ИНФОРМАТИКА

УДК 004.46:655.535.54

АВТОМАТИЗАЦИЯ ГЕНЕРАЦИИ ЗАГОЛОВКОВ НОВОСТНЫХ СТАТЕЙ

И.В. Тимохин, Н.Б. Осипенко

Гомельский государственный университет им. Ф. Скорины

AUTOMATION OF HEADLINE GENERATION FOR NEWS ARTICLES

I.V. Tsimokhin, N.B. Osipenko

F. Scorina Gomel State University

Описывается задача автоматизации создания заголовков. В процессе выполнения работы разработана методика применения подхода seq2seq с механизмом внимания с использованием искусственных нейронных сетей и реализован ее полный цикл от построения обучающей выборки до верификации полученных моделей.

***Ключевые слова**: генерация заголовков, обработка естественного языка, искусственные нейронные сети, реферирование текста.*

The approach for automation of headline generation, using the seq2seq method with the Attention mechanism is considered. The steps from dataset gathering to final model verification are described.

***Keywords**: headline generation, natural language processing, artificial neural networks, text summarisation.*

Введение

Рост объема обрабатываемой информации сделал актуальной задачу реферирования текста, поскольку краткий смысловой аналог исходного текста позволяет облегчить, ускорить и автоматизировать обработку информации. Её частным случаем является задача генерации заголовков, например, по исходному тексту новостной статьи.

1 Выбор подхода

Подходы к решению более общей задачи реферирования текста делятся на две группы: извлекающие (известные также как квазиреферирование, англ. extraction-based summarisation) и генерирующие (англ. abstraction-based summarisation). Извлекающие методы заключаются в поиске фрагментов текста, которые максимально характеризуют весь текст, генерирующие – в преобразовании исходного текста к внутреннему состоянию модели, из которого уже затем строится реферат. Использование извлекающих методов ограничивает множество конечных результатов, а использование генерирующих методов с хорошими моделями позволяет разнообразить получаемый результат.

При решении задач рассматриваемого типа хорошо зарекомендовал себя метод seq2seq (sequence to sequence – из последовательности в последовательность) [1] и его модификация seq2seq with attention с механизмом внимания, предложенная в [2]. Подход seq2seq основан на построении кодировщика и декодировщика, строящих представление входной и выходной последовательностей соответственно. В задаче реферирования входными данными является исходный текст, а выходными – сжатый текст. Механизм внимания основан на идее, что при генерации выходной последовательности на каждый выходной элемент влияет лишь несколько элементов исходной последовательности. Так, например, при переводе для получения итогового слова перевода не обязательно использовать все слова исходной последовательности, обычно необходимо знать лишь несколько слов. При использовании механизма внимания для каждого элемента выходной последовательности считается вектор-контекст, который и используется для создания выходной последовательности.

2 Описание алгоритма построения модели

Для построения модели генерации заголовков по исходному тексту новостной статьи разработана методика, схематично представленная в виде шести приведенных ниже этапов.

Этап 1. *Сбор исходных данных*. Для создания выборки был выбран белорусский портал tut.by, ежедневно публикующий новости с 2001 года. Для парсинга сайта использовалась библиотека Scrapy. Всего было собрано 300 тысяч статей с заголовками, опубликованных между 2001 годом и ноябрём 2019 года. Во время сбора данных текст очищался от html-тегов и непечатаемых символов.

Этап 2. *Обучение SentencePiece*. В некоторых языках для токенизации (процесса разделения текста на отдельные компоненты – токены) можно использовать разбиение текста по пробелу и каждое слово становится токеном. Такой подход неплохо работает в случае английского языка. Иногда для решения задачи анализа текста применяют стемминг (процесс поиска основы слова, которая не обязательно совпадает с его морфологическим корнем) и лемматизацию (процесс приведения слова к словарной форме).

Таблица 2.1 – Параметры обучаемых моделей

Название модели	Токенов SentencePiece	Векторов GloVe	Скрытых слоёв	Нейронов в скрытых слоях
10k-big	10 тысяч	50	2	300
5k-big	5 тысяч	50	2	300
10k-medium-3	10 тысяч	50	3	50
10k-medium-2	10 тысяч	50	2	50

Для токенизации (преобразования исходной последовательности слов к последовательности токенов) текста использовалась библиотека SentencePiece. Размер итогового множества всех токенов (подслов) задаётся перед обучением.

Этап 3. *Токенизация исходных данных с SentencePiece*. Исходные данные преобразовывались с использованием обученной на предыдущем этапе модели SentencePiece.

