Terra Linguistica. 2023, Том 14, № 2. С. 31–44.
Terra Linguistica. 2023, Vol. 14, No. 2. Pp. 31–44.

Научная статья
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.14204

РАЗРАБОТКА И ОЦЕНКА АЛГОРИТМА ЛЕКСИЧЕСКОЙ СУБСТИТУЦИИ ДЛЯ РУССКОГО ЯЗЫКА НА ОСНОВЕ ПРЕДСКАЗЫВАЮЩИХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

А.В. Леонтьев ✉ , **О.А. Митрофанова**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ av_leontyev@mail.ru

Аннотация. Статья посвящена опыту разработки и оценки алгоритма лексической субституции для русского языка. Задача лексической субституции, заключающаяся в подборе подходящей замены для целевого слова в контексте, активно исследовалась в течение последних нескольких десятилетий применительно к английскому и некоторым другим европейским языкам, но не русскому. Кроме того, большинство алгоритмов не принимают во внимание тип семантических отношений, связывающих замены с целевым словом. Алгоритм, рассматриваемый в статье, работает с русским языком и подбирает замены трех типов: синонимы, гиперонимы и гипонимы целевого слова. Для отбора кандидатов используется лексическая база данных RuWordNet, а в основе алгоритма ранжирования кандидатов лежат статичные предсказывающие векторные представления слов fastText. Оценка алгоритма лексической субституции проведена посредством психолингвистического эксперимента, результаты которого анализируются в статье. Полученные результаты могут представлять интерес для специалистов в области компьютерной лингвистики и искусственного интеллекта и могут быть применены в таких задачах обработки и анализа естественного языка, как перефразирование, машинный перевод, упрощение текстов, а также в лингводидактике.

Ключевые слова: лексическая субституция, дистрибутивная семантика, синонимия, векторные представления слов, RuWordNet, fastText, психолингвистический эксперимент.

Для цитирования: Леонтьев А.В., Митрофанова О.А. Разработка и оценка алгоритма лексической субституции для русского языка на основе предсказывающих нейросетевых моделей // Terra Linguistica. 2023. Т. 14. № 2. С. 31–44. DOI: 10.18721/JHSS.14204

Research article
DOI: https://doi.org/10.18721/JHSS.14204

DEVELOPMENT AND EVALUATION OF THE LEXICAL SUBSTITUTION ALGORITHM FOR RUSSIAN BASED ON PREDICTIVE NEURAL NETWORK MODELS

A.V. Leontev ✉ , **O.A. Mitrofanova**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ av_leontyev@mail.ru

Abstract. The paper deals with the lexical substitution task for the Russian language. Lexical substitution is essentially the task of determining the best suiting substitute for a given target word in context. Although the task has been actively researched for English as well as some other European languages, there is little data for Russian. Besides, few studies consider the type of semantic relations between the target word and its substitutes. Our algorithm works with Russian and produces synonym, hypernym and hyponym substitutes. We use the RuWordNet lexical database for predicting substitutes, and fastText word embeddings for the candidate ranking task. The algorithm is evaluated through psycholinguistic experiments, and the results are analyzed in the paper. The research data may be of interest for specialists in the field of computational linguistics and artificial intelligence, and be applied to such NLP tasks as paraphrasing, machine translation, text simplification, as well as linguodidactics.

Keywords: lexical substitution, distributional semantics, synonymy, word embeddings, RuWordNet, fastText, psycholinguistic experiment.

Citation: A.V. Leontev, O.A. Mitrofanova, Development and evaluation of the lexical substitution algorithm for Russian based on predictive neural network models, Terra Linguistica, 14 (2) (2023) 31–44. DOI: 10.18721/JHSS.14204

Введение

Лексическая субституция (lexical substitution) – формальное преобразование текста, определяемое как замещение одного слова другим в контексте и применяемое в практических задачах автоматической обработки текстов, таких как автоматическое разрешение лексико-семантической неоднозначности, перифразирование, упрощение текстов, информационный поиск, построение вопросно-ответных систем и генерация тестовых заданий для изучающих родной или иностранные языки. Основным требованием к лексической субституции является сохранение смысла и грамматической целостности контекста при замене слова на его альтернативу [1].

Лексическая субституция как процедура основывается на описании парадигматических и синтагматических отношений в тексте, которые, в свою очередь, задают критерии лексического выбора и взаимозаменяемости слов в контексте. С позиций лингвистической семантики и лексикографии, лексическое значение слова определяется набором признаков, которые, с одной стороны, обусловлены его парадигматическим противопоставлением какому-то другому слову или нескольким словам, а с другой стороны, определяют синтагматическую совместимость слова с его потенциальным контекстным окружением [2–4].

Процесс лексического выбора в теории лингвистических моделей «Смысл <=> Текст» [2] объясняется при помощи семантических разложений лексем и воплощается в понятии лексических функций. Особый интерес в контексте лексической субституции представляют парадигматические лексические функции (синонимия, антонимия, конверсия, гипонимия и некоторые другие). Понятие субституции как формальной процедуры шире, чем понятие замены как типа лексических функций, поскольку не всегда позволяет осуществить идентификацию семантических

типов отношений между замещающими друг друга лексическими единицами. В частности, это проявляется в неуниверсальности критериев взаимозаменяемости и контекстной нейтрализации различий в определении синонимов [5, 3]. Ясность в данный вопрос вносят исследования, проводимые на корпусном материале [6] и на материале современных дистрибутивно-семантических моделей [7].

В данной статье рассматриваются алгоритмы и наборы данных, созданные для решения задачи лексической субституции, а также предлагается решение задачи лексической субституции применительно к материалу русского языка с использованием предсказаний нейросетевых моделей fastText и информации о семантических связях целевых слов в лексической базе данных RuWordNet.

