АВТОМАТИЗАЦИЯ, МАШИННОЕ ОБУЧЕНИЕ И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ В ЛИНГВИСТИКЕ, ПЕРЕВОДЕ И ПЕДАГОГИКЕ
AUTOMATION, MACHINE LEARNING AND INTELLIGENT SYSTEMS IN LINGUISTICS, TRANSLATION AND PEDAGOGY

Вестник Челябинского государственного университета. 2024. № 8 (490). С. 220–229.
Bulletin of Chelyabinsk State University. 2024;(8(490):220-229.

Научная статья

УДК 81`322.2

doi: 10.47475/1994-2796-2024-490-8-220-229

МАШИННОЕ ОБУЧЕНИЕ ПРИ АДАПТАЦИИ УЧЕБНЫХ ТЕКСТОВ: ЛЕКСИЧЕСКИЙ АСПЕКТ

Ольга Юрьевна Редькина[1]✉, Татьяна Васильевна Карпета[2], Дарья Сергеевна Пономарева[3], Дарья Александровна Вершинина[4]

[1] Челябинский государственный университет, Челябинск, Россия, filolchen@gmail.com, 0000-0002-5012-0866

[2] Южно-Уральский государственный университет, Челябинск, Россия, eroshkinatv@susu.ru, 0000-0001-5079-5970

[3] Челябинский государственный университет, Челябинск, Россия, ponomarevadasha18@yandex.ru, 0009-0002-7941-753X

[4] Челябинский государственный университет, Челябинск, Россия, vershinina.0420@mail.ru, 0009-0000-2959-9262

Аннотация. Обучение в школах инофонов в смешанных классах требует адаптации учебных материалов, особенно по дисциплинам гуманитарного цикла. Процесс адаптации отличается большой трудоёмкостью, однако с учётом того, что адаптация осуществляется под определённый уровень владения русским языком как иностранным, а значит, предполагает соответствие чётким требованиям, может быть отчасти автоматизирован. В статье представлены результаты апробации методов машинного обучения по адаптации учебных текстов на лексическом уровне. Материалом адаптации стал учебник по истории России за 11 класс, в котором были произведены синонимические замены ряда лексем на слова, включённые в лексический минимум уровня А1. Для осуществления замен был создан словарь синонимов, в котором заголовочными словами являются слова из лексического минимума уровня А1; отбор синонимов осуществлялся с учётом их частеречной принадлежности из словарей синонимов, размещённых на сайте https://academic.ru/. Были проведены токенизация текста, морфологический анализ слов, приведение их в начальную форму, замена на синонимы из лексического минимума и дальнейшее приведение синонимов в нужную грамматическую форму. Дальнейшая работа предполагала проверку адекватности произведённых замен и анализ ошибок. Алгоритм показал высокую степень точности при осуществлении замен; допущенные ошибки проанализированы и классифицированы. Выделены пять типов ошибок: 1) ошибки, связанные с семантикой многозначного слова, 2) ошибки, связанные с системными отношениями в лексике, 3) ошибки, связанные со стилистическими особенностями словоупотребления, 4) ошибки, связанные с употреблением устойчивых / лексикализованных сочетаний, 5) ошибки, связанные с культурой оформления печатного текста. Основными причинами возникновения этих ошибок стали: 1) неразличение лексико-семантических вариантов многозначного слова или неразличение омонимов, 2) неразличение стилистически окрашенных и нейтральных синонимов, 3) неузнавание имён собственных (географических наименований, имен и фамилий людей), 4) неузнавание условных сокращений, принятых в специализированных изданиях. Предложены шаги по усовершенствованию работы алгоритма: анализ многозначных слов и омонимов, входящих в минимум, для уточнения их значений; введение критерия вероятности использования слова как стилистически окрашенной единицы расширение словаря синонимов за счёт включения в него лексикализованных сочетаний, имён собственных, наиболее частотных условных сокращений.

Ключевые слова: машинное обучение, парсинг сайтов, парсер на языке Python, русский язык как иностранный, лексический минимум, учебный текст, адаптация

Финансирование. Статья подготовлена при финансовой поддержке Фонда перспективных научных исследований ФГБОУ ВО «Челябинский государственный университет» (приказ № 126-1 от 14.03.2024).

Для цитирования: Редькина О. Ю., Карпета Т. В., Пономарева Д. С., Вершинина Д. А. Машинное обучение при адаптации учебных текстов: лексический аспект // Вестник Челябинского государственного университета. 2024. № 8 (490). С. 220–229. DOI: 10.47475/1994-2796-2024-490-8-220-229.

Original article

MACHINE LEARNING IN ADAPTATION OF EDUCATIONAL TEXTS: LEXICAL ASPECT

Olga Yu. Redkina[1]✉, Tatiana V. Karpeta[2], Daria S. Ponomareva[3], Daria A. Vershinina[4]

[1]Chelyabinsk State University, Chelyabinsk, Russia, filolchen@gmail.com, 0000-0002-5012-0866

[2]South Ural State University, Chelyabinsk, Russia, eroshkinatv@susu.ru, 0000-0001-5079-5970

[3]Chelyabinsk State University, Chelyabinsk, Russia, ponomarevadasha18@yandex.ru, 0009-0002-7941-753X

[4]Chelyabinsk State University, Chelyabinsk, Russia, vershinina.0420@mail.ru, 0009-0000-2959-9262

Abstract. This paper presents the results of approbation of machine learning methods for adapting educational texts for the purpose of teaching non-native learners in Russian. The testing was carried out on the text of the textbook on the history of Russia for the 11th grade, in which synonymic substitutions of a number of lexemes were made for words included in the lexical minimum of the A1 level. The preparatory stage of adaptation is described and the classification of errors made in the text as a result of adaptation using machine learning methods is presented.

