УДК 004.912
DOI: 10.15827/0236-235X.126.263-267
Дата подачи статьи: 20.11.18
2019. Т. 32. № 2. С. 263–267

Сравнение нейросетевых моделей для классификации текстовых фрагментов, содержащих биографическую информацию

А.В. Глазкова [1], *к.т.н., старший преподаватель, a.v.glazkova@utmn.ru*

[1] *Тюменский государственный университет, г. Тюмень, 625003, Россия*

В работе представлены результаты сравнения нейросетевых моделей для решения задачи автоматической классификации текстов. Автор рассматривал задачу тематической классификации предложений, содержащих биографическую информацию. Эксперименты проводились на текстовом корпусе, собранном на основе биографических статей, размещенных в онлайн-энциклопедии Википедия, и снабженном семантической разметкой. Для каждого предложения в корпусе семантическая разметка указывает, содержит ли оно биографические сведения и к какому тематическому классу относятся эти сведения.

В статье описаны результаты для пяти нейросетевых архитектур: сети прямого распространения (feedforward network, FNN), рекуррентной нейронной сети (recurrent neural network, RNN), сети на основе управляемых рекуррентных блоков (gated recurrent unit, GRU), сети долгой краткосрочной памяти (long short-term memory, LSTM) и двунаправленной сети долгой краткосрочной памяти (bidirectional long short-term memory, BLSTM). Эти модели были обучены на векторных представлениях слов (word embeddings), построенных для текстов Википедии, и использованы для бинарной и мультиклассовой классификации предложений корпуса. В случае бинарной классификации целью обучения модели было определение того, содержит ли предложение биографическую информацию. При проведении мультиклассовой классификации на вход нейросетевых моделей подавались предложения корпуса, размеченные как содержащие биографическую информацию. В результате классификации требовалось отнести предложение к одному из тематических классов. Наилучшие результаты были достигнуты с помощью LSTM- и BLSTM-моделей. Полученные результаты можно использовать для автоматического поиска биографической информации в текстах, написанных на естественном языке.

Ключевые слова: *классификация текстов, нейронная сеть, обработка естественного языка, рекуррентная нейронная сеть, биографический факт.*

В работе описываются эксперименты по сравнению архитектур нейронных сетей для решения задачи классификации текстов на естественном языке, которая является одной из важных задач информационного поиска. Пути ее решения разнообразны, однако в последнее время на фоне общей тенденции использования механизмов машинного обучения в слабоформализуемых предметных областях все большую популярность приобретают методы, построенные на применении нейронных сетей.

Целью данной работы является сравнение широко используемых нейросетевых моделей на примере задачи тематической классификации фрагментов текстов. Были рассмотрены два типа классификации: бинарная, в ходе которой определялось, содержит ли данный фрагмент текста биографическую информацию, и мультиклассовая, уточняющая тематику биографической информации. Описанные в работе эксперименты проведены в рамках исследования по автоматическому извлечению биографической информации из текстов, написанных на естественном языке.

Методы

Нейросетевые модели. В данном исследовании использовались следующие архитектуры нейронных сетей:

– сеть прямого распространения (feedforward network, FNN);

– рекуррентная нейронная сеть (recurrent neural network, RNN);

– сеть на основе управляемых рекуррентных блоков (gated recurrent unit, GRU);

– сеть долгой краткосрочной памяти (long short-term memory, LSTM);

– двунаправленная сеть долгой краткосрочной памяти (bidirectional long short-term memory, BLSTM).

Нейронная сеть прямого распространения – это классическая модель нейронной сети, основанная на персептронах, которые были опи-

саны Ф. Розенблаттом [1] в конце 1950-х гг. Слои сети прямого распространения обычно являются полносвязными. Это означает, что все нейроны предшествующего слоя связаны со всеми нейронами текущего слоя, причем эти связи имеют прямое направление и не образуют циклов. Сети прямого распространения часто используются в практических приложениях за счет простоты организации, относительно скромных требований к вычислительным мощностям и довольно высокой точности решения многих типов задач. Архитектуры сетей, основанные на принципе прямого распространения, успешно применяются в различных прикладных областях [2–4].