Исследования в области лексической субституции: алгоритмы и наборы данных

Задача лексической субституции изначально была сформулирована применительно к оценке качества разрешения лексической неоднозначности (WSD, Word Sense Disambiguation) [8]. Первое соревнование систем, выполняющих лексическую субституцию в английских текстах, было проведено в рамках SemEval-2007 [1]. Организаторы собрали набор данных (датасет)[1], состоящий из 2010 предложений: 201 целевое слово, по 10 предложений на каждое. В обучающие данные попали 300 предложений, которые были опубликованы для участников, а 1710 предложений составили тестовый датасет, на основе которого проводилась оценка представленных участниками систем. В качестве целевых слов были отобраны многозначные слова, имеющие не менее одного синонима. Затем данные были вручную размечены силами пяти носителей английского языка, каждый из которых обработал данные целиком, предложив от одного до трех вариантов замены целевого слова. Замены должны были иметь тождественное или несколько более общее значение, то есть в качестве результата рассматривались синонимы и гиперонимы.

Соревнование включало в себя три задания, в каждом из которых результаты оценивалось отдельно: задание **base** предполагало, что системы могли выдавать сколько угодно вариантов, при этом баллы за каждый правильный вариант делились на их общее число, в задании **oot** системы могли выдавать до 10 вариантов, баллы за каждый правильный вариант не делились на их общее число, в задании **mw** системы должны были определить, является ли целевое слово частью коллокации (multiword expression), и выделить эту коллокацию. Все три задания оценивались при помощи метрик точности (precision) и полноты (recall).

Тестовый набор данных для качественной оценки лексической субституции был сформирован на основе компьютерного тезауруса WordNet и включал а) синонимы из первого синсета целевого слова, б) согипонимы из первого синсета целевого слова, в) синонимы всех синсетов целевого слова, г) согипонимы всех синсетов целевого слова, ранжированные на основе частот из Британского национального корпуса (BNC, British National Corpus).

Проанализировав результаты соревнования и опыт ручной разметки данных, организаторы пришли к выводу о высокой сложности поставленной задачи, поскольку по своей сути лексическая субституция предполагает некоторую степень вариативности. Эксперты, занимавшиеся разметкой набора данных, предлагали альтернативные варианты замен; при этом некоторые алгоритмы предлагали допустимые замены, которые не были отмечены экспертами; более того, понятие семантической близости двух лексем вплоть до возможности их взаимного замещения не было в достаточной мере формализовано авторами задания. Этот факт осложняет оценку алгоритмов по сравнению с задачами, в которых предоставляется фиксированный набор вариантов.

Определение задачи лексической субституции, предложенное в [1], стало общепринятым, а собранные данные и метрики оценки качества субституции используются исследователями и по сей день.

[1] Данные и подробное описание задания: http://www.dianamccarthy.co.uk/task10index.html (дата обращения 04.06.2023)

Впоследствии составлялись другие датасеты, ориентированные на задачу лексической субституции. Например, в [9] описывается сбор и разметка данных для решения задачи лексической субституции в немецком языке. По формату, объему и принципам разметки датасет аналогичен представленному в [1]. Предпринимался еще ряд попыток расширить и систематизировать данные [10–12], среди них выделяется датасет CoInCo (Concepts-In-Context), описанный в [13]. CoInCo отличается расширенным набором данных, дающим реалистичную картину частотного распределения целевых слов и их значений, а также его сбалансированностью относительно двух стилей (публицистического и художественного).

Помимо работ, выполненных в русле состязания в рамках SemEval-7, есть еще ряд исследований, связанных с практическим решением задачи лексической субституции. Например, в исследовании [14] данная процедура рассматривается в контексте упрощения текстов англоязычного сегмента Википедии, при этом типовые замены выводятся из вероятностных распределений. В проекте [11] задача лексической субституции решается с опорой на кластеризацию контекстных соседей целевых слов с последующим обучением классификатора для генерации и ранжирования замен. В работе [15] задача лексической субституции сводится к задаче бинарной классификации кандидатов на замены, отбираемых из компьютерного тезауруса WordNet, и оценке соответствия значений кандидатов со значениями целевого слова в контексте. В исследовании [16] предложены три различных подхода, основанных на скрытых марковских моделях, *n*-граммных языковых моделях и грамматических правилах, которые также применяются к кандидатам, отобранным на основе словарных данных.

С развитием дистрибутивно-семантических моделей, в частности, с появлением семейства алгоритмов word2vec [17], исследования в области лексической субституции сделали большой скачок вперед. Так, в работе [18] используются векторные представления слов, обученные по алгоритму Skip-gram, для определения, с одной стороны, семантической близости кандидата на замену к целевому слову, а с другой, его уместности в контексте целевого слова. Кроме того, если более ранние работы были сосредоточены в основном на задаче ранжирования кандидатов, полученных из разного вида лексических источников, авторы данного исследования генерируют их при помощи дистрибутивно-семантических моделей. В [19] представлено расширение модели word2vec с применением двунаправленных рекуррентных нейронных сетей, кодирующее в векторах еще больше синтагматической информации и показывающее еще более высокие результаты на задаче лексической субституции. С появлением архитектуры трансформеров [20] возможности решения задачи лексической субституции сделали очередной шаг вперед. В задаче генерации кандидатов на замену, подход, в основе которого лежит модель BERT, показывает немалый прирост в значениях метрик [21]. В исследовании [22] проводится сопоставительный анализ современных нейросетевых подходов к решению задачи лексической субституции. Проведя серию экспериментов на описанных выше датасетах SemEval и CoInCo, авторы пришли к выводу, что современные нейросетевые модели без дообучения показывают результаты, сравнимые с базовыми традиционными подходами, требующие сложного и трудоемкого обучения. В этой же работе предпринимается первая попытка исследования типов семантических отношений, наблюдаемых между целевыми словами и заменами, которые предлагают различные алгоритмы.