Keywords: machine learning, website parsing, Python parser, Russian as a foreign language, lexical minimum, educational text, adaptation

Funding. The article was prepared with the financial support of the Fund for Advanced Scientific Research of Chelyabinsk State University (Order № 126-1 of 14.03.2024).

For citation: Redkina OYu, Karpeta TV, Ponomareva DS, Vershinina DA. Machine learning in the adaptation of educational texts: lexical aspect. *Bulletin of Chelyabinsk State University.* 2024;(8(490):220-229. (In Russ). DOI: 10.47475/1994-2796-2024-490-8-220-229.

Введение

Преподавание иностранных языков связано с использованием на занятиях большого количества текстов, которые условно можно разделить на три разновидности: специально созданные тексты (как правило, используемые на начальном этапе подготовки), адаптированные тексты (используемые на базовом и пороговом уровнях языковой подготовки) и аутентичные (используемые на продвинутом этапе обучения) [1]. Вторую и третью разновидности составляют тексты, изначально не предназначенные для использования в дидактических целях, которые либо подвергаются определённой трансформации (адаптируются), либо остаются в первозданном виде. Под адаптацией будем понимать вслед за Э. Г. Азимовым и А. Н. Щукиным процесс «упрощения, приспособления, облегчения или усложнения текста в соответствии с уровнем языковой компетенции учащихся» [1]. Предмет нашего исследовательского интереса составляют адаптированные тексты, поскольку их подготовка требует комплексного анализа аутентичного текста (его содержания, лексических и грамматических трудностей) и дальнейшей кропотливой работы по изъятию или замене отдельных единиц без утраты или искажения смысла текста [3; 9].

В современной образовательной среде особую актуальность вопрос адаптации учебных текстов приобретает не столько в преподавании собственно иностранного языка, сколько в преподавании на нём, т. е. при обучении инофонов. Стоит отметить, что существенные затруднения у школьников-инофонов вызывает освоение дисциплин гуманитарного цикла: литературы, истории и обществознания. Учебные и контрольные материалы по этим дисциплинам, как правило, представлены объёмными текстами, неоднородными с точки зрения стилистики, включающими разнообразную лексику и грамматические конструкции и характеризующимися (при наличии) имплицитной авторской оценкой. Кроме того, для адекватного восприятия этих текстов от адресата требуется наличие фоновых знаний, касающихся культуры и истории страны проживания.

Потребность в адаптации учебных текстов для инофонов совершенно очевидна, и, на наш взгляд, она может быть осуществлена путём использования инструментов автоматизированного комплексного анализа и адаптации текстов с учётом фактора адресата-инофона. Целью нашего междисциплинарного прикладного исследования является разработка таких инструментов, а в настоящей статье представлены результаты первого этапа, заключающегося в апробации методов машинного обучения применительно к адаптации учебного текста на лексическом уровне.

На сегодняшний день уже разработан сервис, позволяющий с достаточно высокой степенью точности определить уровень сложности текста, — онлайн-инструмент «Текстометр» (https://textometr.ru/). «Уровень сложности текста по шкале CEFR (от A1 до C2) определяется автоматически на основании результатов работы математической модели, обученной на коллекции из 800 текстов из пособий по РКИ, информация об уровне сложности которых нам уже известна, и более чем 100 лингвистических признаков текста» [6]. Ресурса же, позволяющего произвести не только анализ, но и адаптацию учебного текста, на сегодняшний день не существует.

Материалы и методы

Материалом адаптации послужила первая часть учебника «История. История России. 1946 г. – начало XXI в.» за 11 класс, изданного под редакцией А. В. Торкунова[1]. Текст учебника выдержан в научном стиле, содержит значительное количество терминологической лексики, а кроме того, включает ряд фрагментов, написанных в публицистическом стиле (отрывки из воспоминаний, открытых писем, краткие биографии научных, культурных, политических деятелей и т. п.). Задачей данного этапа исследования стала апробация механизмов адаптации учебного текста на лексическом уровне.

Адаптация текста на лексическом уровне предполагает выявление лексики, не соответствующей уровню языковой подготовки адресата, и либо её замену на адекватный синоним из числа изучаемых на данном уровне слов, либо изъятие из текста, либо вынесение в лексико-грамматический комментарий. Поскольку в нашем случае речь идет об автоматизированном процессе, мы ограничились заменой ряда лексем на имеющиеся в минимуме уровня A1 [5].

С точки зрения лингводидактики такая работа не имеет смысла, поскольку, во-первых, исходный текст уровню A1 не соответствует ни по тематике, ни по уровню лексической и грамматической сложности [2], во-вторых, замене могла подвергнуться далеко не вся лексика из текста. Однако смысл апробации сводится не к созданию адаптированного текста, а к проверке возможностей инструмента адаптации, выявлению ошибок для последующей их коррекции. Уровень A1 как нельзя лучше подходит для этих целей, поскольку лексический минимум включает всего 780 слов, следовательно, лучше поддается систематизации и алгоритмизации.

