Некрасова. – 2016. – № 4. – С. 148–151.

4. *Зализняк, А. А.* Феномен многозначности и способы его описания / А. А. Зализняк // Вопросы языкознания. – 2004. – № 2. – С. 20–45.

5. *Полицына, Е. В.* Реализация двухуровневого семантического представления текста в открытой системе автоматизированной обработки текста / Е. В. Полицына, С. А. Полицын // Проблемы компьютерной лингвистики и типологии, Сборник научных трудов. – 2017. – № 6. – С. 98–105.

6. *Zeng, Xian-Mo.* Semantic relationships between contextual synonyms // US-China Education Review. 2007. Vol. 4. № 9. pp. 33–37.

7. *Белькова, А. Е.* Контекстуальные синонимы как стилистическое средство выразительности в языке поэзии В. А. Мазина / А. Е. Белькова // Вестник НВГУ. – Воронеж, 2014. – № 4 – С. 1–7.

8. *Милованова, Е. Е.* Применение нейронных сетей для распознавания контекстных синонимов / Е. Е. Милованова, Е. В Полицына // Г12 «Гагаринские чтения – 2020»: Сборник тезисов докладов. – Москва, 2020. – С. 1731.

9. *Полицына, Е. В.* Проблема и алгоритм автоматического выделения контекстных синонимов из текстов на русском языке / Е. В. Полицына, С. А. Полицын, А. С. Поречный, Е. Е. Милованова // Информатика: проблемы, методы, технологии: сборник материалов XX международной научно-методической конференции / под редакцией А. А. Зацаринного, Д. Н. Борисова. – Воронеж: Издательство «Научно-исследовательские публикации» (ООО «Вэлборн»), 2020. – С. 1663–1669.

10. *Politsyna, E. V.* The Framework for Hypothesis Verification and Analysis of Natural Language Processing for the Russian Language / E. V. Politsyna, S. A. Politsyn, A. S. Porechny // Supplementary Proceedings of the Seventh International Conference on Analysis of Images,

Social Networks and Texts (AIST-SUP 2018). – 2018. – V. 2268. P. 25–33.

11. *Bisera, K.-S.* The semantic aspect of the acquisition of synonyms, homonyms and antonyms in the teaching process of English as a foreign language // European Journal of Foreign Language Teaching. – 2018. – № 3. – P. 28–43.

12. *Батура, Т. В.* Методы автоматической классификации текстов / Т. В. Батура // Программные продукты и системы. – Тверь: Закрытое акционерное общество Научно-исследовательский институт «Центрпрограммсистем», 2017. – Т. 1. – С. 85–99.

13. *Айвазян С. А.* Классификация многомерных наблюдений // С. А. Айвазян, З. И. Бежаева, О. В. Староверов – Москва. : Статистика, 1974. – 240 с.

14. *Кохонен, Т.* Самоорганизующиеся карты / Ю. В. Тюменцев. – Москва : БИНОМ. Лаборатория знаний, 2008. – 143 с.

15. *Hemming, C.* Using Neural Networks in Linguistic Resources. / C. Hemming. // Department of Languages, University College of Skövde, Swedish National Graduate School of Language Technology, 2003.

16. Сервис автоматического реферирования текста. – Режим доступа: http://abstracts.textanalysis.ru/ (Дата обращения: 02.10.2020).

17. Официальная страница Tools for Automated Work with Text (TAWT). – Режим доступа: https://textanalysis.ru/jce/details/tawt (Дата обращения: 01.10.2020).

Полицына Екатерина Валерьевна — канд. техн. наук, доцент, институт №3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
E-mail: kathrin.beaver@mail.ru
ORCID iD: https://orcid.org/0000-0002-9313-4766

Полицын Сергей Александрович — канд. техн. наук, доцент, институт №3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
E-mail: pul_forever@mail.ru
ORCID iD: https://orcid.org/0000-0002-0744-6035

Поречный Александр Сергеевич — аспирант, институт №3, кафедра 319, Московский авиационный институт (Национальный исследовательский университет).
E-mail: alex.porechny@mail.ru
ORCID iD: https://orcid.org/0000-0003-2280-7406

Милованова Екатерина Евгеньевна — магистрант 2-го года обучения кафедры 319, институт №3, Московский авиационный институт (Национальный исследовательский университет).
E-mail: milovanova.e1997@gmail.com
ORCID iD: https://orcid.org/0000-0003-3869-6851

DOI: https://doi.org/10.17308/sait.2020.3/3046 ISSN 1995-5499
Received 20.09.2020
Accepted 30.09.2020

ANALYSIS OF APPROACHES TO THE AUTOMATIC EXTRACTION OF CONTEXTUAL SYNONYMS FROM TEXTS IN THE RUSSIAN LANGUAGE

© 2020 E. V. Politsyna✉, S. A. Politsyn, A. S. Porechny, E. E. Milovanova

Moscow Aviation Institute (National Research University)
4, Volokolamskoe Highway, 125993 Moscow, , Russian Federation

Annotation. The article addresses the problems of determining contextual synonyms and provides the results of the analysis of approaches to their automatic extraction from texts in the Russian language. Several approaches are proposed. These approaches are based on the use of linguistic algorithms to determine similar semantic-syntactic structures in texts and their combination with machine learning methods. The results from the application of the algorithms were analysed. The algorithms were based on morphological, syntactic, and semantic analysis of the text, filtering the results by keywords, and using various clustering tools. In the conclusion, the applicability of the implemented approaches was discussed, and the directions for the development of a combination of these approaches were determined.
Keywords: contextual synonyms, automated extraction of synonyms, synonymy, semantic-syntactic analysis of the text.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Bender E. M.* Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics / E. M. Bender. – Synthesis Lectures on Human Language Technologies. – London : Morgan & Claypool, 2019 – 268 p. DOI:10.2200 /S00935ED1V02Y-201907HLT043. (in English)