Этап 4. *Обучение GloVe*. Передача последовательности токенов в нейронную сеть требует предварительного преобразования к векторам. На основе каждого словаря, полученного SentencePiece, обучались вектора GloVe (Global Vectors for Word Representation; глобальные вектора для представления слов). Обучение векторов GloVe, так же как и SentencePiece, занимает незначительное время по сравнению со временем, необходимым для обучения нейросети. Время обучения нейросети, описанного на этапе 5, на векторах размера 300 оказалось слишком большим. Поэтому размер вектора GloVe был уменьшен до 50.

Этап 5. *Обучение нейронной сети*. Нейросеть строилась с использованием библиотек Keras и TensorFlow на основе метода seq2seq with attention. Длина исходной последовательности ограничивалась 300 токенами, а исходящей – 40 токенами. Обучение моделей производилось с использованием метода стохастической оптимизации Adam с параметром 0,001. Обучались модели с разным количеством скрытых слоёв и количеством нейронов в этих слоях. Описание параметров обучаемых моделей приводится в таблице 2.1.

Этап 6. *Оценка точности полученной нейросети*. Для оценки релевантности автоматического реферирования часто используют набор метрик ROUGE-n (*Recall-Oriented Understudy for Gisting Evaluation*), как пересечение *n*-граммов между предсказываемым и эталонным результатом. Нами использованы ROUGE-1 и ROUGE-2 для уни- и биграммов. Для подсчёта метрики ROUGE-n используются точность $P = m / w_t$ и полнота $R = m / w_r$, а на их основе считается мера $F_1 = 2 \cdot (P + R) / (P \cdot R)$, где m – число n-граммов в гипотезе, которые также находятся в эталоне, w_t – число n-граммов в гипотезе, w_r – число n-граммов в эталоне. Метрика ROUGE-L подсчитывает длину наибольшей общей подпоследовательности (*LCS*: *Longest Common Subsequence*) слов. Преимуществом этой метрики является то, что она требует не последовательных совпадений. Формула для подсчёта ROUGE-L приводится в [3].

3 Описание результатов апробации

Используя новостные статьи с заголовками, опубликованными на белорусском портале tut.by, получены четыре модели с описанными в таблице 2.1 параметрами. На рисунке 3.1 приводится сравнение времени обучения этих моделей.

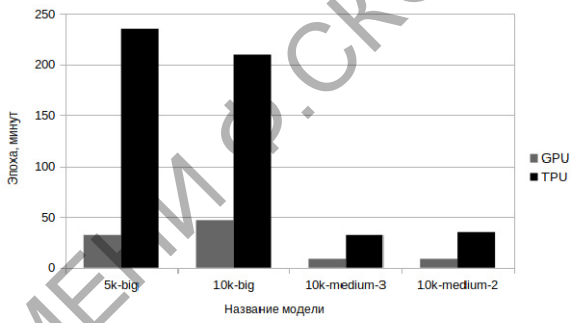

Рисунок 3.1 – Эпоха в минутах для разных сетей на разных устройствах

Выбранные модели обучались до сходимости. Результаты оценки точности полученных моделей на тестовой выборке после заданного числа эпох приводятся в таблице 3.1. Модели 10k-big и 5k-big отличаются между собой количеством токенов, на которые разбивались исходные данные с SentencePiece, и показывают влияние размера словаря на точность полученной модели. Модели 10k-big и 10k-medium-2 отличаются количеством нейронов в скрытых слоях и показывают как повлияло количество нейронов в скрытых слоях на точность моделей. Модели 10k-medium-3 и 10k-medium-2 отличаются количеством скрытых слоёв в нейронных сетях и показывают как повлияло количество скрытых слоёв на точность моделей. Обучение моделей производилось с использованием метода стохастической оптимизации Adam с параметром 0,001. Сравнивая полученные результаты между собой и с опубликованными аналогами, следует учитывать, что они были получены на разных выборках, поэтому имеют условный характер. Т. е. если показанный результат одной модели лучше, чем показанный результат другой модели, то это не обозначает, что одна модель лучше другой сама по себе. Наилучшей моделью из четырех, полученных в этой работе, является 10k-big по значениям F_1 и полноты ROUGE-2 и F_1 ROUGE-L. Наилучший результат среди всех, приведенных в таблице 3.1 по остальным сравниваемым значениям, показывает модель Universal Trans-former on RIA [5].

