ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА “МИФИ”, 2020, том 9, № 3, с. 236–244

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.91

ПРИМЕНЕНИЕ МУЛЬТИЗАДАЧНОЙ МОДЕЛИ ДЛЯ ПРАКТИЧЕСКИХ ЗАДАЧ ГЕНЕРАЦИИ ЗАГОЛОВКА, ОПРЕДЕЛЕНИЯ ЛЕММ И КЛЮЧЕВЫХ СЛОВ

© 2020 г. И. А. Молошников[1,*], А. В. Грязнов[1,], Д. С. Власов[1,***], Р. Б. Рыбка[1,****], А. Г. Сбоев[1,2,*****]**

[1] Национальный исследовательский центр “Курчатовский институт”, Москва, 123182, Россия

[2] Национальный исследовательский ядерный университет “МИФИ”, Москва, 115409, Россия

**e-mail: ivan-rus@yandex.ru*

***e-mail: artem.official@mail.ru*

****e-mail: vfked0d@gmail.com*

*****e-mail: rybkarb@gmail.com*

******e-mail: sag111@mail.ru*

Поступила в редакцию 23.06.2020 г.
После доработки 23.06.2020 г.
Принята к публикации 07.07.2020 г.

В работе проведено комплексное исследование эффективности мультизадачных методов глубокого обучения комбинированных нейросетевых моделей на выбранном наборе задач: генерации заголовков, определения лемм и ключевых слов. Мультизадачная модель строилась на основе слоев Multi-head Attention, на основе нее разрабатывалась генеративная модель для формирования заголовков и модель на основе слоев LSTM для лемматизации. В работе использовались открытые корпуса: РИА Новости, содержащий тексты новостей и заголовки к ним и корпус с морфологической, синтаксической разметкой и леммами СинТагРус из проекта universaldependencies. Для задачи выделения ключевых слов авторами был собран новый корпус на базе новостных текстов с использованием краудсорс платформы. По результатам работы показано повышение точности на 1% в F1 мере для задачи лемматизации при использовании признаков из мультизадачной модели по сравнению с использованием только признаков морфологии, достигнуты state-of-art точности для задачи генерации заголовка (0.42 ROUGE F1 мере). На основе модели для генерации заголовков, полученной в данной работе, предложен алгоритм выделения ключевых слов без дополнительного обучения сети.

Ключевые слова: мультизадачная модель, генерация заголовков, лемматизация, выделение ключевых слов

DOI: 10.1134/S2304487X20030074

ОБЗОР ИСПОЛЬЗУЕМЫХ ПОДХОДОВ

В большинстве задач анализа естественного языка используются различные инструменты предобработки текстов.

Одной из функций предобработки данных является лемматизация – это процесс приведения слова к нормальной форме, что позволяет в статистических моделях учитывать одно и то же слово в различных вариантах его написания. Существуют готовые решения, которые построены на различных методах: словарный (Pymorphy [1]), правила и регулярные выражения (Pattern [2]), машинное обучение (UDPipe [3], NLTK [4]) и другие, но все они опираются только на прямые признаки – саму словоформу и на морфологические признаки. Однако при реальной работе, морфологические признаки могут содержать ошибки. Для улучшения работы этих решений возможно использовать глубокие нейросетевые модели, предварительно обученные на большом объеме неразмеченных данных [5, 6]. При их обучении может использоваться обучение на несколько задач (пример BERT – языковая модель и модель определения, что два предложения идут друг за другом). В нашей работе мы рассматриваем возможность построить модель на дополнительных признаках мультизадачной модели BERT (см. ниже), так как они будут устойчивы и на тренировочной части, и на тестовой, а также при реальной работе, поскольку получены на большом объеме данных.

Описанный выше подход может применяться и к задачам резюмирования текста, т.е. получения для входного текста (последовательности слов) более короткой последовательности, как можно более полно отражающей информацию, которая содержится во входном тексте. При этом выходная последовательность, как и входная, должна быть грамматически связанным текстом. Частным случаем задачи резюмирования является генерация заголовков к новостным статьям.

Сложность задачи заключается в том, что не существует объективной меры того, насколько слово характеризует текст – любой выбор слов является субъективным по причине разницы восприятия текстовой информации у разных людей.

Существует два способа решения задачи генерации заголовка: экстрактивный [7–9] и генеративный [10–15]. В первом случае выходная последовательность (заголовок) является подпоследовательностью входной. То есть ставится задача поиска в тексте фрагмента, который максимально соответствует истинному заголовку. При генеративном подходе заголовок генерируется моделью из слов, последовательно взятых из словаря, и может включать в себя слова, которые не содержит входная последовательность.

