УДК 81'33
DOI 10.25205/1818-7935-2021-19-3-57-68

Автоматическое обнаружение и исправление деривационных ошибок в письменной речи на русском как иностранном

А. С. Выренкова, И. Ю. Смирнов

Национальный исследовательский университет
«Высшая школа экономики»
Москва, Россия

Аннотация

Учебные корпуса представляют собой один из наиболее ценных источников статистических данных об ошибках учащихся. Например, информация из корпусов учащихся, которые изучают язык как иностранный, используется для исследований в области усвоения второго языка [Granger, 1996]. Однако достоверность содержащихся в корпусах данных зависит от качества разметки ошибок, которая чаще всего выполняется вручную и, таким образом, представляет собой трудоемкую и кропотливую процедуру для аннотаторов. Чтобы облегчить процесс разметки, в корпусах используются дополнительные инструменты, в частности спеллчекеры. В данной статье основное внимание уделяется созданию системы автоматического поиска и исправления словообразовательных ошибок. Этот тип ошибок, почти никогда не возникающий у взрослых носителей русского языка, но появляющийся у изучающих русский язык как иностранный [Chernigovskaya, Gor, 2000], был выбран потому, что их исправление вызывает большие сложности у существующих спеллчекеров. В рамках работы на материале Русского учебного корпуса (Russian Learner Corpus, http://www.web-corpora.net/RLC/) было протестировано два подхода, помогающих в решении данной проблемы. Первый, который основывается на принципе конечных автоматов [Dickinson, Herring, 2008], имеет целью обнаружить морфологические нарушения в текстах изучающих русский как иностранный. Второй, в основе работы которого лежит модель шумного канала [Brill and Moore, 2000], обеспечивает исправление выявленных ошибок. После тестирования эффективности этих двух подходов с учетом результатов их работы была предложена собственная система автокоррекции словообразовательных ошибок. В ней используются алгоритм обнаружения морфологических ошибок из подхода Dickinson, Herring и модель Continuous Bag of Words FastText, которая основывается на теории дистрибутивной семантики [Harris, 1954]. В дополнение к ним вводятся правила исправления для распространенных случаев словотворчества, а также словарь парадигм для приведения слова к той грамматической форме, в которой было употреблено исправляемое слово. Результаты работы авторской системы были апробированы на данных Русского учебного корпуса и показали свою валидность.

Ключевые слова

словообразовательные ошибки, словотворчество, машинное обучение, автоматическое обнаружение ошибок, автоматическое исправление ошибок, русский как иностранный, учебный корпус, разметка

Благодарности

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ

Для цитирования

Выренкова А. С., *Смирнов И. Ю.* Автоматическое обнаружение и исправление деривационных ошибок в письменной речи на русском как иностранном // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2021. Т. 19, № 3. С. 57–68. DOI 10.25205/1818-7935-2021-19-3-57-68

A New Approach to Automatic Detection and Correction of Derivational Errors in L2 Russian

Anastasia S. Vyrenkova, Ivan Yu. Smirnov

HSE University
Moscow, Russian Federation

Abstract

Learner corpora serve as one of the most valuable sources of statistical data on learners' errors. For instance, data from foreign-language learners' corpora can be used for the Second Language Acquisition research. However, corpora representativity strongly depends on the quality of its error markup, which is most frequently carried out manually and thus presents a time-consuming and painstaking routine for the annotators. To make annotation process easier, additional tools, such as spellcheckers, are usually used. This paper focuses on developing a program for automatic correction of derivational errors made by learners of Russian as a foreign language. Derivational errors, which are not common for adult Russian native speakers (L1), but occur quite often in written texts or speech of Russian as foreign language learners (L2) [Chernigovskaya, Gor, 2000], were chosen as scope of our research because correction of such mistakes presents a formidable challenge for existing spellcheckers. Using the data from the Russian Learner Corpus (http://www.web-corpora.net/RLC/), we tested two already existing approaches to solve such kind of problems. The first one is based on a finite state automaton principle developed by Dickinson and Herring 2008, and it was tested as algorithm for derivational errors detection. The second one which relies on the Noisy Channel model by Brill and Moore, 2000, was used for studying errors correction. After we analyzed effectiveness of these tests, we developed our own system for autocorrection of derivational errors. In our program the algorithm of Dickinson and Herring was used as word-formation error detection module. The Noisy Channel model has been rejected, and we decided to use instead the Continuous Bag of Words FastText model, based on Harris distributional semantics theory [1954]. In addition, filtering rules have been developed for correcting frequent errors that the model is unable to handle. To restore automatically the correct grammatical word form, dictionary of word paradigms is used. Model results were validated on the data of Russian Learner Corpus.

