DOI: 10.36535/2022-9785945770829-20

ИСПОЛЬЗОВАНИЕ BERT ДЛЯ КЛАССИФИКАЦИИ КОРОТКИХ НАУЧНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Кусакин И.К., Цурупа А.М., Алмакаев А.В., Романов А.Ю.
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия,
ikkusakin@edu.hse.ru, amtsurupa@miem.hse.ru, avalmakaev@yandex.ru, a.romanov@hse.ru

В данной работе рассматриваются подходы к обучению классификаторов научных статей на основе BERT с целью реализации приложения для адаптации лучших моделей для последующего использования в инфраструктуре ВИНИТИ РАН. Для этого лингвистическая модель BERT была обучена на специализированном корпусе научных текстов для последующего использования в качестве встроенной части классификатора. В работе приведены результаты экспериментов по обучению моделей классификации научных статей по первому и второму уровням Российского государственного рубрикатора научно-технической информации (ГРНТИ).

Ключевые слова: *классификация текстов, искусственные нейронные сети, обработка естественного языка, Bidirectional Encoder Representations from Transformers, ruBERT.*

BERT FOR RUSSIAN SHORT SCIENTIFIC TEXTS CLASSIFICATION

Kusakin I.K., Tsurupa A.M., Almakaev A.V., Romanov A.Yu.
HSE University, Moscow, Russian Federation, ikkusakin@edu.hse.ru, amtsurupa@miem.hse.ru,
avalmakaev@yandex.ru, a.romanov@hse.ru

This work is devoted to the study of approaches for training BERT-based classifiers of scientific articles to implement the application with the adoption of the best models for use in the infrastructure of the VINITI RAS. For this purpose, the BERT linguistic model was trained on a specialized corpus of scientific texts for subsequent use as an embedding part of the classifier. The results of experiments carried out to train models for classifying scientific articles according to the first and second levels of the Russian State Rubricator of Science and Technical Information (SRSTI) are provided.

Keywords: *text classification, machine learning, artificial neural network, natural language processing, Bidirectional Encoder Representations from Transformers, ruBERT.*

Введение

В настоящее время сфера обработки естественного языка продолжает активно развиваться благодаря непрерывному совершенствованию средств обработки текстовых данных [1]. К ним относятся набравшие популярность в начале прошлого десятилетия методы векторизации BOW [2], TFIDF [3], и более совершенные с точки зрения семантической полноты Word2Vec [4] и FastText [5]. Кроме методов векторизации текста активно развиваются и сами методы машинного обучения, архитектуры моделей применяемых для решения задач классификации текстов, их суммаризации и генерации. Если в первой половине 2010 годов достаточно активно использовались классические алгоритмы машинного обучения, такие как Logistic Reggression [6], SVM [7], RandomForest [8] и Gradient Boosting [9], то в настоящее время акцент сместился в сторону

специализированных нейросетевых архитектур, таких как рекуррентная LSTM [10] и архитектура-трансформер BERT [11].

ВИНИТИ РАН обладает обширным массивом размеченных научных текстов, при этом ввиду постоянного увеличения потока документов, требуются автоматизированные средства классификации. Поэтому новизна данного исследования обуславливается сразу несколькими факторами. В настоящее время не существует качественного автоматизированного средства классификации научных текстов по кодам рубрикатора ГРНТИ. Также сложность задачи заключается в том, что рубрикатор представляет собой иерархическую структуру, в которой одна статья может относиться сразу к нескольким нодам в дереве научных работ. Основываясь на полученном ранее опыте обучения BERT была выдвинута гипотеза о том, что сеть BERT, которая будет обучена на лингвистические задачи предсказания пропущенных слов и правдоподобия следующего предложения по корпусе научных русскоязычных текстов покажет лучшее качество в целевой задаче классификации аннотаций научных статей по кодам ГРНТИ в сравнении со стандартной моделью, обученной на корпусе текстов на русском языке без заданного домена. Эта гипотеза базируется на результатах референсной работы по англоязычным научным статьям [13], а также результатам полученным ранее на русскоязычных статях с помощью классических алгоритмов машинного обучения [14].