Другим вариантом задачи резюмирования текста является выделение ключевых фраз: определение фраз или слов текста, которые наилучшим образом характеризуют текст, т.е. являются наиболее важными с точки зрения его описания.

Для проверки разрабатываемых алгоритмов обычно используют данные, в которых авторами заранее заданы ключевые слова, например: новостные статьи и аннотации научных публикаций. При таком подходе связь заданных автором слов и текста может быть опосредована, отражать в большей степени личное мнение автора, служить целям, отличным от передачи информации (например, взаимодействие с поисковыми движками или отнесение текста к категории). Одно из возможных решений данной проблемы – сбор корпуса текстов и организация процесса выделения эталонных ключевых слов на основе агрегации данных опроса респондентов-людей.

Проблема автоматического выделения ключевых слов из текста довольно хорошо представлена в литературе. Можно выделить следующие подходы к задаче автоматического выделения ключевых слов из текста: статистический подход, подход на основе графов, машинное обучение.

Задача автоматического выделения ключевых слов может быть решена на основе следующих подходов: статистического (RAKE [16], KPMiner [17], YAKE [18], SBE [19]), подхода на основе графов (TextRank [20], SingleRank [21], TopicRank [22], PositionRank [23], MultipartiteRank [24]), подхода на основе машинного обучения (KEA [25], WINGNUS [26], GenEx [27]).

Статистический подход предполагает определение ключевых слов на основе статистических показателей текста – частоты встречаемости слов, их распределения по документам из выборки, положения слова в тексте, совместной встречаемости слов.

Подход на основе графов предполагает представление текста или его элементов в виде графа. В качестве узлов обычно выступают слова или группы слов, взвешенные связи между ними строят на основе совместной встречаемости в окне фиксированной размерности. Далее используются алгоритмы для ранжирования узлов графа (как правило, PageRank [28] или его модификации).

Подход на основе машинного обучения включает в себя несколько методов обучения с учителем, то есть внутренние коэффициенты модели предварительно настраиваются на основе обучающей выборки, в которой содержатся тексты с уже определенными эталонными ключевыми словами.

МУЛЬТИЗАДАЧНАЯ МОДЕЛЬ BERT

BERT представляет собой уже обученную модель на основе слоев multi head attention, она представлена в работе [6]. Обучение проводилось на корпусе дампов Википедии для 104 языков. При формировании корпуса, использовалось сглаживание по представительности текстов для каждого языка. Модель состоит из 12 слоев, размерность для multi head attention 768 и скрытого слоя 3072, 12 голов в multi head attention.

В процессе обучения BERT решались одновременно две задачи:

– маскируемая языковая модель;

– определение, что два предложения идут последовательно в тексте.

Маскируемая языковая модель

Для 15% токенов, случайно отбираемых каждую эпоху, модель предсказывает на выходе этот токен, при этом сам токен заменяется случайным образом на один из следующих: в 80% случаев на специальный токен-маску [MASK]; в 10% случаев заменяется на случайный токен из словаря; в 10% случаев не заменяется.

Такая замена объясняется тем, что при работе в фазе предсказания токенов [MASK] не будет и для нормальной работы модель должна встречаться с такими примерами при обучении.

