Информационные технологии в образовании. 2022. Вып 5

Исследование возможности автоматической генерации тестовых материалов для проверки знаний правописания слов русского языка

Гареева А.А.[1], Шатская Е.И.[2], Иванов А.А.[3]

[1]*aagareeva@edu.hse.ru*, [2]*eishatskaya@edu.hse.ru*, [3]*ssauivanov@gmail.com*

Национальный исследовательский университет «Высшая школа экономики»

Аннотация. В работе описывается опыт студентов третьего курса бакалавриата в рамках работы над курсовым проектом «Разработка бота-викторины на знание русского языка с автоматической генерацией тестов», выполненным в ВШЭ на факультете компьютерных наук. Целью работы ставится разработка чат-бота, способного в интерактивном режиме, в виде онлайн-викторины, опрашивать участника на предмет знания русского языка средствами вопросов, в которых пропущено правильное написание слова и с ответами в формате одиночного выбора из представленных вариантов. К задачам проекта относится разработка средств автоматизированного создания таких вопросов, а именно: выбор сложного предложения из некоторого текстового корпуса, выбор сложного слова в этом предложении и внесение в него неочевидных искажений для создания вариантов ответов в автоматическом режиме. В работе рассматривается понятие сложности слова и сложности предложения, методы выбора орфограмм и особенности разработки чат-интерфейса на базе одного из популярных мессенджеров. В завершении заключается, что использование наивного подхода создания затруднений в вариантах ответов может служить базовым уровнем для разработки подобных систем, использование статистических моделей на основе цепей Маркова позволяет удалить из генерируемого набора тестов наиболее очевидные варианты неправильных ответов. Область применения разработанного программного продукта ограничивается преподаванием русского языка как иностранного и аудиторией младших классов начальных школ. Предлагается идея внедрения аппарата нейронных сетей для внесения неочевидных ошибок в слова как противопоставление тренду на создание нейронных сетей, исправляющих ошибки.

Ключевые слова: русский язык, тест, викторина, бот, информационные технологии в образовании.

Онлайн-опросы в популярных мессенджерах и социальных сетях играют важную роль, как образовательную, так и развлекательную, обращают внимание; часто интернет-аудиторию привлекают различного рода викторины, когда, помимо правильного ответа на поставленный вопрос, пользователь видит, как отвечали другие участники. Одним из пионеров использования опросов в целях развлечения и маркетинга можно считать медиакомпанию BuzzFeed; используя простые механики множественного и одиночного выбора из нескольких картинок, каждая из которых отвечает на задаваемый вопрос, компания существенно изменила рынок т.н. нативной рекламы, встраивая рекламное сообщение в основной поток потребляемой информации были реализованы успешные кампании для таких компаний как HBO, Mattel, Taco Bell и др. [1]. В работе авторы обратили внимание на популярность специальных каналов в мессенджерах с образовательной направленностью, в которых также проводятся викторины на знание русского языка с различного рода затруднениями: начиная от грамматики сложных слов, правил постановки ударений, словоупотребления и др. Авторы провели анализ трех каналов в

Информационные технологии в образовании. 2022. Вып 5

мессенджере Telegram на предмет популярности формата викторины, в результате можно предположить, что от 10% до 20% аудитории активно участвуют в подобных активностях. Детали исследования представлены на рисунке 1.

