САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-Механический факультет
Кафедра информационно аналитических систем

Суммаризация групп в социальных сетях

Дипломная работа студента 645 группы Чурикова Никиты Сергеевича

Научный руководитель:
к.ф. - м.н., доцент Графеева Н. Г.

Рецензент:
Руководитель департамента вычислительной биологии
Яковлев П. А.

Заведующий кафедрой:
к.ф. - м.н., доцент Михайлова Е. Г.

Санкт-Петербург
2019 г.

Содержание

1 Аннотация **1**

2 Введение **1**
2.1 Постановка задачи 2
2.2 Обзор литературы 2
2.3 Полученные результаты 3

3 Алгоритмы, использованные в работе **3**
3.1 Суммаризация текста 3
3.1.1 TextRank 3
3.1.2 Byte pair encoding 4
3.1.3 Transformer network 4
3.2 Оценки качества 4

4 Эксперименты **4**

5 Заключение **4**

6 Литература **5**

1 Аннотация

Одной из задач обработки естественного языка является задача суммаризации текста. Ее целью является уменьшение размера исходного текста без потери ключевой информации. В данной работе мы решаем схожую проблему, но для информационных ресурсов в социальных сетях. В частности, необходимо рассмотреть задачу суммаризации текстов и картинок, поскольку это два основных источника информации. В тексте мы приводим численное обоснование выбранных методов, а также приводим оценку нашей суммаризации людьми.

2 Введение

В современном мире создается все больше и больше информации, которую мы можем потреблять. Новости, статьи, юмор постоянно меняются и создаются людьми. При таком потоке информации появляется потребность в инструментах, способных давать как можно больше информации с минимальными потерями.

При чтении новостей люди, как правило, не идут дальше новостных заголовков [3], для популярных технических статей создают краткие описания описывающие их достижения и основные моменты [11, 1], а визуальный контент нередко подчиняется единому шаблону.

В данной работе мы показываем, как используя современные достижения в области анализа данных можно извлекать полезную информацию из новостных ресурсов в социальной сети вконтакте [13], приводим оценки людей нашей системы и приводим сравнение с наивными решениями.

2.1 Постановка задачи

Мы поставили перед собой задачу создать систему, которой бы можно было передавать ссылку на новостной ресурс в социальной сети вконтакте, а на выходе получать его краткое описание. В рамках работы мы ограничились новостными ресурсами с высоким содержанием текста.

С алгоритмической точки зрения, задача суммаризации новостного ресурса была рассмотрена нами как две подзадачи:

1. Извлечение ключевых слов, присущих данному источнику информации;

2. Сжатие новостей, используя автоматическое создание заголовков.

Через извлечение данной информации мы хотим добиться эффекта "чтения по диагонали".

Для оценки качества наших алгоритмов, мы воспользовались открытыми датасетами для суммаризации текстов, а также проводили оценку качества людьми.

2.2 Обзор литературы

Задача сжатия текста с малой потерей смысла и сохранением возможности его прочтения имеет название задачи *суммаризации*. При этом, есть два концептуальных подхода к решению: экстрактивный, когда для создания краткого содержания извлекаются целые куски текста вплоть до предложений, и абстрактивная, где в кратком содержании могут быть слова, которых не было в исходном тексте.

В частности, при исследовании абстрактивной генерации заголовков, мы отталкивались от статьи Вконтакте, посвященной данной проблеме [4]. Ими предлагается применять нейронные сети с архитектурой Transformer и предобработкой Byte pair encoding (BPE) [10]. Однако в задаче абстрактивной генерации заголовков существуют дебаты на тему того, что использовать в качестве входа модели. Поскольку долгое время SOTA были модели с архитектурой encoder decoder, то было невозможно использовать длинные входные последовательности. Потому авторы статьи [8] исследуют различные подходы по предварительному извлечению "Topic sentence которое нейронная сеть должна дальше обработать. Это предложение, как говорят авторы, в идеальном случае, должна отвечать на 5W1H. Но достаточно ответов на "что, кто, когда".