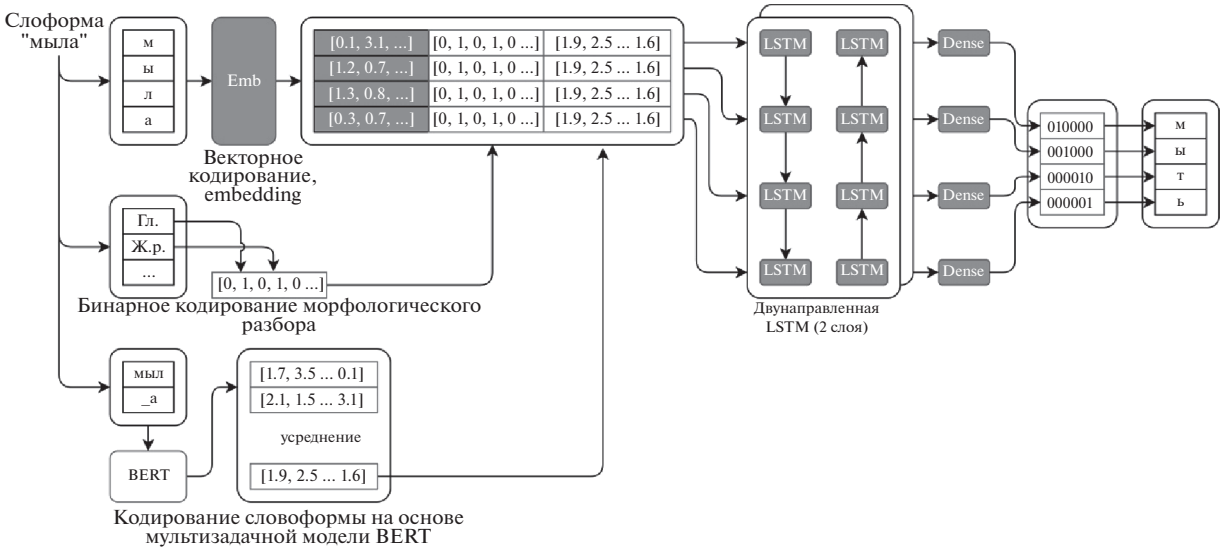

Рис. 1. Схема модели для задачи лемматизации.

Определение, что два предложения идут последовательно в тексте

Для этой задачи документы делятся на предложения, в модель на вход, как одна последовательность, подаются два предложения. Предложения разделяются специальным токеном [SEP]. В начале входной последовательности добавляется специальный токен [CLS] – на выходе модель для него должна дать 1 в случае, если два предложения идут друг за другом в одном документе, или 0, если они из разных документов или частей одного документа. При обучении подается 50% процентов последовательных предложений и 50% случайно перемешанных.

МОДЕЛЬ ДЛЯ ЛЕММАТИЗАЦИИ ТЕКСТОВ

Модель для лемматизации схематично представлена на рис. 1. Она состоит из следующих слоев:

1. Входы сети, их может быть до 3-х:

a. посимвольное представление словоформы, символы кодируются слоем векторного представления (Embedding, размерностью *s*);

b. полный морфологический тэг в бинарной кодировке (опционально);

c. активности из мультизадачной модели BERT, т.к. словоформа может быть представлена несколькими токенами в модели BERT, то берется усреднение выходов по этим токенам (опционально).

2. Признаки морфологии и векторное представление словоформы из BERT копируется несколько раз, по длине слоформы, и конкатенируется с каждым символом словоформы.

3. Далее идет два слоя двунаправленного LSTM, каждый слой имеет размерность $2 \times s$.

4. На выходе, каждая активность декодируется в символ полносвязным слоем с функцией активации softmax, размерность слоя по числу символов в словаре.

Размерность слоев *s* подбирается эмпирически, мы пробовали несколько вариантов: 32, 128 и 256.

Модель для генерации заголовков

Модель нейросети для генерации заголовков состоит из двух частей – энкодер и декодер (рис. 2, блоки энкодера имеют префикс “E_T.”, а декодера “D_T.”). Энкодеру на вход поступает текст (текст разбивается на токены – части слов, пример: слово “студентка” будет представлено двумя токенами: “студент” и “_ка”) и он преобразует его в последовательность векторов для каждого элемента текста. Декодер получает на вход результат работы энкодера и сгенерированную на текущий момент часть заголовка (на первом шаге это специальный токен “[CLS]”, означающий начало последовательности) (см. рис. 2). Заголовок генерируется по токенам, пока сеть не выдаст специальный токен ⟨T⟩ – конец заголовка или не будет достигнута максимальная заданная длина (150 токенов) (подробнее в работе [29]).

МОДЕЛЬ ДЛЯ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ

Предложен метод выделения ключевых слов на основе готовой нейронной сети для задачи генерации заголовка (описанной ранее). Особенность метода заключается в отсутствии обучения