Для решения поставленной задачи мы на основе лексического минимума уровня A1 создали словарь синонимов. Заголовочными словами словарных статей стали слова из минимума, подбор синонимов осуществлялся методом сплошной выборки из пяти словарей синонимов, размещённых на сайте https://academic.ru/[2]. Варианты синонимической замены были отфильтрованы с учётом их частеречной принадлежности.

Для создания словаря синонимов использовался язык программирования Python и библиотека BeautifulSoup. Алгоритм использовал BeautifulSoup для извлечения синонимов из HTML-кода страницы. Полученные синонимы были занесены в словарь.

Учебник «История. История России. 1946 г. – начало XXI в.» был представлен в виде файла PDF. При парсинге PDF-файлов с использованием библиотеки PyMuPDF (Fitz) в Python можно извлечь текст, учитывая различные стили шрифтов, такие как курсив и жирный. Библиотека Fitz позволила получить информацию о каждом фрагменте текста на странице PDF, включая его стиль. В результате работы алгоритма учебник был представлен в формате текстового файла со специальными метками.

Дальнейшая работа с текстом осуществлялась по следующему алгоритму:

1. Токенизация текста и обработка слов. Текст необходимо разделить на отдельные слова (токены) для дальнейшей обработки (рис. 1).

2. Морфологический анализ слов (рис. 2). Проход по каждому слову из текста и получение его нормальной формы (леммы) [10], грамматиче-

[1] История. История России. 1946 г. − начало XXI в. 11 класс. Базовый уровень. В 2 ч. Ч. 1 / под ред. А. В. Торкунова. М., 2020. 112 с.

[2] Словарь синонимов // Академик. URL: https://dic.academic.ru/contents.nsf/dic_synonims/ (дата обращения 11.05.2024).

ских характеристик и другой информации с помощью библиотеки PyMorphy2 — морфологический анализатор для русского и украинского языков [12], разработанный на основе словарей OpenCorpora [14], обеспечивающий достаточно высокую степень точности в анализе словоформы — до 96,43 % [11] и потому часто используемый при автоматической обработке текстов на русском языке [15].

3. Замена слов на слова из списка А1 (использовался метод случайного выбора слова-синонима из подходящих вариантов) (рис. 3).

4. Применение тегов для приведения слов в нужную форму. Грамматические характеристики слова (теги) использовались для приведения его к нужной форме — той, в которой слово стояло в исходном варианте текста (рис. 3).

Примеры реализации пп. 3–4 могут быть представлены в виде последовательности преобразований, в которой (1) — форма исходного слова в тексте, (2) — начальная форма исходного слова, (3) — начальная форма синонима, (4) — словоформа синонима (=1) (рис. 3).

Важно отметить, что были заменены лишь те слова, у которых есть синоним в лексическом минимуме уровня А1, то есть текст не подвергся существенной трансформации. Кроме того, изменение формы слова произошло только у синонимов (начальная форма поменялась на исходную), а контекст употребления этих языковых единиц не менялся. Так, во фрагменте предложения: *национальный ВОПРОС и национальная ПОЛИТИКА в послевоенном СССР* (здесь и далее прописными буквами выделены слова, подвергшиеся замене; во всём остальном тексте, включая имена собственные и начало предложения, используются строчные буквы. — ***Прим. авт.***) — синонимической замене подверглись слова *вопрос*, *политика* и *СССР*. В получившемся фрагменте видим: *национальный ПРОБЛЕМА и национальная КУРС в послевоенном СТРАНЫ*. Согласованные определения с заменой определяемого слова на синоним с другими грамматическими характеристиками форму не поменяли, и, поскольку грамматическая адаптация в наши задачи не входила, этот факт мы как ошибку рассматривать не будем. Наибольший интерес для нас представляют ошибки, связанные с лексическим уровнем текста.

Результаты исследования и их обсуждение

Исходный текст состоит из 27704 слов; общее количество замен — 6431 словоформа (формы слова учтены как отдельные замены, 23 %), уникальных замен — 1164 лексемы (без учёта формы слова, в том числе особых форм глагола, кратких форм причастий и прилагательных, 4 %).

Проанализировав полученный текст, мы смогли классифицировать ошибки: 1) ошибки, связанные с семантикой многозначного слова; 2) ошибки, связанные с системными отношениями

```
'Вопросы', 'и', 'задания', 'внутри', 'параграфов', 'позволят',
```

Рис. 1. Токенизация текста
Fig. 1. Text tokenization

```
Parse(word='вопросы', tag=OpencorporaTag('NOUN,inan,masc plur,accs'),
normal_form='вопрос', score=0.64609, methods_stack=((DictionaryAnalyzer(), 'вопросы', 34, 9),))
Parse(word='и', tag=OpencorporaTag('CONJ'),
normal_form='и', score=0.998263, methods_stack=((DictionaryAnalyzer(), 'и', 20, 0),))
Parse(word='задания', tag=OpencorporaTag('NOUN,inan,neut sing,gent'),
normal_form='задание', score=0.592592, methods_stack=((DictionaryAnalyzer(), 'задания', 77, 2),))
Parse(word='внутри', tag=OpencorporaTag('PREP'),
normal_form='внутри', score=0.817647, methods_stack=((DictionaryAnalyzer(), 'внутри', 24, 0),))
Parse(word='параграфов', tag=OpencorporaTag('NOUN,inan,masc plur,gent'),
normal_form='параграф', score=1.0, methods_stack=((DictionaryAnalyzer(), 'параграфов', 34, 7),))
Parse(word='позволят', tag=OpencorporaTag('VERB,perf,tran plur,3per,futr,indc'),
normal_form='позволить', score=1.0, methods_stack=((DictionaryAnalyzer(), 'позволят', 657, 10),))
```