Обучение лингвистической модели

Задача обучения лингвистической модели BERT представляет собой одновременно две подзадачи, на которые нейросеть одновременно обучается: Masked Language Model (MLM) и Next Sentence Prediction (NSP). После того как входные текстовые данные кодируются алгоритмом BPE [15], 15 % токенов в последовательности для задачи MLM маскируются следующим образом:

- 80% замаскированных токенов обозначаются как [MASK];
- 10% токенов заменяются на другие случайно взятые токены;
- 10% токенов остаются неизменными;

В задаче NSP случайным образом выбираются два предложения, где в 50% случаев токен B действительно является продолжением предложения B, а в остальных случаях представляет собой случайное предложение из корпуса текста. Последовательность, полученная из конкатенации A и B, подается на вход в модель, которая обучается на задачу предсказания совместимости предложений.

Подготовка данных для обучения лингвистической модели

Предоставленный ВИНИТИ РАН корпус из 2 миллионов научных текстов на первом этапе был вычищен от печатной разметки реферативных журналов. После этого для каждого текста было рассчитано отношение количества кириллических букв к латинским. В случаях, когда данный расчетный параметр был меньше 0,05 квантили, текст исключался из обучающегося корпуса. На следующем этапе из корпуса были исключены тексты, количество слов в которых оказалось меньше 0,01 квантили. В завершающий стадии подготовки данных для обучения лингвистической модели все символы, которые не являются буквами и цифрами, были выделены пробелами для корректной работы алгоритма токенизации текста BERT.

Результаты обучения лингвистической модели

После 80 эпох и трех месяцев обучения лингвистической модели было достигнуто качество модели по сумме кросс-энтропийной и логистической потерей, равное 0,93 [16]. В то же время базовая модель ruBERT, не обученная на домене научных текстов, показала на целевом корпусе научных текстов качество, равное 2,38.

Обучение лингвистической модели BERT на корпусе из 2 миллионов документов требует значительных временных затрат, поэтому эксперименты с изменением гиперпараметров происходили по ходу обучения, и, в случае ухудшения качества, веса модели и оптимизатора возвращались к предыдущей точке сохранения.

Анализ и предобработка данных для обучения классификатора

Предоставленный ВИНИТИ РАН набор данных для классификации по 1 и 2 уровням ГРНТИ представляет собой набор из 569928 документов, которые состоят из заголовка статьи, аннотации, ключевых слов и самих кодов рубрикатора ГРНТИ. Всего в предоставленной выборке содержатся 52 класса 1 уровня рубрикатора и 481 класса второго уровня.

На первом этапе датасет текстов был очищен от печатной разметки и спецсимволов ВИНИТИ, а также формул LaTeX и электронных адресов. Далее из набора данных было удалено 51657 повторяющихся статей, 34465 статей с отношением кириллических символов к латинским менее 0,2, а также 6787 статей с длинной аннотации менее 15 слов. Дополнительный парсинг официального сайта ГРНТИ [17] помог выделить и удалить несуществующие в действительности классы на 1 и 2 уровнях рубрикатора.

Каждый документ датасета может одновременно относиться к нескольким кодам ГРНТИ, причем среднее, максимум и дисперсия количества ответов составляет 1,26, 7, 0,25 для 1-го уровня и 1,33, 8, 0,36 для 2 уровня соответственно.

После всех этапов препроцессинга данных в датасете осталось 477018 статей. Распределение частотностей классов которых представлено на Рис. 1 и 2. Выяснилось, что значительное количество классов содержит количество статей, недостаточное для обобщения генеральной совокупности рубрики. Исходя из этого, было решено отбросить малочисленные классы по определенному порогу. Для обучения и замера качества классификации корпус для обучения модели на целевую задачу был разделен на тренировочную и тестовую выборки с соотношением количества объектов четыре к одному.