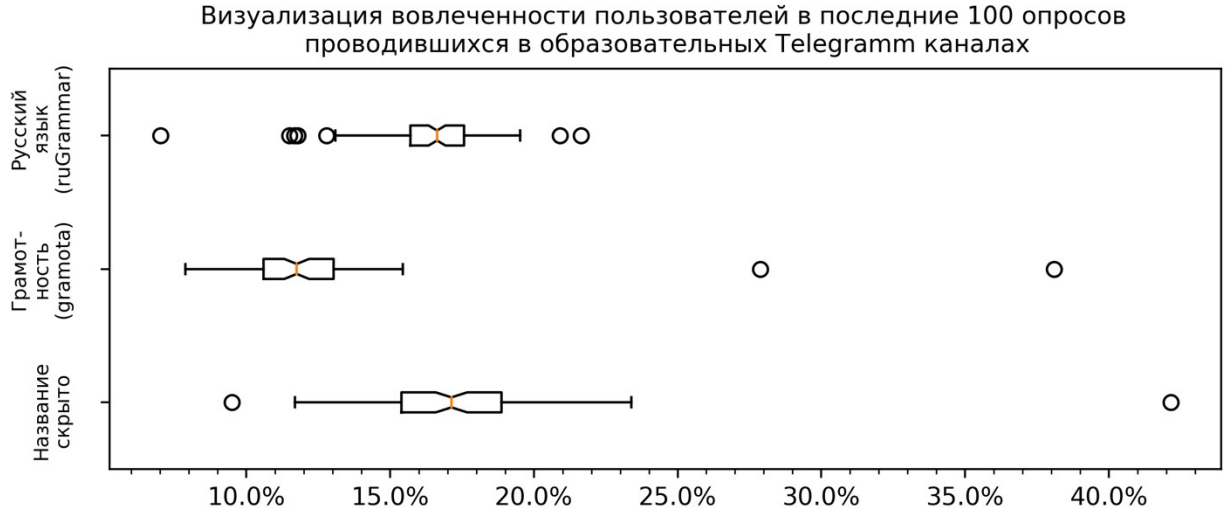

Рис. 1. Результаты анализа ста последних сообщений по механике викторина в каждом из трех рассмотренных каналов.
По оси абсцисс указывается процент от аудитории канала (зафиксированный на момент проведения исследования – август 2022 г.) который участвовал в викторине.
Можно сделать вывод о том, что от 10% до 20% аудитории заинтересована в таком формате, более того, отдельные викторины имели охват более 40%.

Авторы предлагают идею автоматического создания подобных викторин, основанных на некотором текстовом корпусе, что может иметь дидактическое значение при изучении творчества писателей на уроках русского языка и литературы. В качестве тестового корпуса было выбрано произведение Кира Булычева «Путешествие Алисы», который богат неологизмами, окказионализмами, искусственными словами и именами собственными, например, Громозека, инопланетчик, говорун, колеидяне, чумарозец и др. Вместе с этим описание вымышленного мира Булычева наполнено диалогами и сложными предложениями, выбирая особым образом которые можно генерировать затруднения в форме тестов.

Для отбора предложений, участвующих в тестах, предлагается посчитать их сложность во всем текстовом корпусе, основываясь на сложности входящих слов, определяющих диагностирующий потенциал: частотность и степень известности слова, фонетическая сложность, графическая сложность, морфемная сложность, семантическая сложность [3].

Итоговая формальная сложность предложения может быть, например, средним от сложности входящих слов, сложность которых в свою очередь может оцениваться, например, медианой от различных метрик. Используя различные библиотеки и наборы данных (pymorphy2 [4], [5], [6]), разработанные для целей компьютерного анализа русского языка, в таблице 1 представлены следующие примеры оценки сложности слов.

Информационные технологии в образовании. 2022. Вып 5

Таблица 1. Примеры расчета сложности слов по различным критериям: усредненное значение (обобщенная сложность), частотная сложность, графическая сложность, морфемная сложность, фонетическая сложность. Значения не имеют единиц измерения и нормированы.