Рис. 2. Морфологический анализ слов
Fig. 2. Morphological analysis of words

(1) вопросы – (2) вопрос – (3) задача – (4) задачи
(1) текстом – (2) текст – (3) документ – (4) документом

Рис. 3. Замена слов и их приведение в нужную форму
Fig. 3. Replacing words and bringing them into the desired form

в лексике; 3) ошибки, связанные со стилистическими особенностями словоупотребления; 4) ошибки, связанные с употреблением устойчивых / лексикализованных сочетаний; 5) ошибки, связанные с культурой оформления печатного текста. Рассмотрим их более подробно. Для верификации данных мы использовали «Современный толковый словарь русского языка» Т. Ф. Ефремовой[1] (далее — словарь; иные словари указываются с автором или жанровой принадлежностью. — ***Прим. авт.***).

I. Ошибки, связанные с семантикой многозначного слова

Этот наиболее распространенный тип ошибок связан с подбором синонима к лексико-семантическому варианту (далее — ЛСВ) многозначного слова, неуместному в данном контексте.

Пример 1: *социально-экономическое развитие страны в 1960-х – СЕРЕДИНЕ 1980-х гг.* заменено на *социально-экономическое развитие страны в 1960-х – ЦЕНТРЕ 1980-х гг.* Слово *середина* в словаре представлено как многозначное. ЛСВ имеют значения: «1) место, более или менее одинаково удалённое от краев, концов чего-либо; 2) время, более или менее одинаково удалённое от начала и конца чего-либо». В предложенном контексте актуализируется второй ЛСВ, в то время как автоматическая замена приводит к актуализации первого ЛСВ.

Пример 2: *ПОЛИТИКА МИРНОГО сосуществования в 1950-х – первой ПОЛОВИНЕ 1960- х гг.* заменено на *КУРС СПОКОЙНОГО сосуществования в 1950-х – первой ЖЕНЕ 1960-х гг.* В представленном фрагменте произведены три замены, причём если первые две (*политика — курс, мирного — спокойного*) достаточно адекватны, хоть и не вполне уместны в данном контексте, то последняя представляет собой замену ЛСВ многозначного слова в переносном значении, ограниченного в сфере употребления: половина — «перен. разг. муж по отношению к жене или жена по отношению к мужу».

Пример 3: *ОГРОМНОЕ ЗНАЧЕНИЕ при изучении ИСТОРИИ имеет РАБОТА с КАРТАМИ* заменено на *ВЕЛИКОЕ ЦЕНУ при изучении РАССКАЗА имеет УПРАЖНЕНИЕ с ОТКРЫТКАМИ.* Синонимическая замена *огромное — великое* представляет собой не столько ошибку, сколько речевой и стилистический недочёт, связанный с использованием слова с эмоционально-экспрессивной коннотацией и разрушением стереотипного сочетания, принятого в научном стиле. Слово *значение* соотносится, скорее, со словом *ценность*, а не *цена*: «II. 1. Важность, значительность», однако словарь синонимов ASIS В. Н. Тришина[2] приводит среди прочих синонимов и *ценность*, и *цена*. Замена *работа — упражнение* представляется адекватной за исключением грамматической формы. Слово *работа*, использованное здесь в значении «I. 2. занятие», является абстрактным существительным и потому не имеет формы множественного числа. Слово *упражнение* обладает формой множественного числа и в данном контексте должно употребляться в ней. Поскольку алгоритм менял форму синонима на ту же, в которой стояло заменяемое слово, автоматически, подобного рода недочёты необходимо исправлять вручную. Замены *история — рассказ, карта — открытка* произведены вследствие неразличения омонимов; этот тип ошибок описан в соответствующем разделе.

Пример 4: *было сокращено до минимума ЧИСЛО планируемых показателей* заменено на *было сокращено до минимума ДЕНЬ планируемых показателей.* Замена осуществлена в связи с неразличением оттенков смысла ЛСВ многозначного слова. Слово *число* в данном контексте употреблено в значении «I. 1. Понятие количества», заменено на «I. 2. День месяца в порядковом ряду других дней».

II. Ошибки, связанные с системными отношениями в лексике

Ошибки этого типа нарушают парадигматические связи [6] между лексемами и потому могут быть подразделены на подвиды.

1. Ошибки, связанные с омонимией

Ошибки этого вида возникают из-за неразличения омонимов. Как правило, эти ошибки приводят к ошибкам других видов и типов: замене гипонима гиперонимом, использованию стилистически окрашенного синонима и т. д.

Пример 5: *более 11 млн военнослужащих, <...> орудий составляли ГРОЗНУЮ СИЛУ* заменено на *более 11 млн военнослужащих, <...> орудий составляли ГОРОД ЦЕНУ.* На замену *грозная — город* существенное влияние оказала грамматическая трансформация, предшествующая собственно лексической замене слова. Так, прилагательное стоит в форме ж. р. В. п. ед. ч. Начальная форма прилагательного, заложенная

[1] Ефремова Т. Ф. Новый словарь русского языка. Толково-словообразовательный: Св. 136000 словарных статей, около 250000 семантических единиц : В 2 т. М., 2000.