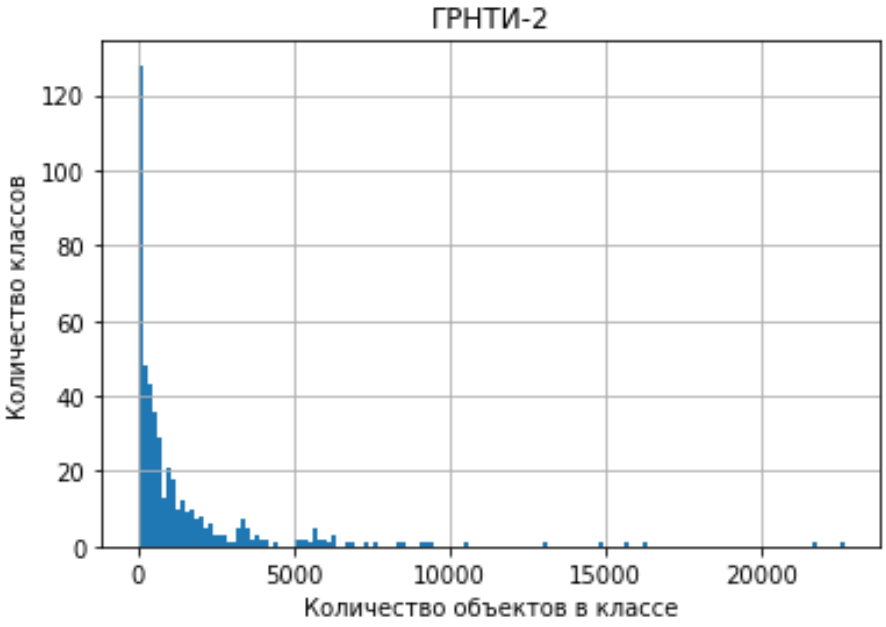

Рис. 1 Зависимость количества классов от их объема второго уровня ГРНТИ

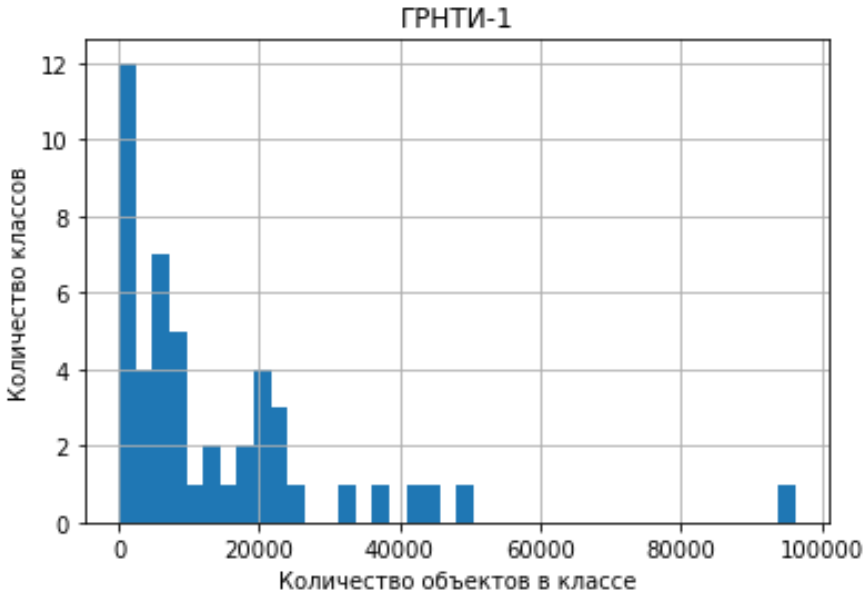

Рис. 2 Зависимость количества классов от их объема первого уровня ГРНТИ

Параметры обучения на целевую задачу

В экспериментах с обучением на целевую задачу все гиперпараметры имели зафиксированные значения кроме коэффициента шага обучения, который подбирался при помощи алгоритма LR Finder.