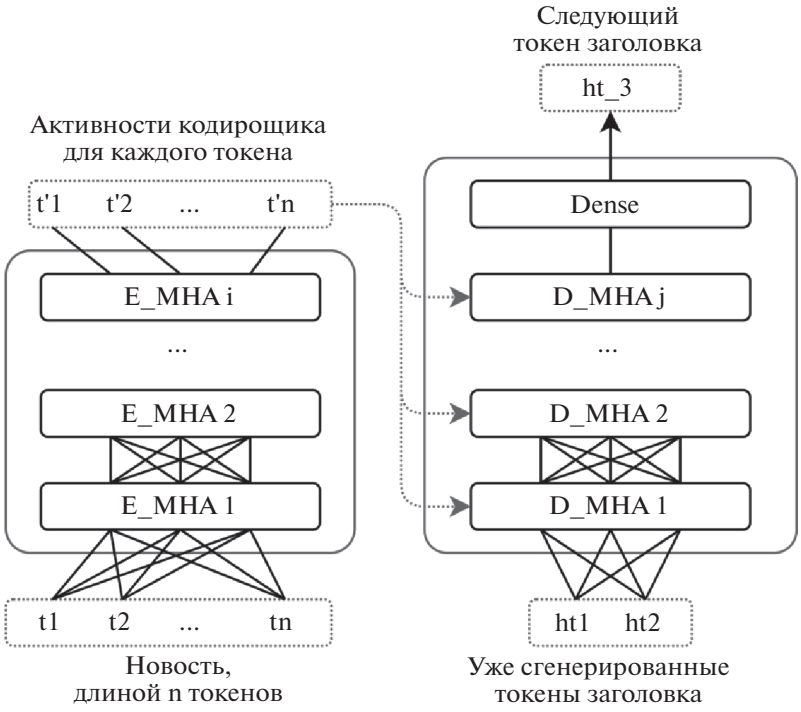

Рис. 2. Схема сети для генерации заголовка.

модели на выделение ключевых слов, вместо этого анализируются активности готовой нейронной сети, обученной на схожей задаче. Гипотеза заключается в том, что после обучения модели генерации заголовков ее внутренние веса, сопоставленные со входными токенами, отражают, какие токены внесли больший вклад в генерацию заголовка, соответственно, какие токены (а значит и слова) наиболее информативны, а задача выделения ключевых слов и генерации заголовка похожие и направлены на краткое резюмирование текста.

Разработанный метод включает в себя следующую последовательность действий:

1) веса последнего слоя внимания (multihead attention) декодера сохраняются, так как являются входными данными алгоритма – получается матрица A размерности $|t| \times m \times n$, где $|t|$ – число токенов входной последовательности, m – число голов слоя внимания (multihead attention), n – число токенов сгенерированного заголовка.

2) для каждого токена t_i рассчитывается сумма весов по головам – получается матрица A' размерности $|t| \times n$;

3) для каждого токена t_i рассчитывается сумма весов по словам заголовка, получается вектор оценок токенов v_{score} размерности $|t|$;

4) для каждого элемента w^j множества слов исходного текста W определяется множество токенов t_w^j, которыми было представлено слово после токенизации текста;

5) для каждого слова w^j получается максимальная оценка на основе набора токенов, которыми слово представлено:

$$score(w^j) = max(t_w^j);$$

6) оценки слов переводятся в шкалу действительных чисел от 0 до 1 по следующей формуле:

$$norm(score(w^j)) = \frac{score(w^j) - min(score(W))}{max(score(W)) - min(score(W))};$$

7) из набора слов W исключаются стоп-слова, производится очистка от символов (чтобы в полученных из текста словах остались только буквы);

8) для оставшихся слов находятся леммы. Если леммы разных слов совпадают (например, такое может быть у слова в разных формах), лемма получает максимальную оценку из представленных;

9) в качестве ключевых слов выбирается 25% слов, имеющих наибольший рейтинг.

ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Оценка модели лемматизации проводилась на корпусе ru_syntagrus из Universal Dependencies v2.2, который использовался на соревновании CoNLL 2018, оценка проводилась скриптом с соревнований. Результаты приведены в табл. 1.

Таблица 1. Результаты оценки модели для лемматизации на корпусе Universal Dependencies 2.2 ru_syntagrus

Номер эксп.	F1	max epochs	s, размер модели	patience	use BERT	use morph
1	87.39	150	32	15	True	False
2	93.42	150	32	15	True	True
3	92.16	150	32	15	False	True
4	85.77	30	32	30	True	False
5	89.74	30	32	30	True	True
6	88.21	30	32	30	False	True
7	95.14	30	128	30	True	False
8	97.70	30	128	30	True	True
9	97.50	30	128	30	False	True
10	95.78	30	256	30	True	False
11	**98.21**	**30**	**256**	**30**	**True**	**True**
12	97.66	30	256	30	False	True
[30]	98.19	–	–	–	False	True

Пояснения к таблице: F1 – точность, полученная официальным скриптом на тестовом множестве, max epochs – максимальное число эпох, которым модель могла обучаться до раннего останова (когда ошибка на валидационном множестве начинает расти), s – размер модели (см. в разделе “Модель для лемматизации текстов”), patience – число эпох, в течение которых может расти ошибка на валидационном множестве до остановки обучения, use BERT – флаг, использовались или нет признаки из мультизадачной модели, use morph – флаг, использовались или нет признаки морфологии. Для борьбы с переобучением использовался ранний останов по валидационному множеству и загрузка весов на лучшую эпоху.