	Численные значения различных сложностей				
Слово	**Обобщ.**	**Частот.**	**Граф.**	**Морф.**	**Фонет.**
из	0,05	0,00	0,00	0,00	0,20
цикл	0,26	0,06	0,20	0,17	0,60
настоящий	0,41	0,00	0,50	0,17	1,00
путешествие	0,60	0,08	0,70	0,83	0,80
неподходящий	0,80	0,75	0,80	0,67	1,00
материализоваться	0,94	0,92	1,00	0,83	1,00

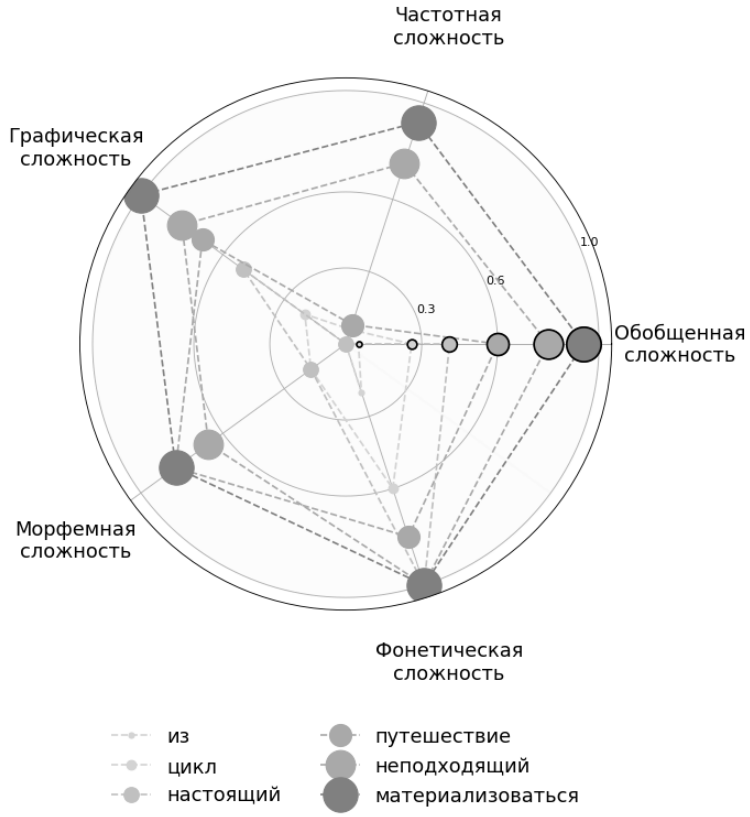

Рис. 2. Визуализация информации из Таблицы 1: сложности шести слов: «из», «цикл», «настоящий», «путешествие», «неподходящий», «материализоваться».

Разработанные программные средства определения обобщенной сложности слова были проверены на корпусах слов [7], распределенных по 6 уровням, предназначенных для студентов, изучающих русский язык как иностранный. Результаты анализа приведены на рисунке 3.

В результате используется следующая процедура подготовки тестов в полуавтоматическом режиме:

1. Выбор текстового корпуса, которым может, например, быть любая книга.
2. Определение сложных слов и предложений, согласно описанным выше критериям сложности. Фильтрация «простых» предложений.
3. Выбор конкретного слова для создания орфограммы, затруднения в «сложном» предложении.