Конфигурация гиперпараметров:

– Оптимизатор Adam с параметрами adafactor=False, adam_beta1=0.9, adam_beta2=0.999, adam_epsilon=1e-08;

– Batch size равный 8 (при больших значениях веса модели и матрицы градиентов не умещаются в 16 Гб видеопамяти);

– Во время обучения градиент обрезался, если его норма была выше 5;

– Бинарная кросс-энтропия как функционал ошибки модели.

Результаты обучения лингвистических моделей

В Таблицах 1 и 2 представлены результаты оценки качества классификации в зависимости от выбора гиперпараметров для 2-го и 1-го уровней ГРНТИ соответственно. Среди рассматриваемых параметров:

1. Порог – минимальное количество статей в классе;
2. LM – используемая лингвистическая модель в классификаторе, где 1 – RuBERT, 2 – BERT, обученная на целевом корпусе научных текстов;
3. Обучение весов LM – параметр обучаемости весов BERT при обучении классификатора;
4. Линейные слои – архитектура классификатора после входов BERT;
5. Skip-connection – наличие skip-connection механизма в архитектуре;
6. Label smoothing – параметр сглаживания истинных ответов, где false – не использовать его, default – использовать стандартный подход, и custom – использовать собственный подход;
7. Dropout – параметр dropout, применяемый к выходам BERT и линейных слоев.

Таблица 1

Качество классификации второго уровня ГРНТИ при различных гиперпараметрах

Порог	LM	Обучение весов LM	Добавление названий	Линейные слои	skip-connection	Label smoothing	Dropout	macro F1	micro F1	weighted F1
700	1	true	False	[768×210]	false	false	0,2	0,617	0,649	0,651
	1	false	False	[768×210]	false	false	0,2	0,524	0,607	0,600
	2	true	false	[768×210]	false	false	0,2	0,654	0,711	0,714
	2	false	false	[768×210]	false	false	0,2	0,582	0,621	0,620
	2	true	true	[768×210]	false	false	0,2	0,670	0,728	0,730
	2	true	true	[768×210]	false	false	0	0,642	0,705	0,703
	2	true	true	[768×512×210]	false	false	0,2	0,673	0,729	0,731
	2	true	true	[768×512×384×210]	false	false	0,2	0,658	0,714	0,719
	2	true	true	[768×512×384,384+768×210]	true	false	0,2	0,643	0,702	0,705
	2	true	true	[768×512×210]	false	default	0,2	0,685	0,737	0,733
	2	true	true	[768×512×210]	false	custom	0,2	0,696	0,748	0,746
	2	true	true	[768×512×384,384+768×210]	true	custom	0,4	0,622	0,672	0,676
	2	true	true	[768×210]	false	custom	0,2	0,689	0,741	0,739
200	2	true	true	[768×512×210]	false	custom	0,2	0,615	0,700	0,697
	2	true	true	[768×210]	false	custom	0,2	0,618	0,705	0,703

Таблица 2

Качество классификации первого уровня ГРНТИ при различных гиперпараметрах

Порог	LM	Обучение весов LM	Добавление названий	Линейные слои	skip-connection	Label smoothing	Dropout	macro F1	micro F1	weighted F1
700	2	true	true	[768×37]	false	false	0,2	0,775	0,807	0,806
	2	true	true	[768×256×37]	false	false	0,2	0,771	0,799	0,797
	2	true	true	[768×256×64×37]	false	false	0,2	0,762	0,785	0,787
	2	true	true	[768×37]	false	default	0,2	0,776	0,805	0,806
	2	true	true	[768×37]	false	custom	0,2	0,788	0,819	0,819
	2	true	true	[768×256×37]	false	custom	0,2	0,783	0,811	0,810
	Лучшая модель для ГРНТИ 2							0,769	0,804	0,804
200	2	true	true	[768×256×42]	false	custom	0,2	0,739	0,795	0,796
	2	true	true	[768×42]	false	custom	0,2	0,741	0,798	0,799
	Лучшая модель для ГРНТИ 2							0,743	0,791	0,791