[2] Словарь синонимов // Академик. URL: https://dic.academic.ru/contents.nsf/dic_synonims/ (дата обращения 11.05.2024).

в библиотеку PyMorphy2, — м. р. Им. п. ед. ч., т. е. *грозный* — омоним к ойкониму *Грозный*. Дальнейшая замена приводит к другому виду ошибок, связанному с нарушением парадигматических связей, как и замена *тысяча — игра*. Эти ошибки рассмотрены в соответствующем разделе.

Пример 6: *выросла заработная ПЛАТА* заменено на *выросла заработная МАМА.* Замена *плата — мама* обусловлена наличием у терминологического словосочетания *материнская плата* жаргонного синонима — *мама*. В словаре это значение слова плата отсутствует. Ограниченный контекст употребления слова мама в этом значении указывает на принадлежность замены к ещё одному типу ошибок — ошибкам, связанным со стилистическими особенностями словоупотребления.

Пример 7: *написание СЦЕНАРИЕВ, РОМАНОВ и пьес* заменено на *написание ПЛАНОВ, ИМЁН и пьес.* Слово *роман* в словаре представлено как два омонима; имя собственное в словаре отсутствует, но оно является омоформой по отношению к любому из двух омонимов (не совпадают формы В. п. ед. и мн. ч.: *Романа — роман*, *Романов — романы*). В контексте предложения слово роман стоит в форме Р. п. мн. ч., совпадающей у всех омонимов, воспринято алгоритмом как имя собственное и заменено на гипероним *имя*. Существенное количество ошибок, допущенных алгоритмом при работе с именами собственными, соотносится с данными, представленными в статье *A Close Look at Russian Morphological Parsers: Which One Is the Best?*, согласно которой, с именами собственными связано до 50 % ошибок в работе парсеров [13]. Замена *сценарий — план* относится к первому типу ошибок.

Пример 8: *«доживём до понедельника», 1969, РЕЖ. с. и. ростоцкий* заменено на *«доживём до понедельника», 1969, ГОРОД. с. и. ростоцкий.* Замена *реж. — город* может быть отнесена к двум типам ошибок: связанным с омонимией и с культурой оформления печатного текста. Использование общепринятых сокращений слов, значение которых легко восстанавливается по контексту, до узнаваемого минимума, является характерной чертой научных и научно-учебных изданий [7]. Алгоритм воспринял графическое сокращение с точкой как лексему в конце предложения, возникла омонимия (Реж — город в Свердловской области), и произошла замена гипонима на гипероним.

2. Ошибки, связанные с синонимией

Ошибки, связанные с синонимией, представляют собой замену слова из текста на слово из лексического минимума не в том значении, в котором оно представлено в минимуме, т. е. либо неподходящего ЛСВ многозначного слова, либо омонима.

Пример 9: *устанавливалось НАКАЗАНИЕ на срок до 20 лет* заменено на *устанавливалось СТАТЬЮ на срок до 20 лет.* В словаре представлено пять значений слова *статья*. На уровне А1 изучается только первое: «1. Публицистическое, научное или научно-популярное сочинение небольшого размера»; в контексте предложения актуализируется пятое: «5. Наказание на основании закона».

Пример 10: *поверхности ЛУНЫ коснулся* заменено на *поверхности МЕСЯЦА коснулся.* Слово *месяц* в словаре представлено как два омонима: «I. 1. Одна двенадцатая часть астрономического года, имеющая каждая самостоятельное название» и «II. Светящийся диск или часть диска ближайшего естественного спутника Земли, видимый на небе ночью; луна». На уровне А1 слово *месяц* изучается только в первом значении.

Пример 11: *ФИЛЬМЫ а. а. тарковского <...> выходили с цензурными купюрами* заменено на *КАРТИНЫ а. а. тарковского <...> выходили с цензурными купюрами.* Слово *картина* в словаре представлено как многозначное. На уровне А1 изучается только первое значение: «1. Произведение живописи в красках»; в контексте актуализируется «2. Кинематографический или телевизионный фильм».

3. Ошибки, связанные с гипо-гиперонимическими отношениями

Ошибки этого вида представляют собой либо замену гипонимов гиперонимами (в роли гипонимов зачастую выступают имена собственные), либо гиперонимов гипонимами.

Пример 12: *СССР* и названия иных стран (*США, Великобритания, Япония, Германия* и пр.) по всему тексту заменены на гипероним *СТРАНА*.

Пример 13: *волжский АВТОЗАВОД* заменено на *волжский ЗАВОД*. Сложносокращенные слова на уровне А1 не изучаются, поэтому произошла замена гипонима *автозавод* на гипероним *завод.*

Пример 14: *в УЧЕБНИК включено два ВИДА ТЕКСТОВ* заменено на *в КНИГУ включено два ГРУППЫ СТИХОВ.* Слова *учебник* и *текст* в замене не нуждались, поскольку они включены в лексический минимум, однако произошедшая замена является весьма репрезентативной. Слово *учебник* заменено на гипероним *книга*, *текст* — на гипоним *стихи*. Замена *вид — группа* представляет собой ошибку, связанную с подбором

синонима из числа неизученных значений омонимов или ЛСВ многозначного слова: на уровне А1 слово группа изучается в значении «I. 2. Совокупность лиц, объединенных общей профессией, какой-либо деятельностью или общностью интересов, взглядов», иначе говоря, в значении «академическая группа».