Рекуррентные нейронные сети – искусственные нейронные сети, некоторые нейроны в которых соединены между собой циклически. Это ключевое отличие RNN от сетей прямого распространения позволяет рекуррентным нейронным сетям лучше справляться с решением задач, данные для которых подаются в виде последовательностей (например, задач, связанных с обработкой естественного языка). Рекуррентные связи позволяют сети хранить информацию о своих предыдущих состояниях и учитывать ее при прогнозировании выхода [5]. Основной трудностью при обучении классических рекуррентных сетей является проблема затухания градиента (vanishing gradient problem), возникающая вследствие использования алгоритма обратного распространения ошибки, когда величина градиента постепенно уменьшается в рекуррентных слоях. Современные рекуррентные сети, призванные решить проблему затухания градиента, – это сети долгой краткосрочной памяти [6] и сети на основе управляемых рекуррентных блоков [7]. Структура сети долгой краткосрочной памяти похожа на структуру классической RNN, однако рекуррентный слой LSTM-сети выполняет ряд дополнительных операций для хранения долгосрочных зависимостей [8]. Архитектуры GRU и двунаправленной LSTM являются модификациями LSTM.

В настоящее время модели нейронных сетей, включающие в себя слои долгой краткосрочной памяти, успешно применяются для решения задач обработки естественного языка и показывают в общем случае сравнимые результаты. Среди недавних работ, связанных с применением рекуррентных нейронных сетей для обработки текстов, можно назвать [9–12].

Представление текстов. Популярной моделью представления текстов в виде, пригодном для машинного обучения, является Word2-Vec – один из способов векторного представления слов, построенных на частоте взаимной встречаемости слов в пределах одного контекста [13].

Результаты, представленные в данной работе, были получены для моделей, обученных на русскоязычных текстах Википедии за 2018 год. В ходе обучения использовался алгоритм обучения Skip-gram, размер результирующего контекстного вектора для слова – 300.

Эксперимент

Данные. Нейросетевые модели были обучены на корпусе биографических текстов. Процесс построения корпуса подробно описан в работе [14].

Построение корпуса проводилось в два этапа. Сначала были автоматически собраны тексты биографических статей, размещенные в свободном доступе в Википедии, затем проведена обработка полученных текстов: удалены короткие тексты, содержащие информацию только об имени, годах жизни и роде деятельности персоны, из оставшихся текстов извлечен раздел «Биография», сделана тематическая разметка полученной текстовой коллекции.

В данной работе минимальным фрагментом текста, пригодным для тематической разметки, выбрано предложение. Текстовая коллекция была разбита на предложения, каждому из которых сопоставлена одна из следующих тематических групп:

– небиографические факты;

– события (личные и профессиональные – встречи, награждения и т.д.);

– неизменяемые личные характеристики: рождение (дата и место рождения), смерть (дата и место смерти), национальность, родители;

– изменяемые личные характеристики: место работы или службы, образование, семья (женитьба, замужество, дети), род занятий, место жительства или пребывания;

– прочие биографические факты.

Тематические группы выделены на основе таксономии биографических фактов, представленной в работе [15]. Под фактом подразумевается любое утверждение (в лингвистическом смысле), содержащееся в тексте в явном виде.

В итоге корпус содержит фрагменты (предложения), относящиеся к 12 классам: «Рождение», «Смерть», «Национальность», «Инфор-

мация о родительской семье», «Место работы», «Образование», «Семья», «Место жительства», «Личные события», «Профессиональные события», «Прочие биографические факты», «Не является биографическим фактом». Итоговый объем корпуса биографических текстов составил 200 текстов со средней длиной 19 предложений (225 слов).