Keywords

derivational errors, machine learning, automatic error detection, automatic error correction, Russian as a foreign language, learner corpus, corpus annotation

Acknowledgements

Research has been completed as a part of the Higher School of Economics Basic Research Program

For citation

Vyrenkova, Anastasia S., Smirnov, Ivan Yu. A New Approach to Automatic Detection and Correction of Derivational Errors in L2 Russian. *Vestnik NSU. Series: Linguistics and Intercultural Communication,* 2021, vol. 19, no. 3, p. 57–68. (in Russ.) DOI 10.25205/1818-7935-2021-19-3-57-68

1. Введение

Учебные корпуса неродной речи представляют собой систематизированную электронную коллекцию устных и письменных текстов, порожденных неносителями языка [Копотев, 2014]. Очевидно, что они предназначены для исследователей и преподавателей L2 [Granger, 2017], однако сегодня как отдельное направление развивается проблематика автоматической обработки естественного языка, также использующая данные, порожденные неносителями [Rudzewitz et al., 2018]. В рамках этого направления, тесно связанного с моделями машинного обучения, для решения задач определения родного языка автора [Jarvis, Paquot, 2015], автоматической идентификации ошибок [Leacock et al., 2014] в качестве обучающего и тестового материала как раз используются данные учебных корпусов, созданных на речевом материале изучающих язык, в частности русский, как иностранный.

Чтобы учебный корпус соответствовал поставленным выше задачам, он должен отвечать нескольким требованиям. Во-первых, содержать большой объем аннотированных текстов. Во-вторых, гарантировать достоверность разметки предоставляемых корпусом данных: аннотация в нем должна быть корректной и следовать единым правилам, соблюсти которые в случае коллективной ручной разметки – к тому же крайне трудозатратной – очень сложно. Поэтому в решении вопроса достоверности часто используются системы автокоррекции, так

называемые спеллчекеры, которые помогают повысить качество и скорость разметки, автоматически исправляя ошибки, или же в случае затруднения предлагают выбрать вручную правильный вариант из ограниченного списка возможных решений.

Актуальная проблема состоит в том, что спеллчекеры еще не способны исправлять все типы ошибок студентов L2, поскольку их модели работы ориентированы на отклонения, аналогичные ошибкам носителей языка (L1). В частности, значительную сложность представляют для них словообразовательные ошибки. Причина носит прагматический характер и состоит в том, что, во-первых, носители их почти не допускают, и, следовательно, их нет в обучающих сетах моделей, а во-вторых, сказывается и то, что расстояние Левенштейна от правильного варианта оказывается в этом случае больше, чем при других типах ошибок. При ручной разметке деривационные ошибки тоже вызывают трудности, так как, несмотря на высокую системность словообразовательных моделей, она не абсолютна, и это вызывает объективные проблемы, выражающиеся в смешении ошибок разных типов.

Цель статьи состоит в том, чтобы предложить более надежный способ автоматической коррекции отступлений от словообразовательной морфологической нормы. Для решения указанной проблемы мы сначала протестировали на материале текстов Русского учебного корпуса [1] алгоритм по обнаружению морфологических ошибок, дополненный системой автоисправления. Эта система основывается на предобученной модели шумного канала; она показала лучший результат среди общедоступных спеллчекеров на соревновании по исправлению ошибок в русскоязычных текстах из социальных интернет-медиа. Проанализировав результаты работы алгоритма и модели, мы предложили собственную систему для выполнения этой задачи.

2. Данные

2.1. Русский учебный корпус

В качестве данных при разработке программы используются тексты Русского учебного корпуса (RLC). Русский учебный корпус представляет собой текстовую коллекцию, в которую входят примеры устной и письменной речи двух нестандартных категорий носителей русского языка: к первой относятся те, кто изучает русский как иностранный (L2), ко второй – так называемые «эритажные» носители, то есть естественные билингвы, которые социализировались в языковой среде неродного языка (L2), но используют родной язык (L1) дома как семейный [Valdes, 2000].

RLC включает в свой состав учебные тексты, например, эссе, сочинения, ответы на вопросы, которые могли быть написаны в условиях временного ограничения или же без такового. Соответственно, теги, используемые для разметки ошибок в этом корпусе, разделены на пять групп: орфографические, морфологические, лексические, синтаксические, ошибки в конструкциях. Ошибки словообразования, то есть деривационные ошибки, автоматическое распознавание которых является предметом исследования в настоящей работе, относятся к категории морфологических. В практических целях в настоящей работе мы рассматриваем ошибки формо- и словообразования в корневых морфемах, префиксах и суффиксах как деривационные.

Например:

(1) *Я **познакомаю** с новыми друзьями, **путушествоваю** по миру, и провожу временю на моем личном исскуствe – кулинария.*

2.2. Словообразовательные ошибки

Все размеченные деривационные ошибки были выгружены из корпуса в формате csv-таблицы, где единичным вхождением является случай разметки аннотатора. Таких случаев оказалось 500.