Пример 15: *большегрузный 180-тонный АВТОМОБИЛЬ белаз* заменено на *большегрузный 180-тонный КОПЕЙКА белаз.* Замена *автомобиль — копейка* — еще один пример комплексной ошибки, связанной с неразличением омонимов, заменой гиперонима гипонимом и стилистической окрашенностью синонима. На уровне А1 слово *копейка* изучается в значении «I. 1. Монета достоинством в одну сотую рубля»; алгоритм же видит в слове омоним со значением «II. *разг*. Легковой автомобиль модели ВАЗ-2101 марки «Жигули», выпускаемый Волжским автозаводом».

III. Ошибки, связанные со стилистическими особенностями словоупотребления

Этот тип ошибок — зачастую весьма курьезных — представляет собой замены общеупотребительных слов или терминов с нейтральной стилевой окраской на слова с разговорной, просторечной окраской, жаргонные слова, а также слова с эмоционально-экспрессивной коннотацией и эвфемизмы.

Пример 16: *КРАСНАЯ АРМИЯ* заменено на *КРАСИВАЯ МОРЕ* — комплексная ошибка, связанная в том числе с употреблением составного наименования. Слово *красный* в словаре имеет пометку: «Употребляется как постоянный эпитет; красивый, прекрасный, ясный, светлый». Слово *армия* представлено как два омонима, второй — в переносном значении: «II. *разг*. Большое количество людей, объединенных каким-либо общим признаком», допускающем формирование отношений контекстуальной синонимии со словом море.

Пример 17: *рассчитывая на ЗАХВАТ ВЛАСТИ* заменено на *рассчитывая на КОШКУ ПРАВИТЕЛЬСТВА*. Слово *кошка* употребляется как жаргонизм с общим значением способа захвата чего-либо. В словаре Т. Ф. Ефремовой это значение не представлено, в словаре С. И. Ожегова их два: «4. *мн*. Род железных шипов (или иных приспособлений), надеваемых на обувь для лазанья на столбы, по отвесным склонам. 5. Небольшой якорь (спец.)»[1].

Пример 18: *xx век <...> стал наиболее СЛОЖНЫМ* заменено на *xx век <...> стал наиболее КИТАЙСКИМ.* Ещё один пример комплексной ошибки, связанной со стилистическими особенностями употребления слова, а также с подбором синонима из числа неизученных значений омонимов или ЛСВ многозначного слова. *Сложный* в словаре обладает рядом значений, в том числе «1. *перен*. Характеризующийся многими переплетающимися явлениями, признаками, отношениями. 2. *перен*. Представляющий трудность для понимания, разрешения, осуществления; трудный». Очевидна связь последнего с фразеологизмом с похожим значением *китайская грамота*. Слово *китайский* на уровне А1 изучается, но как относительное прилагательное, со словами *язык*, *блюдо* и так далее, вне фразеологической единицы и перехода в разряд качественно-относительных прилагательных.

IV. Ошибки, связанные с употреблением устойчивых / лексикализованных сочетаний

Ошибки, отнесённые нами к этому типу, представляют собой синонимические замены слов, приводящие к разрушению производных предлогов, фразеологических единиц, терминологических словосочетаний и иных единиц, являющихся неделимым единством.

Пример 19: *с одной СТОРОНЫ* заменено на *с одной РУБАШКИ.* Синонимическая замена слова *сторона* в значении «5. Одна из поверхностей, один из боков чего-либо» на *рубашка* в значении «II. 1. Оборотная сторона игральных карт» привела к разрушению вводной конструкции.

Пример 20: *в СВЯЗИ с <...> обострением* заменено на *в ЛЮБВИ с <...> обострением*. Синонимическая замена слова *связь* в значении «2. Общение с кем-либо. Любовные отношения, сожительство» привела к разрушению производного предлога *в связи с*.

Пример 21: *имело ВЫСШЕЕ*, *среднее и незаконченное среднее ОБРАЗОВАНИЕ* заменено на *имело БОЛЬШОЕ*, *среднее и незаконченное среднее РОЖДЕНИЕ.* Анализ слов *высший* и *образование* вне контекста их употребления и синонимическая замена (адекватная для другого контекста) приводит к разрушению терминологических словосочетаний *высшее образование*, *среднее образование*, *незаконченное среднее образование.*

V. Ошибки, связанные с культурой оформления печатного текста

Этот тип ошибок представляет собой замену однобуквенных графических сокращений с точкой в обозначении века, года или инициалов

[1] Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка : 72500 слов и 7500 фразеологических выражений. 2-е изд., испр. и доп. М., 1994. 907 с.

на слово *буква*. Менее распространенными являются ошибки по типу представленном в примере 8.

Пример 22: *в 1993 Г.* заменено на *в 1993 БУКВЫ.*

Пример 23: *уже после СМЕРТИ Л. и. брежнева* заменено на *уже после СМЕРТИ БУКВА. и. брежнева.* Здесь стоит отметить, что инициал *И.*, очевидно, воспринимается алгоритмом как лексема — союз *и* и потому не требует замены.

Представленные ошибки имеют комплексный характер, поскольку в конечном счете происходит замена гипонима на гипероним.

Таким образом, в результате синонимических замен произошло существенное количество ошибок, связанных с неудачным выбором синонима из синонимического ряда. Эта проблема отмечается и в других исследованиях, посвященных автоматизации процесса адаптации текста [8].