В ходе бинарной классификации оценивалось, имеется ли в текстовом фрагменте биографическая информация. В качестве элементов класса, включающего в себя предложения, содержащие биографические факты, были приняты предложения текстового корпуса, имеющие соответствующую тематическую разметку (то есть все предложения, кроме помеченных как «Не является биографическим фактом»). Для создания класса фрагментов текста, не включающих в себя текстовую информацию, была использована выборка случайных небиографических статей из Википедии.

Мультиклассовая классификация проводилась на основании предложений, содержащих биографические факты. Из рассмотрения был исключен класс «Национальность» ввиду недостаточного количества обучающих примеров, поэтому классификация проводилась по 10 оставшимся классам.

Результаты. Для реализации нейросетевых моделей была использована библиотека Keras [16]. В качестве функций активации для рекуррентных сетей выбраны гиперболический тангенс на внутренних слоях и функция Softmax для выходного слоя. Для сети прямого распространения – логистическая функция на всех слоях. Размер обрабатываемых фрагментов данных (batch size) – 8. Использованный оптимизационный алгоритм – adaptive moment estimation (the Adam optimization). При обучении сетей проводилась дропаут-регуляризация с вероятностью 0,5. Количество нейронов в рекуррентных слоях варьировалось от 16 до 128, в сети прямого распространения – от 50 до 200 нейронов при глубине сети в 1-2 скрытых слоя. В итоге для каждой архитектуры были обучены несколько сетей, из которых по результатам на обучающей выборке выбрана модель, допущенная до экзамена на тестовой выборке.

Для оценки качества классификации были выбраны следующие метрики: точность (accuracy, количество совпадений фактического и прогнозируемого классов, в %) и F-мера (для бинарной классификации – по точности (precision) и полноте (recall) для класса, включающего в себя предложения, содержащие биографическую информацию; для мультиклассовой классификации – по средним показателям точности и полноты для всех классов).

Расчет точности классификации: $Accuracy = T/N$, где T – количество фрагментов, по которым классификатор принял верное решение; N – общее количество документов.

Расчет F-меры для конкретного класса:
$Precision = TP/(TP + FP)$,
$Recall = TP/(TP + FN)$,
$F\text{-}score = 2*Precision*Recall/(Precision + Recall)$,
где TP – истинно-положительное решение; FP – ложно-положительное решение; FN – ложно-отрицательное решение.

Оценки качества классификации для тестовой выборки приводятся в таблице. Жирным шрифтом выделены модели, имеющие лучшие показатели качества с точки зрения F-меры и точности классификации. Объем тестовой выборки для бинарной классификации – 200 предложений, для мультиклассовой классификации – 35 предложений.

Оценка качества классификации

Classification quality assessment

Тип классификации	**Архитектура сети**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-мера (%)**
Бинарная	FNN	86	86,01	93,89	89,78
	RNN	89,5	89,86	94,66	92,19
	LSTM	**91,5**	**92,18**	**95,66**	**93,89**
	GRU	91	90,65	96,18	93,33
	BLSTM	**91,5**	**91,99**	**95,42**	**93,63**
Мультиклассовая	FNN	73,91	85,14	77,45	81,11
	RNN	82,63	90,56	82,5	86,34
	LSTM	**94,36**	**95,94**	**93,33**	**94,63**
	GRU	86,96	86,43	79,57	82,86
	BLSTM	**92,3**	**97,43**	**94,17**	**95,77**

Заключение

В ходе работы было сделано сравнение качества тематической классификации текстовых фрагментов, проведенной при помощи пяти распространенных архитектур нейронных сетей. Наиболее точные результаты были получены с использованием LSTM- и BLSTM-моделей. Причем, если брать во внимание оценки рассмотренных моделей по F-мере, наилучшее качество в случае бинарной классификации было достигнуто LSTM-сетью, а мультиклассовой классификации – BLSTM-сетью.

В дальнейшем планируется проведение экспериментов по использованию различных спо-

собов представления текстов в векторном виде и автоматическому поиску текстовых фрагментов, содержащих биографическую информацию.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-37-00272 «Автоматизированное извлечение биографических фактов из текстов на естественном языке».

