

Валерия Михайловна Василевская, аспирант
Минский государственный лингвистический
университет, Минск, Беларусь
э-почта: valeria-vas90@mail.ru

Valeria Mihailovna Vasilevskaya, PhD Student
Minsk State Linguistic University, Minsk,
Belarus
e-mail: valeria-vas90@mail.ru

ЛИНГВИСТИЧЕСКИЕ ОШИБКИ НЕЙРОСЕТЕЙ ПРИ РЕФЕРИРОВАНИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Проанализированы автоматические рефераты русскоязычных текстов, созданные автором при помощи сервисов на базе нейросетей, с учетом содержательных критериев. Выделены основные лингвистические ошибки, с которыми сталкиваются нейросети при реферировании текстов на русском языке.

Ключевые слова: автоматическое реферирование; нейросеть; текстовый анализ.

LINGUISTIC FAILS OF NEURAL NETWORKS IN ABSTRACTING RUSSIAN-LANGUAGE TEXTS

Automatic abstracts of Russian-language texts created by the author with the help of services based on neural networks, taking into account the content criteria, are analyzed. The main linguistic fails encountered by neural networks in abstracting texts in Russian are highlighted.

Key words: automatic abstracting; neural network; text analysis.

В современном мире развитие информационных технологий оказывает значительное влияние на различные сферы деятельности человека. В условиях дефицита времени оперативное получение информации становится жизненно необходимым. В этом контексте автоматическое реферирование с применением нейронных сетей приобретает все большую популярность. Однако на сегодняшний день качество вторичных текстов, созданных нейросетями на основе русскоязычных материалов, остается недостаточно изученным с лингвистической точки зрения.

Русский язык, будучи синтетическим и обладающим богатой морфологией, предъявляет особые требования к автоматической обработке. Среди ключевых сложностей можно выделить разнообразие форм слов, свободный порядок слов, многоуровневую вложенность синтаксических конструкций, контекстуально зависимую семантику.

Цель настоящего исследования состоит в выявлении общих тенденций нейросетей к совершению ошибок при передаче содержания русскоязычного текста в процессе реферирования. Для анализа использовались два независимых популярных сервиса на базе нейросетей: 1) YaGPT, 2) myneuralnetworks.

Материалом исследования послужили автоматические рефераты пяти текстов, взятых с официального сайта БелТА. Для каждого текста были созданы по два реферата (по одному на каждом из сервисов). Полученные рефераты сопоставлялись с их оригинальными текстами и между собой. Выявленные нарушения в передаче содержания анализировались по следующим критериям (предложенным Д. В. Белогорской и З. И. Резановой): 1) полнота отражения содержания; 2) точность передачи данных, включая: а) наличие / отсутствие искаженной информации; б) степень полноты представления данных; в) наличие / отсутствие логических нарушений и связности текста; г) наличие / отсутствие избыточных повторов фрагментов текста. Количественные результаты анализа представлены ниже.

Табл. 1 Оценка лингвистических ошибок в работе сервисов YaGPT и myneuralnetworks при реферировании неспециализированных русскоязычных текстов

Критерий оценивания	Количество ошибок, %	
	YaGPT	myneuralnetworks
1. Полнота отражения содержания	17,2	11,1
2. Точность передачи данных	82,8	88,9
а) искажение информации	41,4	27,9
б) неполнота представления информации	20,7	50,0
в) нарушение логики/связности текста	6,9	11,0
г) избыточный повтор	13,8	0,0

Сопоставительный анализ результатов работы двух сервисов показывает, что оба инструмента в целом успешно выделяют микротемы исходного текста. Однако при этом наблюдается общая тенденция к нарушениям точности передачи данных в рефератах. Различие в доминирующих типах ошибок объясняется стратегией обработки сложных мест в тексте: YaGPT стремится к более детальной передаче информации, но допускает искажения, тогда как myneuralnetworks склонен к обобщениям и пропуску трудных для обработки фрагментов.

Оценка выявленных ошибок показывает, что YaGPT демонстрирует недостаточную устойчивость к сложным языковым конструкциям, что приводит к искажению информации в 41,4 % случаев. Несмотря на это, он реже допускает неполноту представления данных (20,7%), чем myneuralnetworks (50,0%), который, напротив, стремится к сокращению содержания, нередко теряя важные детали. Кроме того, YaGPT склонен к избыточному повторению фрагментов текста (13,8%), в то время как myneuralnetworks демонстрирует более логичную структуру рефератов, но с большим количеством пропусков и нарушений связности текста (11,1%).

Наибольшие лингвистические трудности с точностью передачи данных у нейросетей возникают в следующих случаях: 1) при попытке соединить несколько последовательно идущих предложений в одно (в таких случаях нарушение когезии ведет к искажению смысла вплоть до противоположного); 2) при передаче последовательности однородных действий или событий (это приводит к тому, что последовательность воспринимается как простое перечисление: часть элементов в

может быть передана в произвольном порядке, а некоторые из них могут быть утеряны); 3) при интерпретации сложных предложений (в случае подчинительной связи нейросеть часто игнорирует второстепенные предложения, а при сочинительной связи она иногда ошибочно превращается в подчинительную: это делает адекватное восприятие сути оригинала невозможным; 4) при передаче числовых данных (дата, год, вес, время, расстояние и др.) и именованных сущностей. (неспособность нейросетей анализировать контекст и тематику текста приводит к неправильному определению важности этих данных для корректного понимания информации оригинала и, как следствие, частому неправомерному исключению их из реферата); 5) при поиске верного антецедента и разрешении анафоры либо катафоры. Неверное соотнесение с местоимением влечет за собой искажение информации и разрыв логики повествования.

Высокая степень схожести общих результатов работы двух абсолютно разных сервисов в передаче содержания при реферировании позволяет предположить, что подобные ошибки будут иметь место и при реферировании русскоязычных текстов другими сервисами на базе нейросетей.

Вышеописанные проблемы характерны для абстрактного реферирования, при котором генерируется новый текст на основе исходного. Такой подход отличается высокой степенью спонтанности и плохо поддается контролю со стороны пользователя.

Таким образом, оба сервиса имеют свои недостатки, и ни один из них не обеспечивает полностью качественного реферирования русскоязычных текстов. YaGPT подходит для случаев, когда важна детальность, но при этом необходимо учитывать риск искажения информации. Сервис *mynеuralnetworks*, напротив, обеспечивает более точное изложение, но за счет значительного упрощения и пропуска данных, что может быть нежелательно при работе с содержательно насыщенными текстами.

Учитывая высокий процент ошибок при передаче данных в процессе реферирования неспециализированных текстов (82,8% и 88,9% соответственно), результаты исследования позволяют предположить, что рассмотренные сервисы полностью непригодны для реферирования специализированных текстов узких предметных областей, включая информационно-коммуникационные технологии (ИКТ). Такие тексты требуют высокой точности в передаче терминов, числовых данных и алгоритмов, а также однозначности формулировок. Для проверки этой гипотезы нами будут проведены дальнейшие исследования, сфокусированные на текстах по тематике ИКТ, с привлечением большего числа сервисов на базе нейросетей.