Выводы

Анализ результатов автоматической адаптации учебного текста путем замены ряда слов на единицы, представленные в лексическом минимуме уровня А1, показал достаточно высокую долю адекватных замен (5228 словоформ из 6431, 81 %). В то же время при автоматической замене было допущено существенное количество комплексных ошибок, связанных с рядом характеристик лексемы: её семантикой и стилистической окрашенностью.

Основными причинами возникновения этих ошибок стали: 1) неразличение лексико-семантических вариантов многозначного слова или неразличение омонимов, 2) неразличение стилистически окрашенных и нейтральных синонимов, 3) неузнавание имен собственных (географических наименований, имен и фамилий людей), 4) неузнавание условных сокращений, принятых в специализированных изданиях.

Процент ошибок можно снизить путем внесения некоторых изменений в формируемый на подготовительном этапе словарь синонимов независимо от уровня языковой подготовки:

1. Необходимо уточнить, какой именно лексико-семантический вариант многозначного слова или омоним включен в лексический минимум: *карта* как географическая карта или как открытка (уст.); *мир* как отсутствие войны или как Земля; *курс* как этап обучения, дисциплина или стоимость валюты, — то есть ввести критерий вероятности использования слова в том или ином значении.

2. Необходимо исключить из списка потенциальных слов на замену тех ЛСВ, которые обладают разговорной, просторечной окраской, а также жаргонные слова, слова с эмоционально-экспрессивной коннотацией и эвфемизмы, — то есть ввести критерий вероятности использования слова как стилистически окрашенной единицы.

3. Необходимо расширить словарь за счёт включения в него: а) лексикализованных сочетаний: производных предлогов, наиболее частотных вводных конструкций, составных наименований, терминологических словосочетаний и так далее, б) имен собственных (краткий список, данный в лексическом минимуме, явно недостаточен), в) наиболее частотных условных сокращений.

Анализ текста должен осуществляться не пословно, а с учетом ближайшего контекста, позволяющего выявить семантические и стилистические характеристики единицы, что, в свою очередь, позволит исключить слова, не подлежащие замене, а также избежать ошибок, связанных с использованием стилистически и семантически неудачных синонимов.

Внедрение изменений в алгоритм позволит повысить качество базы данных для дальнейшего обучения нейронной сети адаптации учебных текстов для инофонов с уровнем владения русским языком как иностранным А1.

Список источников

1. Азимов Э. Г., Щукин А. Н. Новый словарь методических терминов и понятий (теория и практика обучения языкам). М., 2009. 448 с.

2. Государственный стандарт по русскому языку как иностранному. Элементарный уровень / Владимирова Т. Е. и др. СПб., 2001. 28 с.

3. Коротышев А. В. «Матрица адаптации» как комплекс приёмов для отбора и адаптации художественного текста в аспекте РКИ // Мир русского слова. 2014. № 1. С. 79–85.

4. Лапошина А. Н., Лебедева М. Ю. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. № 3. С. 331–345.

5. Лексический минимум по русскому языку как иностранному: элементарный уровень: общее владение / Н. П. Андрюшина, Т. В. Козлова. СПб., 2012. 79 с.

6. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой. М., 1990. 682 с.

7. Мильчин А. Э., Чельцова Л. К. Справочник издателя и автора. Редакционно-издательское оформление издания. М., 2003. 800 с.

8. Ниценко А. В., Шелепов В. Ю., Большакова С. А., Ивашко К. С. О словесных заменах, сохраняющих смысл русского предложения // Проблемы искусственного интеллекта. 2020. № 1 (16). С. 63−74.

9. Шарафутдинова О. И. Детская литература на уроках РКИ: к проблеме адаптации художественного текста // Проблемы преподавания филологических дисциплин иностранным учащимся. Воронеж, 2010. С. 115−119.

10. Akhmetov I., Krassovitskiy A., Ualiyeva I., Gelbukh A., Mussabayev R. An Open-Source Lemmatizer for Russian Language based on Tree Regression Models // Research on computing science. 2020. URL: https://www.researchgate.net/profile/Iskander-Akhmetov/publication/344473509_An_Open-Source_Lemmatizer_for_Russian_Language_based_on_Tree_Regression_Models/links/5f7af121299bf1b53e0e460a/An-Open-Source-Lemmatizer-for-Russian-Language-based-on-Tree-Regression-Models.pdf. (дата обращения 16.05.2024).

11. Dereza, O. V., Kayutenko, D. A., Fenogenova, A. S.: Automatic morphological analysis for Russian: a comparative study // Proceedings of Student Session of Dialogue-2016. 2016. URL: https://www.dialog-21.ru/media/3473/dereza.pdf. (дата обращения 16.05.2024).

12. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. 2015. P. 320−332.

13. Kotelnikov E., Razova E. and Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? // Communications in Computer and Information Science, 2018.

14. Kuzmenko E. Morphological analysis for Russian: integration and comparison of taggers. In: Proceedings of 5th International Conference on Analysis of Images, Social Networks and Texts (AIST-2016). 2016. P. 162–171. URL: https://www.hse.ru/data/2016/06/10/1117658168/morphological-analysis-russian-1.pdf. (дата обращения 16.05.2024).