Литература

1. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 1958, vol. 65, no. 6, 386 p.

2. Генов А.А., Осипов В.В., Русаков К.Д., Моисеев А.А. Исследование сжатия растровых изображений с использованием искусственных нейронных сетей // Программные продукты и системы. 2018. № 3. С. 430–434. DOI: 10.15827/0236-235X.123.430-434.

3. Пекунов В.В. Извлечение информации из нейронных сетей прямого распространения в виде простых алгебраических моделей // Информационные технологии. 2017. Т. 23. № 1. С. 76–80.

4. Харахинов В.А., Сосинская С.С. Влияние сокращения размерности пространства признаков на результаты классификации листьев различных видов растений // Программная инженерия. 2018. Т. 9. № 2. С. 82–90. DOI: 10.17587/prin.9.82-90.

5. Le J. The 8 Neural Network Architectures Machine Learning Researchers Need to Learn. 2018. URL: https://www.kdnuggets.com/2018/02/8-neural-network-architectures-machine-learning-researchers-need-learn.html (дата обращения: 14.11.2018).

6. Hochreiter S., Schmidhuber J. Long short-term memory. Neural Computation, 1997, vol. 9, no. 8, pp. 1735–1780.

7. Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078. 2014. URL: https://arxiv.org/pdf/1406.1078.pdf (дата обращения: 17.11.2018).

8. Kostadinov S. Understanding GRU Networks. Towards Data Science. 2018. URL: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be (дата обращения: 14.11.2018).

9. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of neural architectures for sentiment analysis of Russian tweets. Proc. Intern. Conf. Dialogue 2016, 2016, pp. 50–58.

10. Ravuri S., Stolcke A. A comparative study of recurrent neural network models for lexical domain classification. Proc. Intern. Conf. IEEE (ICASSP). 2016, pp. 6075–6079. DOI: 10.1109/ICASSP.2016.7472844.

11. Yogatama D., Dyer C., Ling W., Blunsom P. Generative and Discriminative Text Classification with Recurrent Neural Networks. arXiv preprint arXiv:1703.01898. 2017. URL: https://arxiv.org/abs/1703.01898 (дата обращения: 17.11.2018).

12. Chen G., Ye D., Xing Z., Chen J., Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. Proc. IJCNN, IEEE, 2017, pp. 2377–2383. DOI: 10.1109/IJCNN.2017.7966144.

13. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. Proc. Conf. NIPS, 2013, vol. 26, pp. 3111–3119. URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (дата обращения: 17.11.2018).

14. Глазкова А. Формирование текстового корпуса для автоматического извлечения биографических фактов из русскоязычного текста // Современные информационные технологии и ИТ-образование. 2018. URL: http://it-edu.oit.cmc.msu.ru/index.php/SITITO/sitito2018/paper/view/724 (дата обращения: 17.11.2018).

15. da Costa dias Soares S.-F. Extraction of Biographical Information from Wikipedia Texts. 2011. URL: https://fenix.tecnico.ulisboa.pt/downloadFile/395143149863/thesis.pdf (дата обращения: 17.11.2018).

16. Keras: The Python deep learning library. URL: https://keras.io/ (дата обращения: 17.11.2018).

Software & Systems
DOI: 10.15827/0236-235X.126.263-267

Received 20.11.18
2019, vol. 32, no. 2, pp. 263–267

Comparison of neural network models for classifying text fragments containing biographical information

A.V. Glazkova [1], *Ph.D. (Engineering), Senior Lecturer, a.v.glazkova@utmn.ru*

[1] *Tyumen State University, Tyumen, 625003, Russian Federation*

Abstract. The paper presents the results of comparison of neural network architectures to solve the automatic text classification problem. The paper focuses on the topic classification of sentences containing

biographical facts. The data set was collected on the basis of Wikipedia and represented as a corpus of biographical texts with meta tagging. Each sentence in the corpus is marked according to whether it contains any biographical information and what type of information it contains.