Традиционно задача лексической субституции решается в два этапа: поиск или генерация кандидатов на замену и их ранжирование. Более ранние работы в большинстве своем отбирали кандидатов на замену с опорой на лексикографические ресурсы, в частности компьютерный тезаурус WordNet, однако современные подходы используют преимущественно дистрибутивно-семантические модели. Ранжирование кандидатов – задача менее прозрачная, а соответственно, спектр подходов к ее решению значительно шире. Так, например, в работах [23, 24] применяется подход, основанный на правилах, а авторы исследований [15, 11, 25] прибегают к обучению с учителем.

Однако доминирующее положение занимают подходы, использующие дистрибутивно-семантические модели различных типов: счетные [26–29], предсказывающие статические [30, 18], контекстуализированные [21, 22].

Анализ достижений в области лексической субституции приводит нас к двум важным наблюдениям. Во-первых, подавляющее большинство исследований лексической субституции фокусируется исключительно на английском языке. Многие существующие подходы требуют размеченных определенным образом корпусных данных для обучения, и почти все – отдельно подготовленных данных для оценки. Насколько нам известно, попытки решить задачу лексической субституции для русского языка до настоящего времени не предпринимались. Кроме того, нет подходящих данных, необходимых для обучения алгоритмов лексической субституции и для оценки результатов. Во-вторых, долгое время исследователи оставляли без должного внимания семантические типы отношений между заменами и заменяемыми словами. Согласно [1], заменами могут быть либо синонимы, либо гиперонимы. Недавние исследования показывают, что современные алгоритмы подбирают замены безотносительно связей с целевыми словами, которые могли бы быть зарегистрированы в компьютерном тезаурусе WordNet. Однако, в практических приложениях процедуры лексической субституции, таких как перифразирование и упрощение текстов, семантический тип замен оказывается важен.

Разработка нейросетевого алгоритма лексической субституции для русского языка

Предлагаемое нами решение задачи лексической субституции для русского языка предполагает использование компьютерного тезауруса RuWordNet для отбора кандидатов и предобученных векторных представлений слов fastText для их ранжирования. Исходный код решения на языке программирования Python размещен в GitHub-репозитории[2]. Оценка алгоритма проводилась посредством психолингвистического эксперимента.

При выборе источника потенциальных замен мы исходили из возможности их фильтрации по типу семантических отношений, связывающих кандидатов с целевым словом. С этой точки зрения, наилучшим ресурсом оказывается компьютерный тезаурус русского языка RuWordNet в составе библиотеки ruwordnet[3] для языка программирования Python. Тезаурус RuWordNet содержит синсеты трех частей речи: существительные (29297 синсетов), глаголы (7636 синсетов) и прилагательные (12865 синсетов). В общей сложности, текущая версия тезауруса содержит 133 745 уникальных слов и словосочетаний, 154 111 значений. Как и предполагает архитектура WordNet, между синсетами в RuWordNet проводятся связи, соответствующие различным семантическим отношениям в лексике: *гипоним-гипероним, экземпляр-класс, отношение антонимии, часть-целое, причина, логическое следование, предметная область* (*домен*).

Также между синсетами, относящимися к разным частям речи, но выражающими один и тот же смысл, установлены отношения частеречной синонимии, соединяющие разделенные синсеты. Предлагаемый нами алгоритм лексической субституции предлагает пользователю выбрать тип семантических отношений, на основании которого будут отбираться кандидаты: синонимия, гиперонимия и гипонимия.

Рассмотрим алгоритм отбора кандидатов.

1. Целевое слово, поступающее на вход алгоритму, проходит лемматизацию с использованием морфологического анализатора pymorphy2, доступного для использования в виде библиотеки на языке программирования Python.

2. В тезаурусе RuWordNet выполняется поиск синсетов, содержащих целевое слово. Если целевого слова нет в тезаурусе, выбирается ближайшее (по косинусному расстоянию) к целевому слову слово в выбранной модели fastText, которое присутствует в тезаурусе.

[2] URL: https://github.com/zatoulanypes/lexsub (дата обращения: 04.06.2023)
[3] URL: https://github.com/avidale/python-ruwordnet (дата обращения: 04.06.2023)

3. Производится отбор кандидатов на замену целевого слова:

a. если выполняется отбор синонимов, то берутся все значения соответствующих целевому слову синсетов;

b. если выполняется отбор гиперонимов, берутся все значения всех синсетов, которые являются гиперонимами любого из синсетов, соответствующих целевому слову;

c. если выполняется отбор гипонимов, выбираются все значения синсетов, отмеченных в тезаурусе как гипонимы любого из синсетов, соответствующих целевому слову.

4. Алгоритм возвращает множество кандидатов на замену целевого слова, при этом в выдаче могут быть как слова, так и словосочетания.

Из множества кандидатов удаляется само целевое слово, а также словосочетания, содержащие его в именительном падеже.

В предложенном алгоритме ранжирование кандидатов выполняется с использование векторных моделей fastText [31] для русского языка, предобученных и размещенных на портале RusVectōrēs [32]. Обусловлено это тем, что, в отличие от моделей семейства word2vec, модели fastText способны выдавать векторные представления даже для слов и словосочетаний, которые отсутствуют в обучающем корпусе, а также, в отличие от моделей семейства BERT, не требуют дополнительных объемов данных для дообучения. Для нашего исследования были выбраны три модели:

- **geowac_lemmas_none_fasttextskipgram_300_5_2020**, предобученная на корпусе текстов GeoWAC,
- **ruscorpora_none_fasttextskipgram_300_2_2019**, предобученная на текстах Национального корпуса русского языка,
- **tayga_none_fasttextcbow_300_10_2019**, предобученная на корпусе текстов Taiga.