[1] URL: http://www.web-corpora.net/RLC/.

Таблица 1

Пример таблицы, выгруженной из Русского учебного корпуса

Table 1

An example of a table downloaded from the Russian Learner Corpus

Оригинальный текст Original Academic Text	Слово с ошибкой Quote	Норма Correction	Разметка Tags
И читая разные статьи о ресторанах в том городе или о жизни экпатов, я поняла, что по всему миру американцы ишут мексиканскую еду и англичанины ишут индийскую.	англичанины	англичане	Deriv
После осмотрения современного Санкт-Петербурга, я предложила бы туристам пойти именно туда и окунуться в советский Ленинград.	осмотрения	осмотра	Deriv
Люди пошли сначала в поликлинику, нахожденная в районе, где они жили.	нахожденная	находившуюся	Deriv

Валидность разметки, выполненной вручную аннотаторами, была впоследствии проверена также вручную. В окончательную выборку вошли случаи ошибочного формо- или словообразования также содержавшие ошибки других типов, помимо деривационных (см. пример 1).

3. Эксперименты

3.1. Эксперимент на основе статьи «Developing Online ICALL Exercises for Russian»

Одна из немногих работ, посвященных специализированному подходу к обработке морфологических ошибок в русском языке, была написана Маркусом Дикинсоном и Джошуа Херрингом. В работе [Dickinson, Herring, 2008] авторы адаптировали уже существующую ICALL систему автоматической генерации словообразовательных упражнений для изучающих русский язык как иностранный. До этого система генерации аналогичных упражнений была уже разработана для немецкого [Heift, Nicholson, 2001], португальского [2] [Amaral, Detmar, 2006] и японского [Nagata, 1995] языков. Русский язык, обладая богатой морфологией, потребовал от разработчиков создания особого модуля, способного обнаруживать и автоматически исправлять ошибки, которые они отнесли к морфологическим, не дифференцируя их в методах исправления. В основу работы модуля был положен конечный автомат, состоянием которого являлось наличие/отсутствие ошибки, а переходом – буквенные комбинации, составляющие конкретное слово. В общих чертах принцип работы его алгоритма можно описать следующим образом.

Сначала слово разделялось на всевозможные комбинации буквенных *n*-грамм, после чего, проходя по *n*-граммам, алгоритм проверял соответствие *n*-граммы морфеме из морфологического словаря; в качестве словаря использовалась база данных Национального корпуса русского языка для словообразовательного поиска. Если *n*-грамма соответствовала морфеме, то анализировалась следующая *n*-грамма, а также то, что ее комбинация с предыдущей встречается в языке.

[2] Amaral, L., Detmar, M. Putting activity models in the driver's seat: Towards a demand-driven NLP architecture for ICALL. In: Talk given at EUROCALL. University of Ulster, Coleraine Campus, 2007. URL: http://www.sfs.uni-tuebingen.de/~dm/papers/eurocall07-amaral-meurers.pdf.

На первом этапе анализировалась цепочка *n*-грамм либо от начала до конца, либо до того момента, когда *n*-грамма больше не соответствует ни одной из существующих морфем или не встречается в словаре вместе с соседней морфемой. Если не удавалось пройти слово по всей его текстовой длине, то проходил дополнительный анализ цепочек *n*-грамм в обратную сторону, т. е. от конца к началу слова. Такая процедура давала возможность локализовать ошибку: если следующая после места остановки *n*-грамма есть в морфемном словаре, то ошибка считалась морфологической. К сожалению, без собственного эксперимента оценить качество работы алгоритма, описанного в [Dickinson, Herring, 2008], не представляется возможным, так как авторы не предоставили оценки эффективности разработанного ими алгоритма на текстовых данных естественных языков. Кроме того, в статье не была отражена информация о формате данных, что является серьезным упущением.

Отсутствие представленной оценки результатов и желание узнать, насколько успешно система справляется со словообразовательными ошибками побудило нас воссоздать аналогичный алгоритм и провести эксперимент с использованием текстовых данных Русского учебного корпуса.

Тестовая выборка состояла из 170 случайно выбранных слов, содержащих деривационную ошибку, к которым были добавлены 30 правильных слов с тем, чтобы выяснить, насколько корректно алгоритм различает грамматичные и неграмматичные употребления. Воссозданная нами программа Дикинсона и Херринга получала информацию о морфологическом составе слова из словаря, созданного на основе Русского Национального Корпуса. Следующая таблица показывает, в каком виде в словаре представлена морфологическая структура слова.

