ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА НА ПРИМЕРЕ ОПРЕДЕЛЕНИЯ ВОПРОСОВ-ДУБЛИКАТОВ

Шлеменков А.А., Гусак Я.О.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, alex.shlemenkov@gmail.com

Abstract. A deep neural network architecture example was given in this paper. The neural network performance on specified data was measured. Further work and fields where such a system may be applicable were introduced.

Отслеживая тренды современного мира, можно заметить глубокий интерес к области искусственного интеллекта. Одной из его перспективных областей является естественная обработка языков (Natural Language Processing или NLP). Важно заметить, что NLP применима не только в сферах лингвистики и работы с языками, но и в сфере бизнеса. Примером такого употребления может служить реклама (анализ аудитории звезд).

Для исследования была выбрана задача по определению того, являются ли два вопроса с сайта Quora [1] смысловыми дубликатами. Важно отметить, что тексты вопросов имеют максимальную длину в 150 символов.

В связи со значительными успехами нейронных сетей в широком спектре задач (обработка изображений, звука, последовательностей) для реализации модели базовыми стали элементы LSTM (Long Short Term Memory или Долгой краткосрочной памяти) [2]. Структура ячеек LSTM сети представлена на рисунке 1.

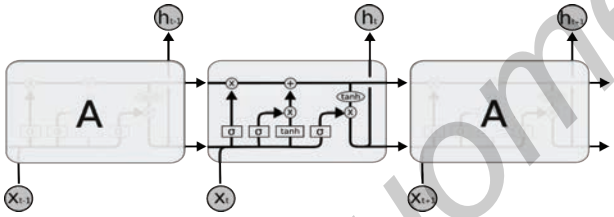

Рисунок 1 – Структура LSTM

LSTM – специальный вид рекуррентных сетей, предназначенный для моделирования долгосрочных зависимостей в данных. Их отличие от обычных RNN (рекуррентных нейронных сетей) в том, что их (LSTM) архитектура предполагает «запоминание» зависимостей на больший промежуток времени с помощью памяти, встроенной в LSTM-элемент, и нескольких «фильтров», которые позволяют, например, «забывать» данные из прошлого или «фильтровать» входные данные.

В качестве данных для обучения были предоставлены только тексты вопросов (пары) на английском языке. Для того, чтобы использовать их в качестве входа нейронной сети, необходимо было провести преобразование, иначе говоря, создать embedding (векторное представление). Так как текст является последовательностью слов, то логично создавать последовательность из их векторных представлений. Для этой задачи был выбран метод word2vec, предложенный в 2013 году как эффективный способ представления слов в векторном пространстве. В качестве модели word2vec была использована GoogleNews, обученная на корпусе размером в 3 миллиарда слов.

Архитектура сети, которая использовалась для определения слов-дубликатов, представлена на рисунке 2.

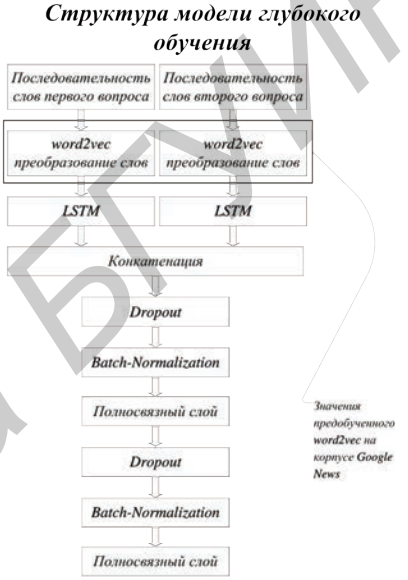

Рисунок 2 – Архитектура модели нейронной сети

В модели для каждого предложения строится его embedding, который после подается в LSTM. Выходы каждой из подсетей, которые обрабатывают вопрос, объединяются и передаются в двуслойную полносвязную сеть. Эта сеть возвращает вероятность того, что вопросы являются дубликатами.

Проверка качества модели на тестовой выборке дала результат по метрике accuracy около 86 %, а по метрике logloss 0.293. В данном случае результаты logloss являются более достоверными, т.к. классы в выборке не сбалансированы. Для сравнения, метрика logloss равная 0.69 имеет смысл случайного предсказания (с учетом баланса классов в выборке).

Таким образом, в результате исследования была разработана модель, которая по входной паре вопросов определяет вопросы-дубли. Такая система может быть использована в качестве инструмента автоматического определения вопросов-дублей и перенаправления пользователя на уже существующие страницы с вопросом, на который уже был дан ответ.

Литература

1. Quora question pairs – Kaggle Kaggle [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/c/quora-question-pairs.

2. Understanding LSTM Networks [Электронный ресурс]. – Режим доступа: http://colah.github.io/posts/2015-08-Understanding-LSTMs.