Вслед за [18] мы используем четыре метрики для определения ранга кандидатов, которые оценивают близость замены как к целевому слову, так и к его контексту:

- Add – арифметическое среднее: $\frac{\cos(s,t)+\sum_{c\in C}\cos(s,c)}{|C|+1}$,
- Mult – геометрическое среднее: $\sqrt[|C|+1]{p\cos(s,t)\cdot\prod_{c\in C}p\cos(s,c)}$,
- BalAdd – «сбалансированное» арифметическое среднее: $\frac{|C|\cdot\cos(s,t)+\sum_{c\in C}\cos(s,c)}{2\cdot|C|}$,
- BalMult – «сбалансированное» геометрическое среднее: $\sqrt[2\cdot|C|]{p\cos(s,t)^{|C|}\cdot\prod_{c\in C}p\cos(s,c)}$,

где s – слово-кандидат на замену, t – целевое слово, C – контекст целевого слова, представленный в виде массива слов c, находящихся в рамках определенного окна вокруг t, а $p\cos(v,v')=\frac{\cos(v,v')+1}{2}$ используется для того, чтобы Mult и BalMult не принимали отрицательные значения. «Сбалансированные» метрики исключают из влияющих на результат факторов величину контекста [18].

Рассмотрим алгоритм ранжирования кандидатов на лексическую субституцию.

1. Исходное предложение токенизируется. В полученном массиве токенов определяется индекс целевого слова, после чего выделяются контексты целевого слова в рамках заданного окна, по умолчанию выбираемого как [–2; +2].

2. Для каждого кандидата на замену:

a. из выбранной пользователем модели берутся векторные представления целевого слова, рассматриваемого кандидата и каждого слова из контекста целевого слова;

b. полученные векторные представления используются для расчета одной из описанных выше метрик, по умолчанию, Add.

3. Все кандидаты ранжируются по значению метрики для данного кандидата в данном контексте.

Таким образом, на выход алгоритма поступает сортированный массив пар целевых слов и соответствующих им значений метрики.

Оценка нейросетевого алгоритма лексической субституции для русского языка

Для оценки качества разработанного нами нейросетевого алгоритма лексической субституции для русского языка был проведен психолингвистический эксперимент, для которого были выбраны девять целевых слов: по три существительных (*вид, подпись, сон*), глагола (*звать, кричать, стоить*) и прилагательных (*близкий, верный, сильный*). Каждое из слов должно иметь несколько значений, которые бы регулярно реализовывались в речи, и для каждого из слов в как минимум в двух его значениях можно было бы подобрать различные замены, значения которых бы не пересекались. Из Национального корпуса русского языка (НКРЯ)[4] были извлечены 18 предложений, по два предложения на каждое целевое слово. В содержательном плане контексты выбирались таким образом, чтобы респонденты концентрировались не на экстралингвистической информации, а на лингвистических признаках целевого слова и его контекстного окружения.

Собранные данные были обработаны алгоритмом автоматической лексической субституции на базе трех векторных моделей с контекстным окном [–2;+2], при этом учитывались три типа семантических отношений для подбора кандидатов: синонимия, гиперонимия, гипонимия. В итоге были получены 54 набора замен, отсортированных алгоритмом согласно их рангу: от наиболее подходящих к наименее подходящим.

В эксперименте приняли участие 30 респондентов. Испытуемым давались предложения с выделенными целевыми словами и до 10 вариантов замен, предложенных алгоритмом. В опросе было 18 разделов, по одному предложению на каждый раздел. Для каждого предложения было три набора замен, сгенерированных алгоритмом на основе трех различных векторных моделей: geowac, ruscorpora, tayga. Требовалось, чтобы респонденты выбрали все замены, которые они считают подходящими в данном контексте. Вопросы, в которых респонденты не находили ни одного подходящего слова, они должны были пропустить. В табл. 1 приведены примеры кандидатов в замены для целевого прилагательного *сильный* в контексте *Эле-Фантик почувствовал себя* ***сильным*** *и взрослым дядькой, вроде того, что видел во сне*.

В результате обработки результатов психолингвистического эксперимента было обнаружено, что для каждого набора замен как минимум один респондент счел хотя бы один вариант приемлемым. Худшими оказались замены для глагола *звать* в контексте *Сюжет блестящий, обычный* (*переводя на современный лад*) *клерк, Анатолий Новосельцев для того, чтобы подняться выше по карьерной лестнице, начинает ухаживать за «непробиваемой» для мужчин директором его фирмы, Людмилой Прокофьевной или, как ее зовут остальные работники, просто Мымра*, предложенные моделями geowac и ruscorpora (8 и 6 голосов соответственно), а также замены для существительного *подпись* в контексте *Например, Ральф Шумахер весьма скептически оценил возможность выступления в одной команде со старшим братом и в дни Гран-при Европы поспешил поставить подпись под контрактом с Уильямсом наближайшие три года*, предложенные всеми тремя моделями (14 голосов для замен, предложенных моделями geowac и ruscorpora, и 13 голосов для замен, предложенных моделью tayga). Лучший результат показали замены, предложенные для имен прилагательных: в среднем, 26 респондентов из 30 выбрали хотя бы одну замену, затем идут глаголы (25)

[4] URL: https://ruscorpora.ru/ (дата обращения: 04.06.2023)