Таблица 2

Морфологическая структура слова «улечься» в используемом словаре

Table 2

The morphological structure of the word *улечься* (to lie down) in the dictionary

Лемма	Морфема	Тип	Позиция в слове	Алломорфы корня слова	Часть речи
улечься	у	префикс	1		V
улечься	леч	корень	2	[1лож]|[2лег]|[2лёг]|[лаг]|[леж]|[лёж]|[леч]|[лог]	V
улечься	ь	суффикс	3		V
улечься	ся	суффикс	4		V

В качестве примера работы воссозданной (локализованной) программы покажем полный процесс обработки текстоформы «простее», которая представляет собой некорректно образованную сравнительную степень прилагательного. На первом шаге она формирует список из всех возможных *n*-грамм: («простее»), («п», «ростее»), («пр», «остее») … («п», «р», «о», «с», «т», «е», «е»). Затем алгоритм проходит по каждой комбинации от начала до конца и, так как полностью пройти их, как правило, не удается, то на третьем шаге повторяет эту процедуру для каждого из них, но уже в обратном порядке. В результате обе итерации сходятся при разбиении слова на две *n*-граммы – «прост» и «ее». И первая, и вторая *n*-граммы присутствуют в словаре морфем, что и является свидетельством выявления деривационной ошибки.

После обработки тестовых данных были получены следующие результаты работы нашего модуля. Ложноположительных случаев срабатывания алгоритма выявлено не было: другими словами, программа не нашла ошибки ни в одном из 30 дополнительных тестовых морфологически правильных слов. В 138 случаях из общего числа 170 (то есть в 81,18 % случаев) морфологически некорректные слова действительно были идентифицированы как таковые. В 24 случаях из 170 (14,12 %) слова были пройдены системой поморфемно до конца. Из этих

24 в 16 случаях слова были созданы с использованием регулярной модели слово- или формообразования («старшее», «свободности», «работчик», «лежая» и т. д.) и должны были быть выявлены нашей программой, но она не справилась с этой задачей. По нашему предположению, причина этого сбоя (система не обнаруживает ошибку там, где она есть), заключается в том, что в этой версии программы проверке подвергаются только две соседние морфемы. Так, программа посчитала слово «работчик» правильным из-за того, что в словаре встречается слово, в котором за корнем «работ» сразу следует суффикс «чик» (например, «разработчик»). Для оставшихся 8 слов (4,7 %) восстановить производящую основу не представлялось возможным из-за их полной ненормативности. К таким словам относятся, например, «чаровны», «конепит».

Таким образом, результаты эксперимента показали, что локализованный алгоритм Дикинсона и Херринга можно достаточно успешно использовать для выявления деривационных ошибок в речи L2, однако для создания полноценной системы автокоррекции одного его недостаточно.

3.2. Эксперимент с моделью шумного канала

Приступая к разработке модуля автоматического исправления слов, было бы полезно сравнить получаемые результаты с показателями аналогичных систем. Однако в свободном доступе не оказалось образца спеллчекера, созданного специально для текстов, написанных изучающими русский как иностранный. Поэтому в качестве объекта сравнения мы выбрали спеллчекер с моделью от команды DeepPavlov [3].

В 2016 году этот спеллчекер участвовал в соревнованиях по автоматическому исправлению письменных ошибок для русскоязычных текстов из социальных медиа SpellRuEval [Sorokin et al., 2016], где показал лучший результат среди общедоступных спеллчекеров (со значением *F*-меры 56,25). Кроме того, для нас особенно важно, что в качестве данных для модели выступают тексты из сети Интернет. Эти тексты часто содержат случаи нестандартного для русского литературного языка словотворчества, которые в определенном смысле сравнимы с деривационными ошибками в речи изучающих русский как иностранный. См. пример (2), взятый из тестового набора соревнования SpellRuEval:

(2) *Музончик конечно зачетный и почемуто до боли знакомый.*

В основе спеллчекера команды DeepPavlov используется довольно распространенный метод для исправления ошибок в письменном тексте – модель шумного канала, разработанная [Brill and Moore, 2000]. Помимо автокоррекции естественного текста [Kernighan et al., 1990, Church, Gale, 1991], модель шумного канала [Shannon, 1948] применяется во многих других задачах обработки естественного языка, например, при создании вопросно-ответных систем, в распознавании речи и в машинном переводе.

Начиная апробацию модели на материале Русского учебного корпуса, мы ожидали, что качество работы модели DeepPavlov не будет достаточно удовлетворительным, так как, во-первых, модель обучена на текстах носителей русского языка (L1), у которых собственно словообразовательные ошибки практически не встречаются даже в текстах сети Интернет, а во-вторых, она нацелена на исправление слов, которые находятся на небольшом расстоянии Левенштейна: 1–2 символа от правильного слова, в то время как некоторые деривационные ошибки могут распространяться на большее количество символов (ср. пример 1).