Таблица 3.1 – Сравнение моделей

Название модели	Значение					
	ROUGE-1		ROUGE-2		ROUGE-L	
	F_1	Полнота	F_1	Полнота	F_1	Полнота
5k-big	34,60	31,46	25,28	23,08	32,71	29,76
10k-big	39,01	36,05	29,91	27,84	37,65	34,85
10k-medium-2	26,14	23,97	16,48	15,21	24,49	22,48
10k-medium-3	24,09	21,18	15,40	13,64	22,84	20,08
Pointer-Gen-Coverage [4]	24,68	–	10,92	–	21,78	–
Universal Transformer w/smoothing on NYT [5]	25,60	23,90	12,92	12,42	23,66	25,27
Universal Transformer on NYT [5]	26,86	25,33	13,48	13,01	24,84	24,38
Universal Transformer w/smoothing on RIA[5]	39,31	37,10	21,82	20,66	36,32	35,37
Universal Transformer on RIA [5]	39,75	37,62	22,15	21,04	36,81	35,91
Moses+ on DUC-2004 [6]	26,50	–	8,13	–	22,85	–
Abs on DUC-2004 [6]	26,55	–	7,06	–	22,05	–
Abs+ on DUC-2004 [6]	28,18	–	8,49	–	23,81	–
Moses+ on Gigaword [6]	28,77	–	12,10	–	26,44	–
Abs on Gigaword [6]	30,88	–	12,22	–	27,77	–
Abs+ on Gigaword [6]	31,00	–	12,65	–	28,34	–

На примере графика обучения сети 5k-big, приведенного на рисунке 3.2, можно видеть, что значения выбранных четырех метрик влияют на сходимость модели сходным образом.

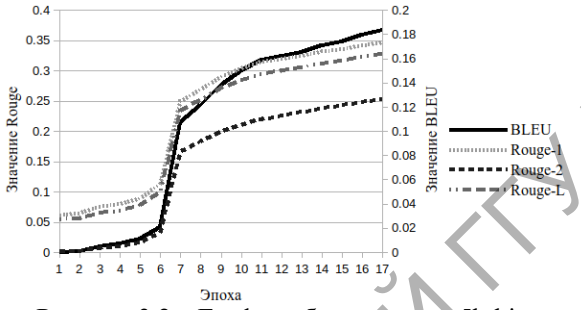

Рисунок 3.2 – График обучения сети 5k-big (метрики BLEU, ROUGE)

В большинстве случаев сгенерированные сетью заголовки не требовали доработки и хорошо отражали смысл статьи, но иногда программа генерирует бессмысленный текст, например, для статьи с заголовком «Минералка в одном из минских магазинов на акции стала дороже, чем стоила до распродажи» был сгенерирован заголовок «В Минске устроили распродажи немецкого минералка». Значит, кроме чисто лингвистического анализа текста, необходимо добавлять семантическую компоненту, учитывающую наличие смысла в сформированном заголовке.

Заключение

С целью построения обученной модели для автоматического реферирования текста на примере русскоязычной новостной ленты разработана методика применения подхода seq2seq with attention с использованием искусственных нейронных сетей. Реализован полный цикл, начиная с этапа сбора выборки, сам процесс обучения искусственных нейросетей и применения полученных моделей. Модели, использующие метод seq2seq with attention, в настоящее время широко используются и описаны например в работах [4]–[6]. Он использован в данном исследовании для автоматического создания заголовков, кратко отражающих суть новостных статей. Разработанная методика решения задачи автоматизации реферирования текста может быть применена без необходимости внесения в неё кардинальных изменений для решения аналогичных задач.

ЛИТЕРАТУРА

1. *Sutskever*, *I*. Sequence to Sequence Learning with Neural Networks [Electronic resource] / I. Sutskever, O. Vinyals, Q.V.Le. – 2014. – Mode of access: https://arxiv.org/pdf/1409.3215.pdf. – Date of access: 01.10.2019.

2. *Bahdanau*, *D*. Neural Machine Translation by Jointly Learning to Align and Translate [Electronic resource] / D. Bahdanau, K. Cho, Y. Bengio. – 2014. – Mode of access: https://arxiv.org/abs/1409.0473. – Date of access: 01.10.2019.

3. *Lin*, *C.-Y*. Rouge: A package for automatic evaluation of summaries / Chin-Yew Lin // Text summarization branches out. – 2004. – P. 74–81.

4. *Xu*, *P*. A novel repetition normalized adversarial reward for headline generation [Electronic resource] / P. Xu, P. Fung. – 2019. – Mode of access: https:// arxiv.org/pdf/1902.07110.pdf. – Date of access: 01.10.2019.

5. *Gavrilov*, *D*. Self-Attentive Model for Headline Generation [Electronic resource] / D. Gavrilov, P. Kalaidin, V. Malykh. – 2019. – Mode of access: https://arxiv.org/pdf/1901.07786.pdf. – Date of access: 01.10.2019.

6. *Rush*, *A.M*. A Neural Attention Model for Abstractive Sentence Summarization [Electronic resource] / A.M. Rush, S. Chopra, J. Weston. – 2015. – Mode of access: https://arxiv.org/pdf/1509.00685.pdf. – Date of access: 01.10.2019.

Поступила в редакцию 16.06.2020.