Выводы

Таким образом, была подтверждена гипотеза о том, что при использовании лингвистической модели, обученной для конкретной задачи, можно повысить точность классификации по сравнению с моделью общего назначения. Анализ полученных результатов позволяет сделать следующие выводы:

1. Обучение лингвистической модели BERT на домене научных русскоязычных текстов дает значительный прирост в качестве классификации на целевой задаче в сравнении с моделью RuBERT, обученной на обобщенном корпусе текстов на русском языке;
2. Качество классификации улучшается, если помимо выходных линейных слоев модели дообучать глубинные слои трансформера;
3. Эффективность классификации возрастает, если на вход подавать конкатенированное название статей и аннотацию;
4. Подбор оптимальных порогов для каждого класса значительно улучшает результаты классификации;
5. Увеличение количества и размерности линейных слоев, skip-connection соединения и вариации с функцией активации не оказывают существенного влияния на качество классификации;
6. Использование метода label smoothing существенно увеличивает скорость обучения и показывает небольшой прирост качества классификации;
7. Зависимость качества классификации по метрике F1 от размера класса имеет нисходящий тренд на рубриках, имеющих менее 500 объектов;
8. Использование выходов классификатора второго уровня ГРНТИ для предсказания первого показывает метрики качества, сравнимые с метриками классификаторов первого уровня.

Список использованной литературы

1. Strubell E., Ganesh A., McCallum A. Energy and policy considerations for deep learning in NLP // arXiv preprint arXiv:1906.02243. – 2019.
2. Zhang Y., Jin R., Zhou Z. H. Understanding bag-of-words model: a statistical framework // International Journal of Machine Learning and Cybernetics. – 2010. – Vol. 1. – № 1-4. – P. 43-52.
3. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. – Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.
4. Goldberg Y., Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method // arXiv preprint arXiv:1402.3722. – 2014.
5. Athiwaratkun B., Wilson A. G., Anandkumar A. Probabilistic fasttext for multi-sense word embeddings // arXiv preprint arXiv:1806.02901. – 2018.
6. Wright R. E. Logistic regression. – 1995.
7. Noble W. S. What is a support vector machine? // Nature biotechnology. – 2006. – Vol. 24. – № 12. – P. 1565-1567.
8. Belgiu M., Drăguţ L. Random forest in remote sensing: A review of applications and future directions // ISPRS journal of photogrammetry and remote sensing. – 2016. – Vol. 114. – P. 24-31.
9. Friedman J. H. Stochastic gradient boosting // Computational statistics & data analysis. – 2002. – Vol. 38. – № 4. – P. 367-378.
10. Mikolov T. et al. Recurrent neural network based language model // Eleventh annual conference of the international speech communication association. – 2010.
11. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213. – 2019.

12. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. – 2018.

13. Beltagy I., Lo K., Cohan A. SciBERT: A pretrained language model for scientific text // arXiv preprint arXiv:1903.10676. – 2019.

14. Romanov A., Lomotin K., Kozlova E. Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts // Data Science Journal. –2019. – Vol. 18. – № 1. – P. 1-17.

15. Bostrom K., Durrett G. Byte pair encoding is suboptimal for language model pretraining // arXiv preprint arXiv:2004.03720. – 2020.

16. MIEM SciBERT – an open-source Russian-science texts linguistic model. – URL: https://github.com/IlyaKusakin/miem-scibert-project

17. Государственный Рубрикатор НТИ России. – URL: http://scs.viniti.ru/rubtree/main.aspx?tree=RGNTI