М. Г. Орел, Д. А. Болдак
(*УО «ГрГУ им. Я. Купалы», Гродно*)

ОБРАБОТКА ТЕКСТА, СФОРМИРОВАННОГО НЕЙРОННОЙ СЕТЬЮ

Пока нейронные сети не научились «читать как люди», существует необходимость коррекции сформированного ими текста, что особенно актуально в связи с интенсивным развитием нейросетей.

В работе была поставлена задача распознавания и выделения слов искаженного текста (полученного в результате обработки нейронной сетью фотографических изображений, сделанных в нестудийных условиях) и их автоматической корректировки. Допустимая корректировка должна была быть согласована с содержимым некоторых словарей, ненужные слова – удалены.

Для первичного анализа текста было использовано модернизированное расстояние Левенштейна (высчитывает минимальное количество операций для преобразования одного слова/словосочетания в другое операциями удаления или вставки) и словари «ненужных слов», «нужных слов» и «нужных словосочетаний», связанных с некоторой предметной областью. Если слово содержит меньше шести букв, нет смысла исправлять в нем даже одну букву, так как это может привести к неверной корректировке. Если слово состоит из 6 – 7 букв, допускаем одну ошибку, если из 8 – 10 букв – 2 ошибки и так далее.

Алгоритм состоит из нескольких проходов. Первый проход удаляет слова или скорректированные слова, удовлетворяющее количеству ошибок, содержащиеся в словаре «ненужных» слов. Второй проход сохраняет слово, если оно найдено в словаре «нужных» слов, и формирует для слова «список подобных слов», получаемых из исходного в результате корректировки. В ходе третьего прохода из текста выделяются фрагменты слов определенной длины, склеиваются в цепочки, и построенные словосочетания ищутся в словаре «нужных словосочетаний».

В результате работы алгоритма, в тексте оставалось до 86% слов и словосочетаний, относящихся к предметной области. Описанный

метод был использован в рамках ИТ-проекта, направленного на обеспечения здорового образа жизни. Планируется его использование для корректировки текстов некоторого языка, с учетом его морфологии, позволяющей сохранять смысл исходного текста.