Для оценки модели генерации заголовка используется метрика Rouge-L-f, основанная на поиске самой длинной общей последовательности токенов [31], результаты оценки на корпусе РИА “Новости” приведены в табл. 2, для обучения, валидации и тестирования использовалась разбивка корпуса из статьи [32].

Число параметров рассчитывалось без учета эмбеддинг слоев и выходного слоя, так как используемая мультиязычная модель BERT имеет большой словарь, что значительно увеличивает количество параметров, однако большая часть мультиязычного словаря не встречается в корпусе.

Для оценки методов выделения ключевых слов был собран валидационный корпус текстов на основе набора данных новостей “РИА новости“ с использованием краудсорсинг-платформы. Корпус включает в себя 50 текстов средней длины 2002 символа. Эталонные ключевые слова для каждого текста выбирались на основе усреднения оценок десяти пользователей.

В качестве метрик оценки методов выделения ключевых слов используется адаптация стандартных метрик *precision*, *recall* и *f1-меры*.

Формулы, по которым вычисляются метрики, представлены ниже, обозначения:

• *words* – множество слов текста, которые являлись кандидатами для оценки на краудсорсинговой платформе;

• kw_{valid} – ключевые слова, которые были определены респондентами на краудсорсинг-платформе, то есть те слова, средняя оценка которых

Таблица 2. Результаты оценки моделей генерации заголовка на корпусе РИА “Новости”

Номер эксп.	Начальная инициализация BERT	Число слоев энк.	Число параметров	Число эпох	rouge-1	rouge-2	rouge-l
1	да	3	48M	3	0.3971	0.2145	0.3735
2	да, замороженные	3	48M	3	0.3587	0.1794	0.3374
3	мультизадач.	3	48M	3	0.0793	0.0016	0.0743
4	мультизадач., заморож.	3	48M	3	0.0702	0.0049	0.0687
5	нет	3	48M	3	0.3673	0.1883	0.3463
6	**да**	**6**	**69M**	3	0.4096	0.2230	0.3867
				6	0.4261	0.2412	0.4013
				13	0.4420	0.2566	0.4156
				23	**0.4453**	**0.2605**	**0.4180**
7	нет	6	69M	3	0.1426	0.0338	0.1346
8	**да**	**12**	**112M**	3	0.4230	0.2381	0.3995
				8	0.4472	0.2625	0.4220
				13	**0.4487**	**0.2654**	**0.4235**
Первое предложение	–	–	0	–	0.2590	0.1113	0.2327
Сторонние реализации							
***PBA-Tranformer [32]**	**нет**	–	**63M**	**10**	**0.4296**	**0.2543**	**0.4002**

Таблица 3. Результаты оценки методов выделения ключевых слов на валидационном корпусе

Метод	Avg. Precision	Avg. Recall	Avg. f1-score
KPMiner	0.57	0.62	0.56
SBE [19]	0.38	0.97	0.53
Основанный на нейронной сети	**0.45**	**0.51**	**0.46**
Position Rank	0.47	0.49	0.45
Term Frequency	0.33	0.71	0.43
RAKE	0.29	0.94	0.42
YAKE	0.34	0.32	0.31
Multipartite Rank	0.33	0.31	0.30
Single Rank	0.43	0.25	0.29
Text Rank	0.35	0.16	0.20

по шкале от 1 до 3 составила больше 2.5 (на основе мнения десяти респондентов);

• $kw_{nonvalid}$ – слова, у которых средняя оценка по шкале от 1 до 3 на основе мнения респондентов составила меньше 2.5.

$$kw_{nonvalid} = words \backslash kw_{valid};$$

• kw_{method} – слова, которые были отнесены к ключевым на основе метода выделения ключевых слов, то есть которым метод присвоил рейтинг выше 75% квантили. В случае, если метод выделяет словосочетания, они разбиваются на отдельные слова, каждому из которых присваивается рейтинг словосочетания;

• $kw_{nonmethod}$ – слова, у которых рейтинг на основе метода имеет значение ниже 75% квантили.

$$kw_{nonmethod} = words \backslash kw_{method};$$

• tp – число истинно положительных примеров;

• fp – число ложно положительных примеров (ошибка второго рода);

• fn – число ложно отрицательных примеров (ошибка первого рода).