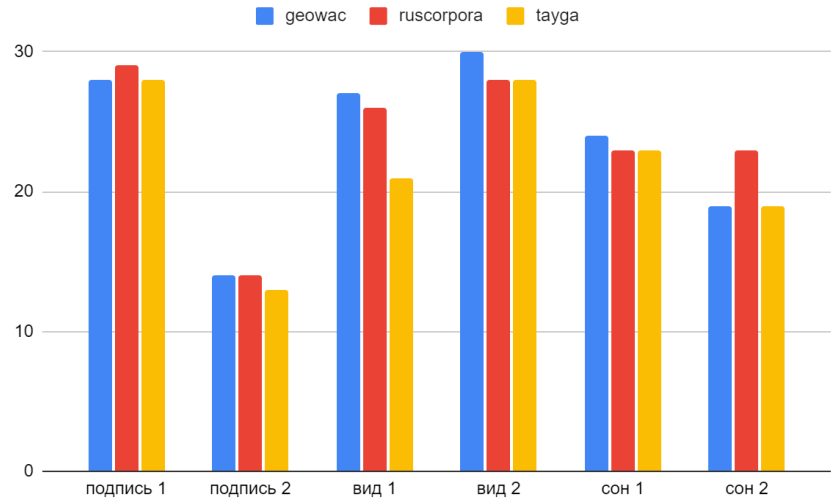

Рис. 1. Имена существительные: число респондентов, выбравших хотя бы один вариант замены из предложенных

Fig. 1. Nouns: the number of respondents who picked at least one substitute candidate

и имена существительные (23). Количественные данные об ответах респондентов представлены на рис. 1–3.

Таблица 1. Кандидаты в замены для целевого прилагательного *сильный* в контексте *Эле-Фантик почувствовал себя сильным и взрослым дядькой, вроде того, что видел во сне*, сгенерированные моделями *geowac*, *ruscorpora*, *taiga*, с учетом *N* – числа респондентов, которые сочли замены приемлемыми

Table 1. Substitute candidates for target adjective *сильный* (*strong*) in the context *Эле-Фантик почувствовал себя сильным и взрослым дядькой, вроде того, что видел во сне* (*Ele Fantik felt like a strong grown-up man, like the one he saw in the dream*) generated by the *geowac*, *ruscorpora*, and *taiga* models, where *N* is the number of respondents who marked the candidate fitting

geowac		ruscorpora		tayga	
Замены	**N**	**Замены**	**N**	**Замены**	**N**
страшным	2	физически мощным	20	непереносимым	0
неистовым	1	жестоким	2	физически мощным	18
жестоким	2	несносным	0	страстным	0
нестерпимым	0	жесточайшим	2	невыносимым	0
страстным	0	нестерпимым	0	физически крепким	22
резковатым	0	физически крепким	22	жестоким	1
яростным	2	выносливым	7	нестерпимым	0
мощным	26	страшным	2	глубоким	0
пронзительным	0	пылким	1	сильнодействующим	0
значительным	12	неотразимым	1	страшным	2

Наиболее стабильные результаты в ранжировании замен показывает модель ruscorpora: в среднем, десять респондентов выбрали первую предложенную ей замену как корректную. Затем идет geowac (9) и tayga (8). Что касается частей речи, первые предложенные алгоритмами замены ока-

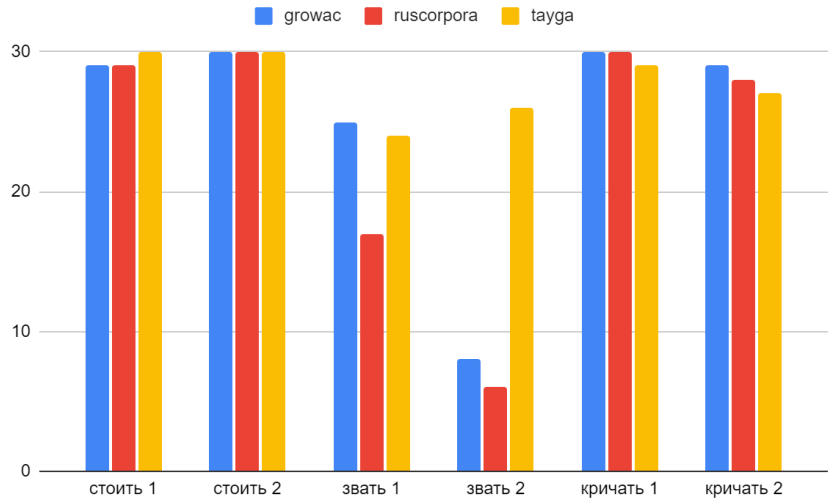

Рис. 2. Глаголы: число респондентов, выбравших хотя бы один вариант замены из предложенных
Fig. 2. Verbs: the number of respondents who picked at least one substitute candidate

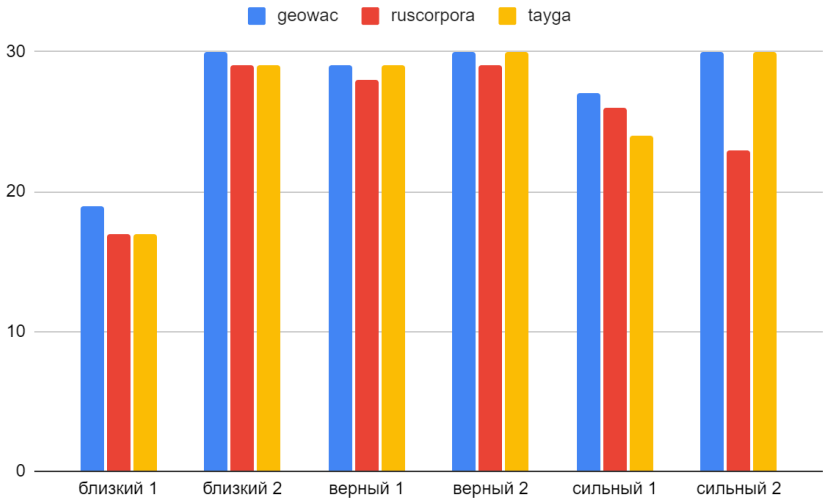

Рис. 3. Имена прилагательные: число респондентов, выбравших хотя бы один вариант замены из предложенных
Fig. 3. Adjectives: the number of respondents who picked at least one substitute candidate

зались наиболее успешными для прилагательных: в среднем, 13 респондентов посчитали их подходящими. Следом идут глаголы (8) и существительные (7).