Информационные технологии в образовании. 2022. Вып 5

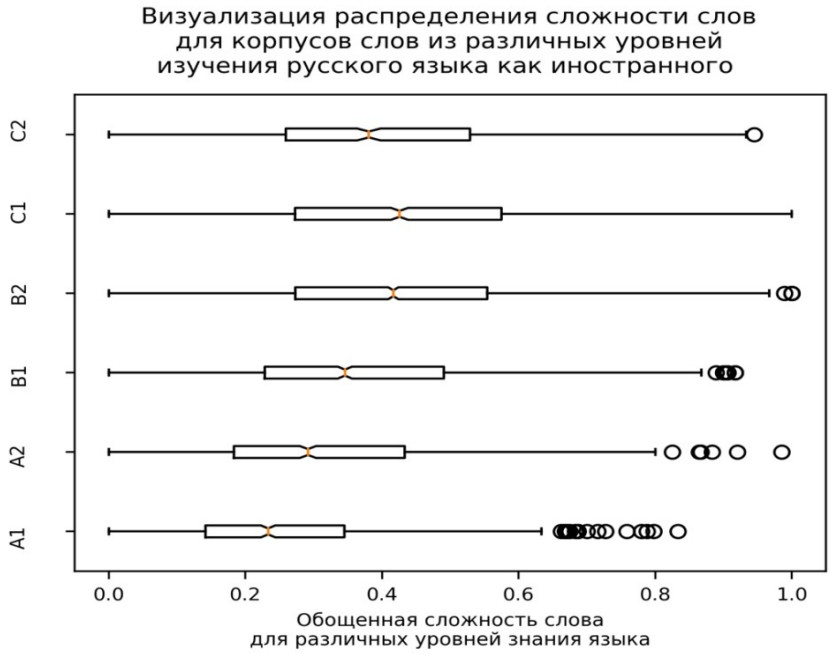

Рис. 3. Демонстрация распределений разработанной обобщенной оценки сложности слов на корпусах слов различных уровней при изучении русского как иностранного. Определенно, слова из уровня B2 сложнее в используемой метрике, чем на уровне A1, однако для уровней C1 и C2 это не является очевидным

Если с анализом сложности на уровне слов и предложений все представляется довольно очевидным и ранее разработанным, то автоматическое внесение искажений (орфограмм) в правильно написанное слово определённо не алгоритмизировано. Более того, вносимые в слова затруднения должны не быть очевидными, в том числе и для носителей языка.

Таблица 2. Примеры генерируемых затруднений для тестов с одиночным выбором правильного написания слова в предложении

Очевидный пример тестового задания	**Усложненный пример тестового задания**
Ил поднялся со дна, застилая <*выберите правильное слово*> пятно удаляющегося фонаря.	В Питере мне нужно <*выберите правильное слово*> где-то пару ночей.
расплывчотое	*переконтоваться*
разплывчатое	*перекантаваться*
расплывчатое	*перекантоваться*
разплывчотое	*переконтаваться*

Авторы в работе предлагают в первую очередь вносить затруднения, основанные на экспертных правилах, например, таких:

1. Парные по глухости и звонкости согласные: *гриб – гриП*.
2. Чередование безударных гласных в корне: *отворится – отвПрится*.
3. Правописание приставок: *бесприданница – беЗприданница*.
4. Гласные «Ы» и «И» после приставок: *межирригационный – межЫрригационный*.
5. Правописание морфем *«ться»* и *«тся»*.
6. Двойные согласные: *беззаконный – беззакоНый – беЗаконный - беЗакоНый*.

Информационные технологии в образовании. 2022. Вып 5

7. Непроизносимые согласные: *ужасный – ужасТный*.
8. Употребление твердого знака: *объем – обЬем*.
9. Дефисные написания: всё-таки – *всётаки - всё таки*.

Использование подобных экспертных правил в автоматическом режиме средствами анализа языка почти всегда приводит к сильно очевидным затруднениям, в особенности, – для носителей языка. Проявляемые очевидности предлагается отфильтровать, используя статистические идеи, основанные на цепях Маркова. На рисунке 4 продемонстрирован граф переходов для однородной марковской цепи, построенной для событий следования одной буквы за другой для текстового корпуса из [2]. В качестве метрики очевидности искажения можно ориентироваться на величину перплексии для конкретного слова.

Таблица 3. Примеры генерируемых затруднений и рассчитанные значения перплексии для них. Значение Inf описывает, что данное затруднение является невозможным для исследуемого корпуса слова, а значит – наиболее очевидным, как неправильный вариант ответа в тесте. Меньшие значения перплексии соответствуют неочевидным затруднениям или верному написанию.

