УДК 004.934.5

АВТОМАТИЧЕСКАЯ РАССТАНОВКА УДАРЕНИЙ В НЕЗНАКОМЫХ МЕДИЦИНСКИХ ТЕРМИНАХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Перминова А. А.[1]
Научный руководитель – канд. физ.-мат. наук Рыбин С. В.[1]
[1]Университет ИТМО
nastyaperminova2002@yandex.ru

Введение

В эпоху цифровизации системы синтеза речи озвучивают медицинские диагнозы, учебные материалы и помогают людям с ограниченными возможностями. Однако медицинские термины, как правило, являются незнакомыми (out-of-vocabulary) для стандартных алгоритмов. Отсутствие систем, ориентированных на специализированную медицинскую терминологию, приводит к ошибкам в синтезе профессиональной речи для цифровых медицинских сервисов и образовательных платформ. Ошибка в ударении в медицинском контексте может не только снизить доверие к сервису, но и повлечь недопонимание.

Основная часть

Эволюция подходов к решению задачи акцентуации прослеживается от точных, но ограниченных словарных методов, через статистические модели, где была доказана критическая важность морфемных признаков [1], к современным нейросетям, которые учатся на символьных последовательностях.

Подход на основе правил был реализован в работе Reynolds & Tyers [2]. Система на базе словаря Зализняка [3] и морфосинтаксического анализатора (Constraint Grammar) для разрешения неоднозначностей в контексте показала точность 93,21 %. Однако ее критическим недостатком стала невозможность обработки незнакомых (out-of-vocabulary, OOV) слов, что ограничивает применение данного подхода на медицинской терминологии.

В работе Ponomareva et al. [4] была реализована посимвольная двунаправленная рекуррентная нейронная сеть с LSTM. Работа показала превосходство модели, обученной на реальном корпусе текстов (~97,7-97,9 %), над моделью, обученной на словарных парадигмах (75,1 %).

Закономерным итогом эволюции стало возникновение гибридных архитектур, стремящихся объединить достоинства подходов, перечисленных выше. Практическая реализация подобной архитектуры для поэтических текстов представлена в работе Ю. О. Коротковой [5]. Комбинированная система проверяет слово по словарю и только в случае его отсутствия или наличия неоднозначностей передает его нейросети. Такой подход позволил достичь точности 97,5-99 %, существенно превзойдя каждый из методов по отдельности.

Выводы

В ходе работы был проведен анализ существующих методов расстановки ударений, и на его основе, в качестве наиболее перспективного, выбран гибридный подход для решения поставленной задачи. Также был запущен процесс формирования набора данных по медицинской терминологии для обучения и оценки будущей модели.

Литература

1. Hall K., Sproat R. Russian stress prediction using maximum entropy ranking // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013. P. 879–883.
2. Reynolds R., Tyers F. Automatic word stress annotation of Russian unrestricted text // Proceedings of the 20th Nordic Conference of Computational Linguistics. Vilnius, Lithuania, 2015. P. 173–180.
3. Зализняк А. А. Грамматический словарь русского языка: словоизменение. – Москва: Русский язык, 1977. 880 с.
4. Ponomareva M., Milintsevich K., Chernyak E. Automated Word Stress Detection in Russian // Proceedings of the First Workshop on Subword and Character Level Models in NLP. Copenhagen, Denmark, 2017. P. 31–35.
5. Короткова Ю. О. Комбинированный словарно-нейросетевой акцентуатор для разметки русского поэтического текста // Труды института русского языка имени В. В. Виноградова. 2022. № 3 (33). С. 181–190.