Наибольшее число замен, признанных респондентами подходящими, было предложено моделями tayga и ruscorpora (в среднем, 4), а модель geowac в среднем предлагала 3 подходящие замены. Среди частей речи в первенстве по разнообразию подходящих замен лидируют глаголы (в среднем, 4), следом идут существительные (3,5) и прилагательные (3).

Факт лидирования глаголов по числу приемлемых замен можно объяснить их богатой морфологией. Среди предложенных алгоритмами замен много приставочных образований и видовых пар (*кричать → накричать, закричать, крикнуть*). Весьма богатая морфология имен существи-

тельных также может объяснить широкий выбор замен-существительных (*во сне → в дремоте, дреме, полудреме*).

Важно заметить, что предложенный нами метод позволяет производить лексическую субституцию с сохранением морфосинтаксических параметров целевого слова (например, *близких – близстоящих*, род. п. мн. ч.) и возможностью подбора словосочетаний в качестве замен (например, *близких – близко расположенных, ближайших по расстоянию*). Словоизменение выполнялось с использованием Python-библиотеки pymorphy2.

Перспективы дальнейшего исследования

Полученные результаты показывают, что алгоритм справляется с поставленной задачей. Хотя эксперимент оказался интересным и показательным методом оценки работы алгоритма, для сравнения различных подходов и оценки динамики качества при развитии алгоритма необходим неизменный проверочный набор данных. Перспективы дальнейшей работы включают подготовку проверочных данных, а также оценку потенциала контекстуализированных векторных представлений слов как применительно к задаче ранжирования кандидатов на замену, так и к задаче их отбора. Открытым, однако, остается вопрос возможности контроля семантического типа замен при подобном подходе.

Заключение

В данной статье описан алгоритм лексической субституции для русского языка с использованием компьютерного тезауруса RuWordNet для отбора кандидатов на замену и векторных представлений слов fastText для их ранжирования, а также проведен анализ результатов оценки алгоритма посредством психолингвистического эксперимента. Материалом для эксперимента послужили восемнадцать предложений, отобранных вручную из НКРЯ, содержащих девять целевых слов: три имени существительных, три имени прилагательных и три глагола. Для данных целевых слов в данных контекстах были предложены замены при помощи разработанного алгоритма на основе трех векторных моделей fastText, предобученных на данных различных корпусов русского языка: НКРЯ, GeoWAC, Taiga. Приемлемость полученных замен была оценена тридцатью респондентами. На основе данных эксперимента были сделаны выводы о качестве разработанного алгоритма и выявлены перспективы его развития.

Результаты данного исследования могут быть полезны для специалистов в области компьютерной лингвистики и искусственного интеллекта, а также могут послужить толчком дальнейших исследований по лексической субституции в русском языке.

СПИСОК ИСТОЧНИКОВ

1. **McCarthy D., Navigli R.** SemEval-2007 Task 10: English Lexical Substitution Task // Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) SemEval 2007. Prague, Czech Republic: Association for Computational Linguistics, 2007. SemEval-2007 Task 10. Pp. 48–53.

2. **Мельчук И.А.** Язык: от смысла к тексту. М.: Языки славянской культуры, 2012. 176 с.

3. **Шмелев Д.Н.** Проблемы семантического анализа лексики. М.: Наука, 1973. 280 с.

4. **Уфимцева А.А.** Лексическое значение: Принцип семиологического описания лексики. М.: Наука, 1986. 240 с.

5. **Апресян Ю.Д.** Исследования по семантике и лексикографии. Т. I: Парадигматика. М.: Языки славянских культур, 2009. 568 с.

6. **Белов В.А.** Взаимозаменяемость как критерий синонимии (экспериментальное и корпусное исследование) // Вестник Санкт-Петербургского университета. Язык и литература. 2018. Т. 15, № 3. С. 390–411.

7. **Тимофеева М.К.** Типология семантических отношений, выделяемых посредством инструмента RusVectōrēs // Научный диалог. 2018. № 8. С. 74–87.

8. **McCarthy D.** Lexical Substitution as a Task for WSD Evaluation // Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. – Association for Computational Linguistics, 2002. Pp. 89–115.

9. **Cholakov K., Biemann C., Eckle-Kohler J., Gurevych I.** Lexical Substitution Dataset for German // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. Pp. 1406–1411.

10. **Sinha R., Mihalcea R.** Explorations in lexical sample and all-words lexical substitution // Natural Language Engineering, 2014, Vol. 20, No. 1, Pp. 99–129.

11. **Biemann C.** Co-Occurrence Cluster Features for Lexical Substitutions in Context // Proceedings of TextGraphs-5 – 2010 Workshop on Graph-based Methods for Natural Language Processing TextGraphs 2010. Uppsala, Sweden: Association for Computational Linguistics, 2010. Pp. 55–59.

12. **McCarthy D., Sinha R., Mihalcea R.** The cross-lingual lexical substitution task // Language resources and evaluation, 2013, Vol. 47, Pp. 607–638.

13. **Kremer G., Erk K., Padó S., Thater S.** What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics EACL 2014. – Gothenburg, Sweden: Association for Computational Linguistics, 2014. Pp. 540–549.

14. **Yatskar M., Pang B., Danescu-Niculescu-Mizil C., Lee L.** For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2010. Los Angeles, California: Association for Computational Linguistics, 2010. Pp. 365–368.

15. **Dagan I., Glickman O., Gliozzo A., Marmorshtein E., Strapparava C.** Direct Word Sense Matching for Lexical Substitution // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics COLING-ACL 2006. – Sydney, Australia: Association for Computational Linguistics, 2006. Pp. 449–456.