Тестовая выборка, которую мы сформировали на основе Русского учебного корпуса, включала 170 предложений, содержащих те же слова с деривационной ошибкой, что были использованы для первого эксперимента, описанного выше. Выборку было решено не расширять ввиду необходимости ручной оценки результатов.

[3] URL: http://docs.deeppavlov.ai/en/master/features/models/spelling_correction.html.

В результате модель оставила 134 слова (78,8 %) без каких-либо изменений, а в 19 случаях (11,2 %) слова с деривационной ошибкой были исправлены на нормативный вариант. Стоит отметить, что во всех этих случаях расстояние Левенштейна до правильного слова не превышало 2. 17 слов (10 %) были исправлены на неправильный вариант. Как уже говорилось выше, мы связываем такое низкое качество работы спеллчекера DeepPavlov с лингвистическими характеристиками материала (датасета), на котором она обучалась, а также с тем, что некоторые слова, содержащие ошибку, очень сильно отличаются от нормативного варианта. Но в целом результаты эксперимента 2 позволили получить конкретные значения для оценки качества работы альтернативной системы автоматического исправления словообразовательных ошибок.

4. Автоматическое исправление деривационных ошибок

При создании собственной системы автоматического исправления словообразовательных ошибок мы решили обратиться к идеям дистрибутивной семантики. Дистрибутивная семантика – область, которая занимается развитием и изучением теории и методов вычисления и категоризации семантической близости лингвистических единиц на основании их распределения в больших массивах лингвистических данных. В ее основе лежит идея о том, что, если слова употребляются в схожих контекстах, они имеют схожую семантику [Harris, 1954]. В дистрибутивной семантике слова обычно представляются в виде векторов в многомерном пространстве их контекстов. Семантическое сходство вычисляется как косинусная близость между векторами двух слов и может принимать значения в промежутке от –1 до 1 (на практике часто используются только значения выше 0). Чем ближе значение косинусного расстояния к 1, тем чаще слова встречаются в схожем контексте и являются более близкими семантически, и наоборот, если значение ближе к 0, то у слов практически нет похожих контекстов.

Поскольку задача состоит в том, чтобы уметь предсказывать слово исходя из его окружения, мы выбрали модель типа Continuous Bag of Words. Кроме того, в нашу задачу входит поиск семантически схожих лексем для тех слов, которые не встретились в обучающей выборке модели или встретились там небольшое количество раз, поэтому мы сочли необходимым обратиться к FastText модели, строящей векторное представление не слова целиком, а всех его *n*-грамм, что позволяет получать векторные представления для малочастотных слов, опираясь на их части (см. [Bojanowski et al., 2017]). В силу того, что обучение дистрибутивных моделей требует очень большого объема данных и, следовательно, вычислительных мощностей, было решено использовать готовую модель. В качестве таковой была выбрана модель проекта RusVectores Continuous Bag of Words FastText, созданная в 2019 году [Kutuzov, Kuzmenko, 2017]; она дает возможность использовать *n*-граммы длиной 3–5 символов. Кроме того, учитывалось, что она была обучена на интернет-корпусе «Тайга» [Shavrina, Shapovalova, 2017], причем его размер на момент обучения составлял почти 5 миллиардов слов. Оригинальная модель [Kutuzov, Kuzmenko, 2017], получая на вход слово с деривационной ошибкой, возвращает список из 10 наиболее близких по значению слов. Никакого дополнительного обучения модели не проводилось.

Таким образом, для получения правильного слова в нашу задачу на текущем этапе входит, во-первых, поиск подходящего слова-кандидата из полученного списка, и во-вторых, приведение его к правильной грамматической форме. Для этого в списке были оставлены только те лексемы, которые находятся на расстоянии Левенштейна, не превышающем половины их длины от корректного варианта. Затем, исходя из некоторых предполагаемых грамматических признаков неправильного слова, полученных с помощью морфологического анализатора Pymorphy2, мы провели фильтрацию. Например, для существительного таким признаком стала одушевленность. После того как из списка были убраны неподходящие кандидаты, выбирался семантически ближайший, и уже на основе полного списка предполагаемых грамма-

тических признаков неправильного слова находилась нужная форма в словаре «Полная парадигма. Морфология. Орфоэпия. Частотность».