Для текста i:

$$tp_i = |kw^i_{method} \cap kw^i_{valid}|;$$

$$fp_i = |kw^i_{method} \cap kw^i_{nonvalid}|;$$

$$fn_i = |kw^i_{nonmethod} \cap kw^i_{valid}|;$$

$$precision_i = \frac{tp_i}{(tp_i + fp_i)}; \quad recall_i = \frac{tp_i}{(tp_i + fn_i)};$$

$$f1_i = \frac{2 \times precision_i \times recall_i}{(precision_i + recall_i)}.$$

Агрегированная оценка для метода:

$$precision_{avg} = \frac{\sum_{i=1}^{|precision|} precision_i}{|precision|};$$

$$recall_{avg} = \frac{\sum_{i=1}^{|recall|} recall_i}{|recall|};$$

$$f1_{avg} = \frac{\sum_{i=1}^{|f1|} f1_i}{|f1|}.$$

В табл. 3 представлены результаты оценки методов выделения ключевых слов на собранном валидационном корпусе.

ВЫВОДЫ

В работе проведено комплексное исследование эффективности мультизадачных методов глубокого обучения комбинированных нейросетевых моделей на выбранном наборе задач: генерации заголовков, определения лемм и ключевых слов. В результате:

1) в задаче определения лемм добавление новых признаков к модели нейросети дает прирост на 1% по метрике F1 по сравнению со стандартным набором признаков;

2) разработанный метод резюмирования текста на задаче генерации заголовков при валидации на корпусе русскоязычных статей проекта “РИА новости” позволил достичь State-of-the-art точности 0.42 по метрике ROUGE-L, превосходит аналоги на 2%;

3) для задачи выделения ключевых слов метод на базе нейросетевой модели показывает точности 0.46 по метрике f1-score, превосходя наиболее популярные методы статистического анализа, уступая методам SBE и KPMiner.

Таким образом проведенное исследование на примере перечисленных задач продемонстрировало пользу использования векторного представления текста, полученного на базе глубоких нейросетевых моделей, предварительно обученных на большом корпусе текстовых данных с последующим обучением для целевой задачи для создания эффективных моделей решения задач текстового анализа.

БЛАГОДАРНОСТИ

Работы выполнены при поддержке гранта РФФИ № 18-37-00331 “мол_а” и с использованием вычислительных ресурсов ОВК НИЦ “Курчатовский институт”, http://computing.nrcki.ru.

