УДК 004.852

РУБРИКАЦИЯ ТЕКСТОВ С ПОМОЩЬЮ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА

© Рахметов И.П., Солдатова О.П.

Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Российская Федерация

e-mail: raxmetov2003@gmail.ru

Рубрикация текстов может осуществляться как вручную человеческим трудом, так и в автоматическом режиме на основе содержимого статьи. Простым в реализации и интуитивно понятным алгоритмом рубрикации является рубрикация текстов на основе частотного анализа ключевых слов, принадлежащих заранее известным рубрикам. Альтернативным способом является использование при рубрикации нейронных сетей, в частности многослойного персептрона, показывающего хорошие результаты при решении задач классификации.

В целях упрощения задачи было решено проводить рубрикацию текстов для двух рубрик: «Наука» и «Спорт». Тексты, принадлежащие данным рубрикам и использовавшиеся для обучения и тестирования рубрикаторов, были извлечены из датасета News dataset from Lenta.Ru [1], содержащего более восьми сотен тысяч новостей с новостного сайта Lenta.ru.

Каждый текст был лемматизирован с помощью морфологического анализатора с поддержкой снятия морфологической неоднозначности MyStem [2]. Из полученного списка лемм были удалены повторяющиеся элементы, стоп-слова (слова, не несущие смысловой нагрузки) [3] и имена [4], списки которых были найдены в сети Интернет. Затем для каждой рубрики были отобраны сто наиболее часто встречающихся слов, после чего эти слова были объединены в единый словарь, где каждому слову соответствовала одна из рубрик. Если слово принадлежало более чем одной рубрике, оно не включалось в словарь. Таким образом, был получен словарь ключевых слов для рубрик «Спорт» и «Наука».

Рубрикатор на основе частотного анализа ключевых слов был реализован в виде алгоритма на языке программирования Lua. Алгоритм выполняет обработку входного текста по алгоритму, аналогичному алгоритму генерации словаря, вплоть до этапа удаления стоп-слов и имен. Затем леммы из полученного списка ищутся в словаре ключевых слов. По результатам анализа текст относится к той рубрике, к которой принадлежит большинство найденных слов.

Для рубрикатора на основе многослойного персептрона входной текст необходимо представить в числовом виде. Для этого слова из словаря ключевых слов сортируются в алфавитном порядке (необязательно упорядочивать слова именно по алфавиту, но порядок должен сохраняться для любых используемых текстов). Затем входной текст обрабатывается тем же образом, как и для рубрикатора на основе частотного анализа. Наконец, для каждого слова из словаря подсчитывается количество его вхождений в обработанном тексте.

В результате для каждого входного текста получается вектор из чисел, каждое из которых больше или равно нулю, с длиной, равной количеству ключевых слов в словаре. При генерации обучающих и тестовых данных каждый вектор дополнялся унитарным кодом рубрики.

Из базы текстов было получено 100 000 векторов, по 50 000 для каждой рубрики. 2500 из них были использованы для обучения персептрона, а 97 500 – для его тестирования.

Архитектура многослойного персептрона состоит из двух скрытых слоев размерностью 16 и 4 нейрона соответственно и выходного слоя размерностью 2 нейрона.

Веса нейронов инициализированы с помощью инициализации Xe [5]. Все слои сети имеют функцию активации Softmax. В качестве обучающего алгоритма был выбран стохастический градиентный спуск. Обучение проводилось в течение 600 эпох с коэффициентом скорости обучения, равным 0,005.

Тестирование рубрикатора на основе метода частотного анализа ключевых слов проводилось на 100 000 текстах, которые после кодирования были использованы для обучения и тестирования персептрона. Из них 50 000 принадлежали рубрике «Спорт», другие 50 000 – рубрике «Наука».

Для рубрики «Спорт» 48 895 текстов были классифицированы корректно, 1105 были ошибочно отнесены к категории «Наука», следовательно, процент ошибки для данной категории равен 2,21 %.

Для рубрики «Наука» 48 716 текстов были классифицированы корректно, 1284 были ошибочно отнесены к категории «Спорт», процент ошибки равен 2,57.

Тестирование рубрикатора на основе нейронной сети проводилось на том же наборе из 100 000 текстов, однако 2500 из них были использованы для обучения нейронной сети, таким образом, они были исключены из тестового набора. Следовательно, тестовыми были только 97 500 текстов, 48 750 из которых принадлежали рубрике «Спорт», а другие 48 750 – «Наука».

Для рубрики «Спорт» 48 315 текстов были классифицированы корректно, 435 были ошибочно отнесены к категории «Наука», следовательно, процент ошибки для данной категории равен 0,89%.

Для рубрики «Наука» 48 408 текстов были классифицированы корректно, 342 были ошибочно отнесены к категории «Спорт», процент ошибки равен 0,70.

В результате рубрикатор на основе многослойного персептрона показал большую точность по сравнению с рубрикатором на основе частотного анализа текстов.

Однако при изменении словаря ключевых слов требуется генерация новых обучающих данных и переобучения всей сети (или даже изменения ее топологии). При этом рубрикатор на основе частотного анализа ключевых слов может быть использован для рубрикации текстов непосредственно после обновления словаря без каких-либо изменений.

Библиографический список

1. News dataset from Lenta.Ru. URL: https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-fromlenta (дата обращения: 19.05.2023).

2. Морфологический анализатор текста MyStem с поддержкой снятия неоднозначности. URL: https://yandex.ru/dev/mystem (дата обращения: 19.05.2023).

3. Список русских стоп-слов. URL: https://snipp.ru/seo/stop-ru-words (дата обращения: 19.05.2023).

4. База имен, написанных русскими буквами. URL: https://github.com/may-cat/rusnames (дата обращения: 19.05.2023).

5. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2018. 480 с.