Данный словарь, составленный на основе «Полной акцентуированной парадигмы по Зализняку», содержит 181 770 лемм (5 074 140 словоформ) и расширен за счет таких словарей, как «Полный орфографический словарь русского языка» В. В. Лопатина, «Словарь иностранных слов» (Москва: Русский язык, 1988), «Новый толково-словообразовательный словарь русского языка» Т. Ф. Ефремова (2000), «Толковый словарь» под ред. С. И. Ожегова и Н. Ю. Шведовой (Москва: Азъ, 1992) и Викисловарь. После того как подходящая форма слова найдена, она выводится системой как результирующая. Если же на этапе фильтрации результатов не остается ни одного кандидата, то алгоритм сообщает, что в этом слове присутствует деривационная ошибка, требующая ручного исправления. Кроме того, для идентификации некоторых распространенных ошибок, которые представляют непреодолимую сложность для модели, но при этом являются частотными, были написаны особые правила обработки. Такие правила активируются, когда система сталкивается с четко прописанным условием. Например, если система устанавливает, что перед ней слово в просторечной форме «ихний», производная от него форма или очень близкое к этому слово, то она исправляет его на соответствующую ему нормативную форму «их».

(3) *До: Пока турист пытается прочитать это письмо група дитей окружает их и крадёт* ***ихнену*** *сумку или куртку.*

После: Пока турист пытается прочитать это письмо група дитей окружает их и крадёт ***их*** *сумку или куртку.*

Особые правила введены также для словоформ с частицей *не-*. В случаях некорректного образования отрицательных форм приоритет отдается не структурному, а семантическому фактору, так как вероятность использования автором форманта с негативным значением без намерения выразить отрицание, крайне мала. По этой причине отсутствуют варианты исправления, в которых семантика отрицания убирается полностью. Поэтому, например, версия исправления «неможно» на «можно», «нужно», не рассматривается, хотя эти варианты чаще всего считаются моделью как самые близкие.

(4) До: *Всё зависит от людей и в общем* ***неможно*** *установить правила, чтобы его объяснить.*

После: *Всё зависит от людей и в общем* ***нельзя*** *установить правила, чтобы его объяснить.*

Таким образом, разработанная нами система автоматического исправления ошибок в формо- или словообразовании имеет следующую структуру:

1) программе на вход поступает слово;

2) скрипт, реализованный на основе алгоритма Дикинсона и Херринга, определяет присутствие в слове словообразовательной ошибки;

3) при помощи модели Continuous Bag of Words FastText составляется список из 10 семантически близких форм-кандидатов;

4) этот список проходит постобработку для фильтрации слов-кандидатов по грамматическим признакам и частям речи;

5) отобранное слово-кандидат с помощью словаря парадигм приводится к нужной форме;

6) если система не смогла предложить никакого нового варианта, она выводит сообщение о необходимости ручного исправления;

7) некоторые распространенные виды ошибок, неразрешимые для модели, исправляются с помощью правил (см. пример 3).

5. Результаты работы системы автоматического исправления словообразовательных ошибок

Описанная выше система была протестирована на 500 деривационных ошибках из Русского учебного корпуса. Верным мы считали исправление, полностью совпадающее с корректной формой, предложенной разметчиком. Показатель качества работы системы составил 0,616 – она исправила 308 ошибочных написаний из 500. Это примерно в 6,3 раза лучше, чем исправление подобных ошибок специально обученным на интернет-текстах и протестированным нами спеллчекером, в основе которого лежит проверенный метод шумного канала. Все неисправленные случаи были классифицированы.

Наибольшие трудности представляли для системы ошибки в корне слова, то есть собственно словообразовательные (19 случаев). В каждом из этих случаев модель не нашла вариантов исправления (см. примеры ниже) и предложила исправить ошибку вручную.

(5) *Они **искует** пассажера для сваём машину.*
(6) *Это очень счастливая **вспаминяне** иметь.*
(7) *Бабушка делает мне визу в Россию так что думаю и туда **паежу** на пару недель.*

Кроме того, система не смогла корректно исправить ошибки в словах, образованных путем словосложения – объединения корневых элементов двух слов. В этих случаях оказалось невозможным исправить ошибку на эквивалент с одним корнем без изменения значения. Встретилось 17 таких слов.

(8) *Самознание народа и истории происходит от самих людей*
(9) *Но не может "защищать" его **новонайденное** счастье.*
(10) *Этот факт вместе с тем, что население в России испытывает убыль знаменательный того, что России **понадобудет** принять меры чтобы не потерять свой родной язык.*

Третью относительно многочисленную группу сложных случаев составили ошибки, которые объясняются использованием корня вместе с недопустимой для него словообразовательной моделью. Таких случаев было 14.

(11) *Или **устраивание** какого-либо мироприятия, где всем было безумно весело.*
(12) *После этого **пропадения** калитка всегда была заперта.*

Проблемы также возникали в случае замены, удаления или добавления приставки, так как в русском языке они достаточно свободно присоединяются к слову. Из-за этого в 13 случаях были предложены варианты с неподходящей приставкой или совсем без ее исправления.