СПИСОК ЛИТЕРАТУРЫ

1. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. 2015. P. 320–332.
2. *De Smedt T., Daelemans W.* Pattern for Python. Journal of Machine Learning Research. 2012. № 13. P. 2031–2035.
3. *Straka M., Straková J.* Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2017.
4. NLTK, NLTK 3.3 release: May 2018. http://www.nltk.org
5. *Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.* Deep contextualized word representations. arXiv preprint arXiv:1802.05365. 2018.
6. *Devlin J., Chang M.W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018. Oct 11.
7. *Liu Y., Lapata M.* Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345. 2019. Aug 22.
8. *Liu Y.* Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318. 2019. Mar 25.
9. *Chen X., Gao S., Tao C., Song Y., Zhao D., Yan R.* Iterative document representation learning towards summarization with polishing. arXiv preprint arXiv:1809.10324. 2018. Sep 27.
10. *Stepanov M.A.* Generaciya zagolovkov novostnih statei, ispolzuya stemi, lemmi i grammemi. Kompyuternaya lingvistika i intellectual'nie tehnologii (po materialam ezhegodnoi mezhdunarodnoi konferencii "Dialog"). 2019. № 18. Additional vol.
11. *Gusev I.* Importance of copying mechanism for news headline generation // arXiv preprint arXiv:1904.11475. 2019.
12. *Gavrilov D., Kalaidin P., Malykh V.* Self-attentive model for headline generation // European Conference on Information Retrieval. 2019. C. 87–93.
13. *Sokolov A.M.* Phrase-based attentional transformer dlya generacii zagolovkov. Kompyuternaya lingvistika i intellectual'nie tehnologii (po materialam ezhegodnoi mezhdunarodnoi konferencii "Dialog"). 2019. № 18. Additional vol.
14. *Dong L., Yang N., Wang W., Wei F., Liu X., Wang Y., Gao J., Zhou M., Hon H.W.* Unified language model pre-training for natural language understanding and generation. InAdvances in Neural Information Processing Systems 2019 (P. 13042–13054).
15. *Song K., Tan X., Qin T., Lu J., Liu T.Y.* Mass: Masked sequence to sequence pre-training for language generation. arXiv preprint arXiv:1905.02450. 2019. May 7.
16. *Rose S., Engel D., Cramer N., Cowley W.* Automatic keyword extraction from individual documents. Text mining: applications and theory. 2010. P. 1–20.
17. *El-Beltagy S.R., Rafea A.* KP-miner: Participation in semeval-2. Proceedings of the 5th international workshop on semantic evaluation. 2010. P. 190–193.
18. *Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A.* YAKE! Keyword extraction from single documents using multiple local features. Information Sciences. 2020. № 509. P. 257–89.
19. *Gydovskikh D.V., Moloshnikov I.A., Naumov et al.* A probabilistically entropic mechanism of topical clusterisation along with thematic annotation for evolution analysis of meaningful social information of internet sources. Lobachevskii Journal of Math. 2017. V. 38. P. 910–913. https://doi.org/10.1134/S1995080217050134
20. *Mihalcea R., Tarau P.* Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing. 2004. P. 404–411.
21. *Wan X., Xiao J.* CollabRank: towards a collaborative approach to single-document keyphrase extraction. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). 2008. P. 969–976.
22. *Bougouin A., Boudin F.* TopicRank: Topic ranking for automatic keyphrase extraction. 2014. № 55. P. 45–69.
23. *Florescu C., Caragea C.* Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. V. 1. P. 1105–1115.
24. *Boudin F.* Unsupervised keyphrase extraction with multipartite graphs. arXiv preprint arXiv:1803.08721. 2018.
25. *Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G.* Kea: Practical automated keyphrase extraction. IGI global. Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. 2005. P. 129–152.
26. *Nguyen T.D., Luong M.T.* WINGNUS: Keyphrase extraction utilizing document logical structure. Association for Computational Linguistics. Proceedings of the 5th international workshop on semantic evaluation. 2010. P. 166–169.
27. *Turney P.D.* Learning algorithms for keyphrase extraction. Information retrieval. 2000. № 2 (4). P. 303–36.
28. *Page L., Brin S., Motwani R., Winograd T.* The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
29. *Moloshnikov I.A., Gryaznov A.V., Vlasov D.S., Sboev A.G.* Vibor effectivnogo neirosetevogo metoda formirovaniya zagolovkov. NIYaU MIFI. VI Mezhdunarodnaya konferenciya "Lazernie, pazmennie issledovaniya i tehnologii-LaPlaz-2020", sbornik nauchnih trudov. 2020. V. 1. P. 80–81.
30. *Kanerva J., Ginter F., Miekka N., Leino A., Salakoski T.* Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies. 2018. P. 133–142.
31. *Lin C.Y., Hovy E.* Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003. P. 150–157.
32. Headline Generation Shared Task on Dialogue'2019, http://www.dialog-21.ru/media/4661/cameraready-submission-157.pdf

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2020, vol. 9, no. 3, pp. 236–244

Application of a Multitasking Model for Practical Tasks of Heading Generation, Definition of Lemmas and Keywords

I. A. Moloshnikov[a,#], A. V. Gryanov[a,##], D. S. Vlasov[a,###], R. B. Rybka[a,####], and A. G. Sboev[a,b,#####]

[a] *National Research Center Kurchatov Institute, Moscow, 123182 Russia*

[b] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*

[#]*e-mail: ivan-rus@yandex.ru*

[##]*e-mail: artem.official@mail.ru*

[###]*e-mail: vfked0d@gmail.com*

[####]*e-mail: rybkarb@gmail.com*

[#####]*e-mail: sag111@mail.ru*

Received June 23, 2020; revised June 23, 2020; accepted July 7, 2020

Abstract—The efficiency of multitask deep learning methods for combined neural network models is comprehensively studied in application to a selected set of tasks: generating headers, defining lemmas, and keywords. The multitask model is built using Multi-head Attention layers and is used to develop models for generating headers and a model based on LSTM layers for lemmatization. Open corpuses RIA Novosti, containing news texts and headings for them, and a corpus with morphological, syntactic markup and lemmas for word forms SynTagRus from the universal dependencies project are used. For the task of highlighting keywords, we have assembled a new corpus consisting of news texts using the crowdsourcing platform. The results of the work show an increase in the accuracy by 1% in the F1 score for the lemmatization problem when using features from the multitask model compared to using only morphological features, state-of- art accuracy (0.42 ROUGE F1 score) is achieved for the title generation task. An algorithm for highlighting keywords without additional network training is proposed on the basis of the model for generating headings obtained in this work.

