Сравнительный анализ средств проверки орфографии литературного арабского языка

Полина П. Кремнева
Российский государственный гуманитарный университет,
Москва, Россия, kremneva2003@yandex.ru

Данная работа посвящена анализу средств проверки орфографии (далее – спеллчекер[1]) литературного арабского языка. В настоящее время все больше арабоязычных пользователей Интернета используют именно литературный арабский язык, т.к. это делает информацию доступной для более широкого круга пользователей. Орфографические особенности арабского языка влекут за собой большое количество орфографических ошибок, что делает актуальными исследования в данной области. Однако задача автоматической коррекции ошибок связана не только с орфографическими особенностями арабского, но также с проблемой оценки лингвистической приемлемости текста и с задачей понимания естественного языка (ведь если модель не “понимает” язык, она не может верно исправлять связанные с семантикой и прагматикой ошибки).

В арабском языке выделяются следующие виды орфографических ошибок [Gheith A. Abandah, Ashraf Suyyagh, Mohammed Z. Khedher 2021]:

1. дискурсивные орфографические ошибки;
2. прагматические ошибки;
3. морфо-синтаксические ошибки;
4. лексические ошибки;
5. контекстные / семантические ошибки;
6. орфографические ошибки, связанные с вариативностью написания букв арабского алфавита

[1] Спеллчекер - компьютерная программа, осуществляющая проверку заданного текста на наличие в нём орфографических ошибок.

Целью работы было выяснить, какие виды орфографических ошибок в арабском языке вызывают больше сложностей у спеллчекеров а также предположить, с чем это может быть связано и подготовить предложения по улучшению их работы.

Для анализа материала были использованы правиловый спеллчекер Ar-corrector 1.1.6[2], использующий для поиска и исправления ошибок редакционные расстояния (позволяют формально сравнивать предложения или отдельные слова и определять, какие из них наиболее похожи друг на друга) и языковую модель с N-граммами (для исправления слов в зависимости от предыдущего контекста), а также нейросетевой спеллчекер Arabisc[3] (LSTM рекуррентная сеть), обученный на материале 554622 арабских предложений, взятых из корпусов News-Commentary Corpus (OPUS) v11 и MultiUn (OPUS) и использующий комбинацию статистического и нейросетевого методов определения ошибок.

Исследование проводилось на материале 500 записей из двух открытых аккаунтов арабоязычных пользователей сети Twitter[4]. Тексты данных аккаунтов отличались по стилю и лексическому наполнению, один из аккаунтов принадлежит девушке-художнице, другой – журналисту.

Для каждого типа (с 3 по 6) описанных выше орфографических ошибок, а также отдельно для каждого спеллчекера мы посчитали F-меру (мера точности теста) и составили таблицы контингентности.

	precision	recall	F
морфо-синтаксические	0,355	0,88	0,506
лексические	0,902	0,836	0,868

[2] https://pypi.org/project/ar-corrector/#description
[3] Arabic Spelling Checker - a Hugging Face Space by mohamedabdullah
[4] https://twitter.com

контекстные/семантические	0,355	0,828	0,497
soft-spelling	0,85	0,156	0,264

Таблица 1. Ar-corrector. F-мера

	precision	recall	F
морфо-синтаксические	0,09	0,6	0,157
лексические	0,397	0,855	0,537
контекстные/семантические	0,135	0,8	0,23
soft-spelling	0,813	0,238	0,369

Таблица 2. Arabisc. F-мера

Основываясь на полученной статистике, мы пришли к следующим выводам: лучше всего средства проверки орфографии справляются с лексическими ошибками, это можно объяснить тем, что для их обнаружения не всегда нужен даже контекст. Кроме того, спеллчекер, основанный на нейронной сети, не дает значимого прироста качества, как мы предполагаем, из-за обучающих данных. Именно из-за них показатели качества работы нейросети зачастую гораздо ниже, чем показатели языковой модели с N-граммами. Также у спеллчекеров наблюдаются проблемы с определением части речи, обработкой слитных местоимений (они в арабском языке могут выполнять функцию прямого или косвенного дополнения при глаголе, а также несогласованного определения при имени), определением тропов, удалением повторяющихя символов, проблемы с междометиями и диалектной лексикой, а также с “охранным нуном” (ставится между основой глагола и слитным местоимением 1-го лица ед.ч.).

Итак, хуже всего спеллчекеры справляются с ошибками, хоть немного затрагивающими семантику, а как следствие, и прагматику. Это наблюдение натолкнуло нас на мысль о необходимости улучшения работы нейросетевого спеллчекера.

Мы подготовили предложения по улучшению его работы, связанные с задачей понимания естественного языка:

— Использование архитектуры трансформер[5].

— Работа с обучающими данными – автоматические аугментации[6] и морфо-синтаксическая разметка[7].

Следует отметить, что данная работа - один из первых шагов в сторону задачи понимания естественного языка, однако выявленные нами проблемы и предложенные возможные варианты их решения в перспективе помогут сделать работу средств автоматической проверки орфографии для арабского языка более качественной.

Литература:

1. Крачковская В. А. Новое исследование по истории арабской письменности // «Учёные записки ЛГУ». 1974. № 374
2. Ярцева В. Н. Лингвистический энциклопедический словарь. М., 1990
3. Gheith A. Abandah, Ashraf Suyyagh, Mohammed Z. Khedher. Correcting Arabic Soft Spelling Mistakes using BiLSTM-Based Machine Learning // International Journal of Advanced Computer Science and Applications — 2021. — vol 13, № 5. — 2-4.
4. Rasha Altarawneh. Spelling Detection Errors Techniques in NLP: A Survey // International Journal of Computer Applications — 2017. — vol 172, № 4.

[5] Трансформер (transformer) – архитектура глубоких нейронных сетей, работающая без использования рекуррентности (последовательных вычислений). Данная архитектура основана на механизме внимания (attention), который позволяет ей эффективно работать с контекстом, благодаря выделению релевантной части входных данных

[6] Аугментация (augmentation) – это построение дополнительных данных из исходных при решении задач машинного обучения. Аугментация текстовых данных обучающих наборов заключается в изменении содержания их образцов. Среди общих техник аугментации можно выделить следующие: замена слов в предложении, перестановка слов в предложении, генерация новых предложений на основе изменения структуры исходных.

[7] Морфо-синтаксическая разметка – тип разметки, при котором каждому слову текста приписывается морфологическая информация, а также задается его синтаксическая структура.

5. Ritika Mishra, Navjot Kaur. A Survey of Spelling Error Detection and Correction Techniques // International Journal of Computer Trends and Technology. — 2013. — vol 41. — 372.
6. Yasmin Moslem, Rejwanul Haque and Andy Way. Arabisc: Context-Sensitive Neural Spelling Checker // Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications — 2020. — 11–19.
7. Arabic, Standard [Ethnologue 25th edition]. - URL: Arabic, Standard | Ethnologue
8. Arabisc. - URL: Arabic Spelling Checker - a Hugging Face Space by mohamedabdullah
9. Ar-corrector 1.1.6. - URL: ar-corrector · PyPI