The paper describes the results of topic classification of the corpus sentences for five neural network architectures. They are: feedforward network (FNN), recurrent neural network (RNN), gated recurrent unit RNN (GRU), long short-term memory RNN (LSTM) and bidirectional LSTM (BLSTM). These models were trained on the word embedding features obtained for the corpus sentences words. We have tried two types of classification: binary (depending on whether a sentence contain biographical information or not) and multi-class classification (definition of the type of biographical fact). LSTM and BLSTM models have shown the best results. The obtained results will be used to implement biographical information retrieval methods for natural language texts.

Keywords: text classification, neural network, natural language processing, recurrent neural network, biographical fact.

Acknowledgements. *The study has been supported by RFBR, research project no. 18-37-00272.*

References

1. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*. 1958, vol. 65, no. 6, 386 p.
2. Genov A.A., Osipov V.V., Rusakov K.D., Moiseev A.A. Research on compression of raster images using artificial neural networks. *Software & Sistems*. 2018, no. 3, pp. 430–434 (in Russ.).
3. Pekunov V.V. The Derivation of the information from artificial feed-forward neural networks in the form of simple algebraic models. *Information Technologies*. 2017, vol. 23, no. 1, pp. 76–80 (in Russ.).
4. Kharakhinov V.A., Sosinskaya S.S. The effect of dimension reducing on classification results of leaves of various plant species. *Software Engineering*. 2018, vol. 9, no. 2, pp. 82–90 (in Russ.).
5. Le J. *The 8 Neural Network Architectures Machine Learning Researchers Need to Learn*. 2018. Available at: https://www.kdnuggets.com/2018/02/8-neural-network-architectures-machine-learning-researchers-need-learn.html (accessed November 14, 2018).
6. Hochreiter S., Schmidhuber J. Long Short-term Memory. *Neural Computation*. 1997, vol. 9, no. 8, pp. 1735–1780.
7. Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. *Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation*. arXiv preprint arXiv:1406.1078. 2014. Available at: https://arxiv.org/pdf/1406.1078.pdf (accessed November 17, 2018).
8. Kostadinov S. *Understanding GRU Networks. Towards Data Science*. 2018. Available at: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be (accessed November 14, 2018).
9. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of neural architectures for sentiment analysis of Russian tweets. *Computational Linguistics and Intellectual Technologies: Proc. Intern. Conf. "Dialogue 2016"*. 2016, pp. 50–58 (in Russ.).
10. Ravuri S., Stolcke A. A comparative study of recurrent neural network models for lexical domain classification. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE Intern. Conf*. 2016, pp. 6075-6079.
11. Yogatama D., Dyer C., Ling W., Blunsom P. *Generative and Discriminative Text Classification with Recurrent Neural Networks*. arXiv preprint arXiv:1703.01898. 2017. Available at: https://arxiv.org/abs/1703.01898 (accessed November 17, 2018).
12. Chen G., Ye D., Xing Z., Chen J., Cambria E. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. *IEEE 2017 Intern. Joint Conf. on Neural Networks (IJCNN)*. 2017, pp. 2377–2383.
13. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. *Proc. Conf. NIPS*, 2013, vol. 26, vol. 26, pp. 3111–3119. Available at: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed November 17, 2018).
14. Glazkova A. Building a text corpus for automatic biographical facts extraction from Russian texts. *Modern Information Technologies and IT-Education*. 2018. Available at: http://it-edu.oit.cmc.msu.ru/index.php/SITITO/sitito2018/paper/view/724 (accessed November 17, 2018).
15. da Costa dias Soares S.-F. *Extraction of Biographical Information from Wikipedia Texts*. 2011. Available at: https://fenix.tecnico.ulisboa.pt/downloadFile/395143149863/thesis.pdf (accessed November 17, 2018).
16. *Keras: The Python Deep Learning Library*. Available at: https://keras.io/ (accessed November 17, 2018).