Keywords: multitask learning model, heading generation, lemmatization, key word highlighting

DOI: 10.1134/S2304487X20030074

REFERENCES

1. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, 2015, pp. 320–332.
2. De Smedt T., Daelemans W. Pattern for Python. Journal of Machine Learning Research, 2012, № 13, pp. 2031–2035.
3. Straka M., Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2017.
4. NLTK, NLTK 3.3 release: May 2018. http://www.nltk.org
5. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations. arXiv preprint arXiv:1802.05365. 2018.
6. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018. Oct 11.
7. Liu Y., Lapata M. Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345, 2019. Aug 22.
8. Liu Y. Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318. 2019. Mar 25.
9. Chen X., Gao S., Tao C., Song Y., Zhao D., Yan R. Iterative document representation learning towards summarization with polishing. arXiv preprint arXiv:1809.10324. 2018. Sep 27.
10. Stepanov M.A. Generaciya zagolovkov novostnih statei, ispolzuya stemi, lemmi i grammemi. Kompyuternaya lingvistika i intellectual'nie tehnologii (po materialam ezhegodnoi mezhdunarodnoi konferencii "Dialog"). 2019. № 18. Additional vol.
11. Gusev I. Importance of copying mechanism for news headline generation //arXiv preprint arXiv:1904.11475. 2019.
12. Gavrilov D., Kalaidin P., Malykh V. Self-attentive model for headline generation // European Conference on Information Retrieval. 2019. C. 87–93.
13. Sokolov A.M. Phrase-based attentional transformer dlya generacii zagolovkov. Kompyuternaya lingvistika i intellectual'nie tehnologii (po materialam ezhegodnoi

mezhdunarodnoi konferencii “Dialog”). 2019. № 18. Additional vol.

14. Dong L., Yang N., Wang W., Wei F., Liu X., Wang Y., Gao J., Zhou M., Hon H.W. Unified language model pre-training for natural language understanding and generation. InAdvances in Neural Information Processing Systems 2019 (pp. 13042–13054).
15. Song K., Tan X., Qin T., Lu J., Liu T.Y. Mass: Masked sequence to sequence pre-training for language generation. arXiv preprint arXiv:1905.02450. 2019. May 7.
16. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. Text mining: applications and theory, 2010. pp. 1–20.
17. El-Beltagy S.R., Rafea A. KP-miner: Participation in semeval-2. Proceedings of the 5th international workshop on semantic evaluation, 2010. pp. 190–193.
18. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 2020. № 509, pp. 257–89.
19. Gydovskikh D.V., Moloshnikov I.A., Naumov et al. A probabilistically entropic mechanism of topical clusterisation along with thematic annotation for evolution analysis of meaningful social information of internet sources. Lobachevskii Journal of Math, 2017. vol. 38. pp. 910–913. https://doi.org/10.1134/S1995080217050134
20. Mihalcea R., Tarau P. Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing, 2004. pp. 404–411.
21. Wan X., Xiao J. CollabRank: towards a collaborative approach to single-document keyphrase extraction. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 2008. pp. 969–976.
22. Bougouin A., Boudin F. TopicRank: Topic ranking for automatic keyphrase extraction. 2014. № 55. pp. 45–69.
23. Florescu C., Caragea C. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, Vol. 1, pp. 1105–1115.
24. Boudin F. Unsupervised keyphrase extraction with multipartite graphs. arXiv preprint arXiv:1803.08721. 2018.
25. Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G. Kea: Practical automated keyphrase extraction. IGI global. Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, 2005. pp. 129–152.
26. Nguyen T.D., Luong M.T. WINGNUS: Keyphrase extraction utilizing document logical structure. Association for Computational Linguistics. Proceedings of the 5th international workshop on semantic evaluation, 2010. pp. 166–169.
27. Turney P.D. Learning algorithms for keyphrase extraction. Information retrieval, 2000. № 2(4). pp. 303–36.
28. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
29. Moloshnikov I.A., Gryaznov A.V., Vlasov D.S., Sboev A.G. Vibor effectivnogo neirosetevogo metoda formirovaniya zagolovkov. NIYaU MIFI. VI Mezhdunarodnaya konferenciya “Lazernie, pazmennie issledovaniya i tehnologii-LaPlaz-2020”, sbornik nauchnih trudov. 2020. Vol. 1. pp. 80–81.
30. Kanerva J., Ginter F., Miekka N., Leino A., Salakoski T. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies, 2018, pp. 133–142.
31. Lin C.Y., Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003, pp. 150–157.
32. Headline Generation Shared Task on Dialogue'2019, http://www.dialog-21.ru/media/4661/cameraready-submission-157.pdf