Слово	Значение перплексии	Вывод о затруднении
расплывчотое	Inf	Слишком очевидное затруднение
разплывчатое	Inf	Слишком очевидное затруднение
расплывчатое	**20,2**	**Верное написание**
разплывчотое	Inf	Слишком очевидное затруднение
переконтоваться	10,7	Неочевидное затруднение
перекантаваться	10,9	Неочевидное затруднение
перекантоваться	**10,1**	**Верное написание**
переконтаваться	11,5	Неочевидное затруднение

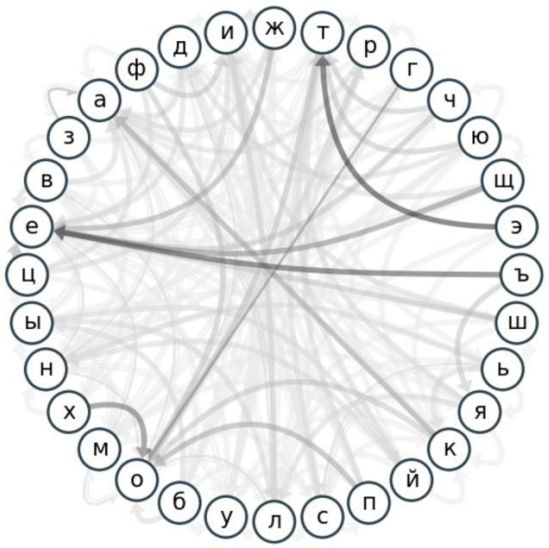

Рис. 4. Ориентированный граф переходов для цепи Маркова, в которой каждое состояние – определенная буква в слове для текстового корпуса из [2].
Из графа видно, например, правило написания твердого знака перед буквами «е» и «я», наиболее часто после буквы «э» следует буква «т», а после буквы «х» - буква «о» и т.д.

Информационные технологии в образовании. 2022. Вып 5

Рассмотренный авторами подход нашел свою реализацию в виде Telegram бота, в котором пользователь (участник викторины) может почти бесконечно получать различные автоматически сгенерированные задания на проверку навыков правописания. Однако определенная часть затруднений все еще остаются очевидными и соответствующие тесты – слишком простыми. Остается открытым вопрос как данную «очевидность» измерять априорно, предложенный авторами подход с использованием перплексии полностью не решает задачу, хотя остается потенциал для аналогичного подхода но, – с триграммами и т.п.

Для полноты исследования и его завершения необходимо рассмотреть современные нейросетевые подходы, применяющиеся в задачах исправления ошибок в текстах, с целью их использования для обратных целей – внесения ошибок.

Список литературы

[1] Реклама под прикрытием: Секреты эффективности BuzzFeed. Разоблачение формата «натив».- [Электронный ресурс] URL: https://secretmag.ru/opinions/reklama-pod-prikrytiem-sekrety-efektivnosti-buzzfeed.htm (дата обращения: 27.08.2022).

[2] *Булычев К.* Путешествие Алисы. Litres, 2022 – 240 с.

[3] *Душейко А.С., Резанова З.И., Аленина Е.А.* Параметр сложности слова в диагностике лексического компонента языковой способности // Вестник Томского государственного университета. 2019. №. 442. – С. 22-31.

[4] *Коробов М.* Морфологический анализатор pymorphy2 // Морфологический анализатор pymorphy2 [Электронный ресурс] URL: https://pymorphy2.readthedocs.io/en/stable/index.html (дата обращения: 22.08.2022).

[5] *Ляшевская О.Н., Шаров С.А.* Новый частотный словарь русской лексики // Словари на основе национального корпуса русского языка [Электронный ресурс] URL: http://dict.ruslang.ru/freq.php (дата обращения: 22.08.2022).

[6] *Тихонов А.Н.* Словообразовательный словарь русского языка в двух томах [Электронный ресурс] URL: http://speakrus.ru/dict2/index.htm (дата обращения: 22.08.2022).

[7] Русские слова по уровням. [Электронный ресурс] URL: http://www.tolstyslovar.com/ru/a1 (дата обращения: 22.08.2022).