16. **Preiss J., Coonce A., Baker B.** HMMs, GRs, and N-Grams as Lexical Substitution Techniques – Are They Portable to Other Languages? // Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning. Borovets, Bulgaria: Association for Computational Linguistics, 2009. Pp. 21–27.

17. **Mikolov T., Chen K., Corrado G., Dean J.** Efficient Estimation of Word Representations in Vector Space/arXiv:1301.3781 [cs]. – arXiv, 2013.

18. **Melamud O., Dagan I., Goldberger J.** Modeling Word Meaning in Context with Substitute Vectors // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2015. Denver, Colorado: Association for Computational Linguistics, 2015. Pp. 472–482.

19. **Melamud O., Goldberger J., Dagan I.** context2vec: Learning Generic Context Embedding with Bidirectional LSTM // Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning CoNLL 2016. Berlin, Germany: Association for Computational Linguistics, 2016. Pp. 51–61.

20. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.** Attention Is All You Need/arXiv:1706.03762 [cs]. – arXiv, 2017.

21. **Zhou W., Ge T., Xu K., Wei F., Zhou M.** BERT-based Lexical Substitution // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics ACL 2019. Florence, Italy: Association for Computational Linguistics, 2019. Pp. 3368–3373.

22. **Arefyev N., Sheludko B., Podolskiy A., Panchenko A.** A Comparative Study of Lexical Substitution Approaches based on Neural Language Models/arXiv:2006.00031 [cs]. – arXiv, 2020.

23. **Yuret D.** KU: word sense disambiguation by substitution // Proceedings of the 4th International Workshop on Semantic Evaluations: SemEval '07. USA: Association for Computational Linguistics, 2007. Pp. 207–213.

24. **Hassan S., Csomai A., Banea C., Sinha R., Mihalcea R.** UNT: SubFinder: Combining Knowledge Sources for Automatic Lexical Substitution // Proceedings of the Fourth International Workshop on Se-

mantic Evaluations (SemEval-2007) SemEval 2007. Prague, Czech Republic: Association for Computational Linguistics, 2007. Pp. 410–413.

25. **Szarvas G., Busa-Fekete R., Hüllermeier E.** Learning to Rank Lexical Substitutions // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing EMNLP 2013. Seattle, Washington, USA: Association for Computational Linguistics, 2013. Pp. 1926–1932.

26. **Erk K., Padó S.** A Structured Vector Space Model for Word Meaning in Context // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing EMNLP 2008. Honolulu, Hawaii: Association for Computational Linguistics, 2008. Pp. 897–906.

27. **Dinu G., Lapata M.** Measuring Distributional Similarity in Context // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing EMNLP 2010. Cambridge, MA: Association for Computational Linguistics, 2010. Pp. 1162–1172.

28. **Thater S., Fürstenau H., Pinkal M.** Contextualizing Semantic Representations Using Syntactically Enriched Vector Models // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL 2010. Uppsala, Sweden: Association for Computational Linguistics, 2010. Pp. 948–957.

29. **Apidianaki M.** Vector-space models for PPDB paraphrase ranking in context // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing EMNLP 2016. Austin, Texas: Association for Computational Linguistics, 2016. Pp. 2028–2034.

30. **Yuret D.** FASTSUBS: An Efficient and Exact Procedure for Finding the Most Likely Lexical Substitutes Based on an N-gram Language Model // IEEE Signal Processing Letters, 2012, Vol. 19, FASTSUBS, No. 11, pp. 725–728.

31. **Bojanowski P., Grave E., Joulin A., Mikolov T.** Enriching Word Vectors with Subword Information/ arXiv:1607.04606 [cs]. – arXiv, 2017.

32. **Kutuzov A., Kuzmenko E.** WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers/ eds. D.I. Ignatov et al. Cham: Springer International Publishing, 2017. Pp. 155–161.

REFERENCES

[1] **D. McCarthy, R. Navigli,** SemEval-2007 Task 10: English Lexical Substitution Task, Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) SemEval 2007. – Prague, Czech Republic: Association for Computational Linguistics, 2007. SemEval-2007 Task 10. Pp. 48–53.

[2] **I.A. Melchuk,** Yazyk: ot smysla k tekstu [Language: from Sense to Text]. M.: Yazyki slavyanskoy kultury, 2012.

[3] **D.N. Shmelev,** Problemy semanticheskogo analiza leksiki [Problems of Lexical Semantic Analysis]. M.: Nauka, 1973.

[4] **A.A. Ufimtseva,** Leksicheskoye znacheniye: Printsip semiologicheskogo opisaniya leksiki [Lexical Meaning: The Principle of Semiological Lexical Description], M.: Nauka, 1986.

[5] **Yu.D. Apresyan,** Issledovaniya po semantike i leksikografii. T. I: Paradigmatika [Studies in Semantics and Lexicography. V. I: Paradigmatics]. M.: Yazyki slavyanskikh kultur, 2009.

[6] **V.A. Belov,** Vzaimozamenyayemost kak kriteriy sinonimii (eksperimentalnoye i korpusnoye issledovaniye) [Interchangeability as a Criterion of Synonymy (Experimental and Corpus Study)], Vestnik Sankt-Peterburgskogo universiteta. Yazyk i literatura, 15 (3) (2018) 390–411.

[7] **M.K. Timofeyeva,** Tipologiya semanticheskikh otnosheniy, vydelyayemykh posredstvom instrumenta RusVectōrēs [Typology of semantic relations extracted by the RusVectōrēs tool], Nauchnyy dialog, 8 (2018) 74–87.

[8] **D. McCarthy,** Lexical Substitution as a Task for WSD Evaluation, Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. – Association for Computational Linguistics, 2002. Pp. 89–115.

[9] **K. Cholakov, C. Biemann, J. Eckle-Kohler, I. Gurevych,** Lexical Substitution Dataset for German, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. Pp. 14062–1411.