15. Litvinova T., Seredin P., Litvinova O., and Zagorovskaya O. Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts // Proceedings of the Workshop on Stylistic Variation. Association for Computational Linguistics. Copenhagen, Denmark, 2017. P. 69–73. URL: https://aclanthology.org/W17-4909.pdf. (дата обращения 16.05.2024).

References

1. Asimov EG, Shchukin AN. New Dictionary of Methodological Terms and Concepts (Theory and Practice of Language Teaching). M., 2009. 448 p. (In Russ.)

2. Vladimirova TE et al. State standard for Russian as a foreign language. Elementary level. St. Petersburg, 2001. 28 p. (In Russ.)

3. Korotyshev AV. «Matrix of adaptation» as a set of techniques for the selection and adaptation of an artistic text in the aspect of RCT. *Mir russkogo slova = The World of Russian Word.* 2014;(1):79-85. (In Russ.)

4. Laposhina AN, Lebedeva MYu. Textometer: an online tool for determining the level of text complexity in Russian as a foreign language. *Rusistika = Russistics.* 2021;19(3):331-345. (In Russ.)

5. Lexical minimum in Russian as a foreign language: elementary level: general knowledge. Ed. by NP Andryushina, TV Kozlova. St. Petersburg, 2012. 79 p. (In Russ.)

6. Linguistic Encyclopedic Dictionary. Ed. by VN Yartseva. M., 1990. 682 p. (In Russ.)

7. Milchin AE, Cheltsova LK. Handbook of publisher and author. Editorial and publishing design of the edition. M., 2003. 800 p. (In Russ.)

8. Nitsenko AV, Shelepov VYu, Bolshakova SA, Ivashko KS. About word substitutions preserving the meaning of the Russian sentence. *Problemy iskusstvennogo intellekta = Problems of Artificial Intelligence.* 2020;1(16):63-74. (In Russ.)

9. Sharafutdinova OI Children's literature at the lessons of RCT: to the problem of adaptation of the art text. Problemy prepodavaniya filologicheskih disciplin inostrannym uchashchimsya = Problems of teaching philological disciplines to foreign students. Voronezh, 2010. P. 115-119. (In Russ.)

10. Akhmetov I, Krassovitskiy A, Ualiyeva I, Gelbukh A, Mussabayev R. Lemmatizer for Russian Language based on Tree Regression Models. Research on computing science 2020. Available from: https://www.

researchgate.net/profile/Iskander-Akhmetov/publication/344473509_An_Open-Source_Lemmatizer_for_Russian_Language_based_on_Tree_Regression_Models/links/5f7af121299bf1b53e0e460a/An-Open-Source-Lemmatizer-for-Russian-Language-based-on-Tree-Regression-Models.pdf.

11. Dereza OV, Kayutenko DA, Fenogenova AS. Automatic morphological analysis for Russian: a comparative study. Proceedings of Student Session of Dialogue-2016. Available from: https://www.dialog-21.ru/media/3473/dereza.pdf.

12. Korobov M. Morphological Analyzer and Generator for Russian andUkrainian Languages // Analysis of Images, Social Networks and Texts, 2015, pp. 320-332. (In Russ.)

13. Kotelnikov E, Razova E and Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? Communications in Computer and Information Science, 2018.

14. Kuzmenko E. Morphological analysis for Russian: integration and comparison of taggers. Proceedings of 5th International Conference on Analysis of Images, Social Networks and Texts (AIST-2016), 2016, pp. 162-171. Available from: https://www.hse.ru/data/2016/06/10/1117658168/morphological-analysis-russian-1.pdf.

15. Litvinova T, Seredin P, Litvinova O, and Zagorovskaya O. Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts. Proceedings of the Workshop on Stylistic Variation, Association for Computational Linguistics. Copenhagen, Denmark, 2017, pp. 69-73. Available from: https://aclanthology.org/W17-4909.pdf.

Информация об авторах

О. Ю. Редькина — кандидат филологических наук, доцент кафедры русского языка и литературы.

Т. В. Карпета — кандидат физико-математических наук, доцент кафедры прикладной математики и программирования.

Д. С. Пономарева — магистрант 2 курса по направлению подготовки 45.04.01 Филология.

Д. А. Вершинина — студент 2 курса по направлению подготовки 01.03.02 Прикладная математика и информатика.

Information about the authors

O. Yu. Redkina — Candidate of Philological Sciences, Associate Professor at the Department of Russian Language and Literature.

T. V. Karpeta — Candidate of Physical and Mathematical Sciences, Associate Professor at the Department of Applied Mathematics and Programming.

D. S. Ponomareva —Master of Arts (Philology).

D. A. Vershinina — Master of Science (Mathematics and Informatics).

Статья поступила в редакцию 21.06.2024; одобрена после рецензирования 28.08.2024; принята к публикации 24.09.2024.

The article was submitted 21.06.2024; approved after reviewing 28.08.2024; accepted for publication 24.09.2024.

Вклад авторов: Редькина О. Ю. — систематизация ошибок, подготовка текста публикации. Пономарева Д. С. — обработка адаптированного текста, систематизация ошибок. Карпета Т. В., Вершинина Д. А. — подготовка и проведение адаптации текста, их описание в тексте статьи.
Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: Redkina O. Yu. — systematization of errors, preparation of the publication text. Ponomareva D. S. — processing of the adapted text, systematization of errors. Karpeta T. V., Vershinina D. A. — preparation and implementation of text adaptations, their description in the text of the article.
The authors declare no conflicts of interests.