(13) *До: Они **проездут** по лесу.*
*После: Они **проездят** по лесу.*
(14) *До: Она не хотет **вехать** на машину но волки ее введут.*
*После: Она не хотет **поехать** на машину но волки ее введут.*

Успешнее всего модель исправляла ошибки в суффиксах. Если в неправильном слове не было каких-либо других ошибок, то оно почти всегда получало нормативную форму.

(15) *До: Также, некоторые из них считаются **межфакультетными** кафедрами*
*После: Также, некоторые из них считаются **межфакультетскими** кафедрами*
(16) *До: Лев гоняет волка долго, а волк в мешке упадает и опускается низ **скалинки**, потом лезает на дерево и убегает из дома.*
*После: Лев гоняет волка долго, а волк в мешке упадает и опускается низ **скалы**, потом лезает на дерево и убегает из дома.*

6. Заключение

Суммируем основное. Предложена новая система автоматического исправления слово- и формообразовательных ошибок в письменной речи на L2, обозначаемых обобщенно как деривационные. На первом этапе для обнаружения деривационных ошибок она использует модель на основе конечных автоматов [Dickinson and Herring, 2008], описанную в разделе 3.1, на втором – для их исправления – модель Continuous Bag of Words FastText, дополненную новыми правилами отбора слов-кандидатов и словарем парадигм для их последующего преобразования в нужную текстоформу.

Новый подход обладает двумя принципиальными преимуществами по сравнению с существующими методами автокоррекции морфологии: возможностью исправлять ошибки, находящиеся на большем расстоянии Левенштейна от правильного варианта, и возможностью обучения на текстах носителей языка (L2) без значительного ухудшения качества работы.

Эффективность работы модели подтверждена сравнением с автокорректором, функционирующем на основе модели шумного канала. Этот последний справляется с задачей исправления деривационных ошибок только в 11,2 % случаев, оставляя, таким образом, большинство ошибок без изменений, что говорит о низком уровне его эффективности. У разработанной нами системы эффективность исправления ошибок равна 61,6 %, что для такого типа задач является приемлемым результатом, сопоставимым с уровнем эффективности, достигаемым стандартными спеллчекерами на текстах носителей языка (L1).

Список литературы

Копотев М. Введение в корпусную лингвистику: электрон. учеб. пособие для студентов филологических и лингвистических специальностей университетов. Praha: Animedia, 2014.

Amaral, L., Detmar, M. Where does ICALL Fit into Foreign Language Teaching? In: Talk given at CALICO Conference. University of Hawaii, 2006.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017, vol. 5, p. 135–146.

Brill, E., Moore, R. An Improved Error Model for Noisy Channel Spelling Correction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, p. 286–293.

Chernigovskaya T., Gor K. The Complexity of Paradigm and Input Frequencies in Native and Second Language Verbal Processing: Evidence from Russian. *Language and Language Behavior (Eds. Erling Wande & Tatiana Chernigovskaya)*, 2000, p. 20–37.

Church, K., Gale, W. Probability scoring for spelling correction. *Statistics and Computing*, 1991, vol. 1, p. 93–103

Dickinson, M., Herring, J. Developing Online ICALL Resources for Russian. *The 3rd workshop on innovative use of NLP for building educational applications*, Columbus, OH, 2008, p. 1–9.

Granger, S. From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In: Languages in Contrast. Text-based cross-linguistic studies, Lund University Press, 1996, p. 37–51.

Granger, S. Learner Corpora in Foreign Language Education. In: Language, Education and Technology, 2017, p. 1–14. DOI 10.1007/978-3-319-02328-1_33-1.

Harris, Z. Distributional Structure. *WORD*, 1954, vol. 10, iss. 2–3, p. 146–162. DOI 10.1080/00437956.1954.11659520

Heift, T., Devlan, N. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, 2001, vol. 12 (4), p. 310–325.

Kernighan, M., Church, K., Gale, W. A Spelling Correction Program Based on a Noisy Channel Model. COLING-90, 1990, p. 205–210. DOI 10.3115/997939.997975.

Kutuzov, A., Kuzmenko, E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. *Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science*, 2017, vol. 661. Springer, Cham.

Leacock, C., Chodorow, M., Gamon, M., Tetreau, J. Automated Grammatical Error Detection for Language Learners, 2nd ed. *Synthesis Lectures on Human Language Technologies*. 2014, vol. 7, p. 1–185. DOI 10.2200/S00562ED1V01Y201401HLT025

Nagata, N. An Effective Application of Natural Language. *Processing in Second Language Instruction*. CALICO Journal, 1995.

Paquot, M., Jarvis, S. Learner corpora and native language identification, 2015. DOI 10.1017/CBO9781139649414.027.

Rudzewitz, B., Ziai, R., De Kuthy, K., Möller, V., Nuxoll, F., Detmar, M. Generating Feedback for English Foreign Language Exercises. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2018, p. 127–136.