[10] **R. Sinha, R. Mihalcea,** Explorations in lexical sample and all-words lexical substitution, Natural Language Engineering, 20 (1) (2014) 99–129.

[11] **C. Biemann,** Co-Occurrence Cluster Features for Lexical Substitutions in Context, Proceedings of TextGraphs-5 – 2010 Workshop on Graph-based Methods for Natural Language Processing TextGraphs 2010. Uppsala, Sweden: Association for Computational Linguistics, 2010. Pp. 55–59.

[12] **D. McCarthy, R. Sinha, R. Mihalcea,** The cross-lingual lexical substitution task, Language resources and evaluation, 47 (2013) 607–638.

[13] **G. Kremer, K. Erk, S. Padó, S. Thater,** What Substitutes Tell Us – Analysis of an "All-Words" Lexical Substitution Corpus, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics EACL 2014. Gothenburg, Sweden: Association for Computational Linguistics, 2014. Pp. 540–549.

[14] **M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, L. Lee,** For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2010. Los Angeles, California: Association for Computational Linguistics, 2010. Pp. 365–368.

[15] **I. Dagan, O. Glickman, A. Gliozzo, E. Marmorshtein, C. Strapparava,** Direct Word Sense Matching for Lexical Substitution, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics COLING-ACL 2006. Sydney, Australia: Association for Computational Linguistics, 2006. Pp. 449–456.

[16] **J. Preiss, A. Coonce, B. Baker,** HMMs, GRs, and N-Grams as Lexical Substitution Techniques – Are They Portable to Other Languages?, Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning. – Borovets, Bulgaria: Association for Computational Linguistics, 2009. Pp. 21–27.

[17] **T. Mikolov, K. Chen, G. Corrado, J. Dean,** Efficient Estimation of Word Representations in Vector Space/arXiv:1301.3781 [cs]. – arXiv, 2013.

[18] **O. Melamud, I. Dagan, J. Goldberger,** Modeling Word Meaning in Context with Substitute Vectors, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL-HLT 2015. Denver, Colorado: Association for Computational Linguistics, 2015. Pp. 472–482.

[19] **O. Melamud, J. Goldberger, I. Dagan,** context2vec: Learning Generic Context Embedding with Bidirectional LSTM, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning CoNLL 2016. Berlin, Germany: Association for Computational Linguistics, 2016. Pp. 51–61.

[20] **A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin,** Attention Is All You Need/arXiv:1706.03762 [cs]. – arXiv, 2017.

[21] **W. Zhou, T. Ge, K. Xu, F. Wei, M. Zhou,** BERT-based Lexical Substitution, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics ACL 2019. – Florence, Italy: Association for Computational Linguistics, 2019. Pp. 3368–3373.

[22] **N. Arefyev, B. Sheludko, A. Podolskiy, A. Panchenko,** A Comparative Study of Lexical Substitution Approaches based on Neural Language Models/arXiv:2006.00031 [cs]. – arXiv, 2020.

[23] **D. Yuret,** KU: word sense disambiguation by substitution, Proceedings of the 4th International Workshop on Semantic Evaluations: SemEval '07. USA: Association for Computational Linguistics, 2007. Pp. 207–213.

[24] **S. Hassan, A. Csomai, C. Banea, R. Sinha, R. Mihalcea,** UNT: SubFinder: Combining Knowledge Sources for Automatic Lexical Substitution, Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) SemEval 2007. Prague, Czech Republic: Association for Computational Linguistics, 2007. Pp. 410–413.

[25] **G. Szarvas, R. Busa-Fekete, E. Hüllermeier,** Learning to Rank Lexical Substitutions, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing EMNLP 2013. Seattle, Washington, USA: Association for Computational Linguistics, 2013. Pp. 1926–1932.

[26] **K. Erk, S. Padó,** A Structured Vector Space Model for Word Meaning in Context, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing EMNLP 2008. Honolulu, Hawaii: Association for Computational Linguistics, 2008. Pp. 897–906.

[27]. **G. Dinu, M. Lapata,** Measuring Distributional Similarity in Context, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing EMNLP 2010. Cambridge, MA: Association for Computational Linguistics, 2010. Pp. 1162–1172.

[28] **S. Thater, H. Fürstenau, M. Pinkal,** Contextualizing Semantic Representations Using Syntactically Enriched Vector Models, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL 2010. Uppsala, Sweden: Association for Computational Linguistics, 2010. Pp. 948–957.

[29] **M. Apidianaki,** Vector-space models for PPDB paraphrase ranking in context, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing EMNLP 2016. Austin, Texas: Association for Computational Linguistics, 2016. Pp. 2028–2034.

[30] **D. Yuret,** FASTSUBS: An Efficient and Exact Procedure for Finding the Most Likely Lexical Substitutes Based on an N-gram Language Model, IEEE Signal Processing Letters, 19 (11) (2012) 725–728.

[31] **P. Bojanowski, E. Grave, A. Joulin, T. Mikolov,** Enriching Word Vectors with Subword Information/arXiv:1607.04606 [cs]. – arXiv, 2017.

[32] **A. Kutuzov, E. Kuzmenko,** WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7–9, 2016, Revised Selected Papers/ eds. D.I. Ignatov et al. Cham: Springer International Publishing, 2017. Pp. 155–161.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Леонтьев Алексей Вячеславович
Aleksei V. Leontev
E-mail: av_leontyev@mail.ru

Митрофанова Ольга Александровна
Olga A. Mitrofanova
E-mail: o.mitrofanova@spbu.ru
ORCID: https://orcid.org/0000-0002-3008-5514

Поступила: 02.06.2023; Одобрена: 27.06.2023; Принята: 29.06.2023.
Submitted: 02.06.2023; Approved: 27.06.2023; Accepted: 29.06.2023.