2. *Reformatskij A. A.* (ed.) Vvedenie v jazykovedenie [Introduction to linguistics]. Moscow, Aspekt-Press. 1999. (in Russian)

3. *Putjatina E. I.* Kontekstual'naja sinonimija v tekste i ego diskurse [Contextual synonymy in the text and its discourse] Vestnik KGU im. N.A. Nekrasova. 2016. 4. P. 148–151. (in Russian)

4. *Zaliznjak A. A.* Fenomen mnogoznachnosti i sposoby ego opisanija [The phenomenon of polysemy and ways to describe it] Voprosy jazykoznanija. 2004. 4. P. 20–45. (in Russian)

5. *Politsyna E. V. & Politsyn S. A.* Realizacija dvuhurovnevogo semanticheskogo predstavlenija teksta v otkrytoj sisteme avtomatizirovannoj obrabotki teksta [The two-level semantic text representation in the open automated text processing system] Problemy komp'juternoj lingvistiki i tipologii, Sbornik nauchnyh trudov. 2017. 6. P. 98–105. (in Russian)

6. *Zeng Xian-mo* Semantic relationships between contextual synonyms: US-China Education Review. 2007. 4 (9). P. 33–37. (in English)

7. *Bel'kova, A. E.* Kontekstual'nye sinonimy kak stilisticheskoe sredstvo vyrazitel'nosti v jazyke pojezii [Contextual synonyms as a stylistic means of expression in the language of poetry] Vestnik NVGU. 2014. 4. P. 1–7. (in Russian)

8. *Milovanova E. E.* Primenenie nejronnyh setej dlja raspoznavanija kontekstnyh sinonimov [Using neural networks to recognize contextual synonyms] XLVI Gagarin Science Conference. Collection of abstracts. Moscow, Moscow Aviation Institute (National Research Univesity). 2020. P. 1731. (in Russian)

9. *Politsyna E. V., Politsyn S. A., Porechny A. S. & Milovanova E. E.* Problema i algoritm avtomat-

✉ Politsyna Ekaterina V.
e-mail: kathrin.beaver@mail.ru

icheskogo vydelenija kontekstnyh sinonimov iz tekstov na russkom jazyke [The problem and the algorithm of automatic extraction of contextual synonyms from the texts in Russian language] IPMT-2020. Voronezh, Izdatel'stvo «Nauchno-issledovatel'skie publikacii» (OOO «Vjelborn»). 2020. P. 1663–1669. (in Russian)

10. *Politsyna E. V., Politsyn S. A. & Porechny A. S.* The Framework for Hypothesis Verification and Analysis of Natural Language Processing for the Russian Language (2018) Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018). 2018. V. 2268. P. 25–33. (in English)

11. *Bisera, K.-S.* The semantic aspect of the acquisition of synonyms, homonyms and antonyms in the teaching process of English as a foreign language. European Journal of Foreign Language Teaching. 2018. 3. P. 28–43. (in English)

12. *Batura T. V.* Metody avtomaticheskoj klassifikacii tekstov [Methods for automatic text classification] Tver, Zakrytoe akcionernoe obshhestvo Nauchno-issledovatel'skij institut «Centrprogrammsistem». 2017. V. 1. (in Russian)

13. *Ajvazjan S. A., Bezhaeva Z. I. & Staroverov O. V.* Klassifikacija mnogomernyh nabljudenij [Classification of multidimensional observations]. Moscow, Statistika. 1974. (in Russian)

14. *Kohonen T.* Samoorganizujushhiesja karty [Self-Organizing Maps]. Translated from English by Tjumencev6 Ju. V. (2008) Moscow, BINOM. 2001. (in Russian)

15. *Hemming C.* Using Neural Networks in Linguistic Resources. Department of Languages, University College of Skövde, Swedish National Graduate School of Language Technology. 2003. (in English)

16. Portal "Avtomatizirovannyj analiz teksta". Automatic text summarization service. 2020 Available at: http://abstracts.textanalysis.ru/ (accessed: 2nd october 2020)

17. Portal "Avtomatizirovannyj analiz teksta". Tools for Automated Work with Text (TAWT). 2020. Available at: https://textanalysis.ru/jce/details/tawt (accessed: 2nd october 2020)

Politsyna Ekaterina V. — PhD in Technical Sciences, Associate Professor, Institute 3, Department 319, Moscow Aviation Institute (National Research University).
E-mail: kathrin.beaver@mail.ru
ORCID iD: https://orcid.org/0000-0002-9313-4766

Politsyn Sergey A. — PhD in Technical Sciences, Associate Professor, Institute 3, Department 319, Moscow Aviation Institute (National Research University).
E-mail: pul_forever@mail.ru
ORCID iD: https://orcid.org/0000-0002-0744-6035

Porechny Alexander S. — PhD student, Institute 3, Department 319, Moscow Aviation Institute (National Research University).
E-mail: alex.porechny@mail.ru
ORCID iD: https://orcid.org/0000-0003-2280-7406

Milovanova Ekaterina E. — second year master's student, Institute 3, Department 319, Moscow Aviation Institute (National Research University).
E-mail: milovanova.e1997@gmail.com
ORCID iD: https://orcid.org/0000-0003-3869-6851