Shannon, C. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948, vol. 27, p. 379–423.

Shavrina, T., Shapovalova, O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. *Proceedings of international conference CORPORA2017*, 2017, p. 78–84.

Sorokin, A., **Baytin, A., Galinskaya, I., Rykunova, E., Shavrina, T.** SpellRuEval: the First Competition on Automatic Spelling Correction for Russian. *Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue"*, 2016, p. 660–673.

Valdes, G. The teaching of heritage languages: an introduction for Slavic-teaching professionals. *Slavica, Bloomington*, 2000, p. 375–403.

References

Amaral, L., Detmar, M. Where does ICALL Fit into Foreign Language Teaching? In: Talk given at CALICO Conference. University of Hawaii, 2006.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017, vol. 5, p. 135–146.

Brill, E., Moore, R. An Improved Error Model for Noisy Channel Spelling Correction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, p. 286–293.

Chernigovskaya T., Gor K. The Complexity of Paradigm and Input Frequencies in Native and Second Language Verbal Processing: Evidence from Russian. *Language and Language Behavior (Eds. Erling Wande & Tatiana Chernigovskaya)*, 2000, p. 20–37.

Church, K., Gale, W. Probability scoring for spelling correction. *Statistics and Computing*, 1991, vol. 1, p. 93–103

Dickinson, M., Herring, J. Developing Online ICALL Resources for Russian. *The 3rd workshop on innovative use of NLP for building educational applications*, Columbus, OH, 2008, p. 1–9.

Granger, S. From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In: Languages in Contrast. Text-based cross-linguistic studies, Lund University Press, 1996, p. 37–51.

Granger, S. Learner Corpora in Foreign Language Education. In: Language, Education and Technology, 2017, p. 1–14. DOI 10.1007/978-3-319-02328-1_33-1.

Harris, Z. Distributional Structure. *WORD*, 1954, vol. 10, iss. 2–3, p. 146–162. DOI 10.1080/00437956.1954.11659520

Heift, T., Devlan, N. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, 2001, vol. 12 (4), p. 310–325.

Kernighan, M., Church, K., Gale, W. A Spelling Correction Program Based on a Noisy Channel Model. COLING-90, 1990, p. 205–210. DOI 10.3115/997939.997975.

Kopotev, M. Introduction to Corpus linguistics: Course-book for students of arts subjects with emphasis on the Russian language. Praha, Animedia, 2014. (in Russ.)

Kutuzov, A., Kuzmenko, E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. *Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science*, 2017, vol. 661. Springer, Cham.

Leacock, C., Chodorow, M., Gamon, M., Tetreau, J. Automated Grammatical Error Detection for Language Learners, 2nd ed. *Synthesis Lectures on Human Language Technologies*. 2014, vol. 7, p. 1–185. DOI 10.2200/S00562ED1V01Y201401HLT025

Nagata, N. An Effective Application of Natural Language. *Processing in Second Language Instruction*. CALICO Journal, 1995.

Paquot, M., Jarvis, S. Learner corpora and native language identification, 2015. DOI 10.1017/CBO9781139649414.027.

Rudzewitz, B., Ziai, R., De Kuthy, K., Möller, V., Nuxoll, F., Detmar, M. Generating Feedback for English Foreign Language Exercises. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2018, p. 127–136.

Shannon, C. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948, vol. 27, p. 379–423.

Shavrina, T., Shapovalova, O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. *Proceedings of international conference CORPORA2017*, 2017, p. 78–84.

Sorokin, A., **Baytin, A., Galinskaya, I., Rykunova, E., Shavrina, T.** SpellRuEval: the First Competition on Automatic Spelling Correction for Russian. *Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue"*, 2016, p. 660–673.

Valdes, G. The teaching of heritage languages: an introduction for Slavic-teaching professionals. *Slavica, Bloomington*, 2000, p. 375–403.

Материал поступил в редколлегию
Date of submission
29.03.2021

Сведения об авторах / Information about the Authors

Выренкова Анастасия Сергеевна, кандидат филологических наук, доцент школы лингвистики, факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва, Россия)

Anastasia S. Vyrenkova, PhD in Applied and Mathematical Linguistics, associate professor of School of Linguistics, Faculty of Humanities, HSE University (Moscow, Russian Federation)

anastasia.marushkina@gmail.com
ORCID 0000-0003-1707-7525

Смирнов Иван Юрьевич, аспирант Школы лингвистики, факультет гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва, Россия)

Ivan Yu. Smirnov, PhD student of School of Linguistics, Faculty of Humanities, HSE University (Moscow, Russian Federation)

smirnof.van@gmail.com
ORCID 0000-0001-8361-0282