Для экстрактивной суммаризации чаще всего используют алгоритм TextRank [6]. Идея

2.3 Полученные результаты

Что является результатом работы (будет веб сервис, куда можно закинуть ссылку на группу), как оценивали качество (продолжить результаты работы алгоритмов толокерам), а также оценка качества по автоматизированным метрикам, и как они коррелируют с оценками людей. Сравниться с бэйзлайном.

3 Алгоритмы, использованные в работе

Нами были использованы как классические подходы, так и новые, основанные на нейронных сетях. В следующих секциях мы опишем их основные принципы, а также приведем ссылки на их реализации.

3.1 Суммаризация текста

Для суммаризации текста мы воспользовались алгоритмом экстрактивной суммаризации основанном на TextRank [12, 9, 6], и моделью трансформера [2], обученной на датасете РИА новостей [4]. Для предобработки данных модели трансформера мы использовали byte pair encoding [10]. Помимо этого мы извлекали первое предложение из новости. Для TextRank и извлечения первого предложения не требуется обучающая выборка, что делает их очень удобными в использовании. При этом, исследования показывают, что в задаче генерации заголовков, первое предложение в новости – это очень сильный бэйзлаин [4], который трудно побить как экстрактивной, так и абстрактивной суммаризацией.

3.1.1 TextRank

TextRank является является адаптацией идеи алгоритма PageRank [7] с задачи рекомендации страниц в интернете на задачу рекомендации лучшего предложения или набора слов в тексте. Сам алгоритм состоит в том, что мы текст превращаем в граф, где узлы – это предложения, а для каждого ребра подсчитывается вес, где вес определяется по количеству совпавших слов в двух предложениях.

Таким образом, получается, что можно выбрать предложение с самыми тяжелыми ребрами в качестве предложения, которое описывало бы исходный текст.

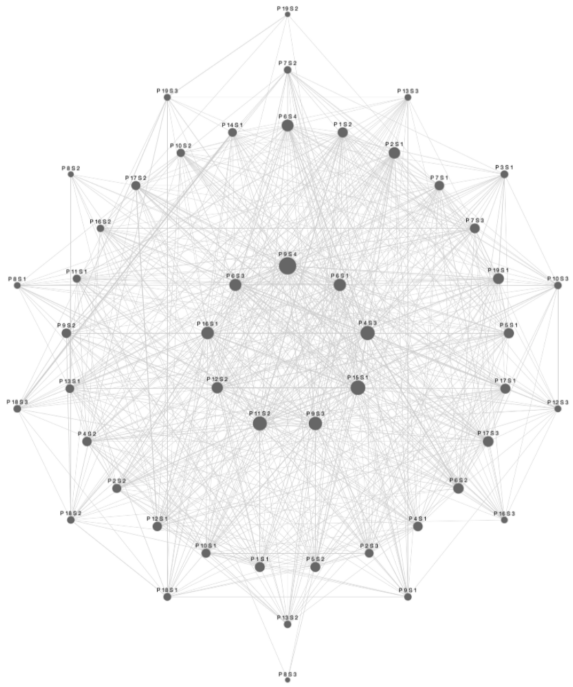

Рис. 1: Пример результирующего графа textrank.

3.1.2 Byte pair encoding

3.1.3 Transformer network

3.2 Оценки качества

Для оценки качества текстовых моделей мы использовали метрику ROUGE-L F1 [5], при этом мы считали ее на датасете РИА новостей [4].

Помимо этого, как для текстовых данных, так и для изображений, мы использовали Яндекс.Толоку [14], чтобы привлечь людей к оценке качества наших результатов.

4 Эксперименты

Для обучения моделей были использованы 8 Tesla K80.

5 Заключение

На февраль 2019:

В данной работе мы ожидаем показать, что предложенные нами решения не хуже, а даже лучше предложенных бэйзлайнов как по автома-

тическим оценкам, так и по оценкам людей. Мы также представим код и ссылку на сервис, куда можно отправить ссылку на интересующую группу и оценить получившийся результат.

6 Литература

Список литературы

[1] article essence. Article essence. — 2019. — feb. — https://opendatascience.slack.com/messages/C5VQ222UX.

[2] Attention is all you need / Ashish Vaswani, Noam Shazeer, Niki Parmar et al. // CoRR. — 2017. — Vol. abs/1706.03762.

[3] DeMers, J. 59 percent of you will share this article without even reading it. — 2016. — aug. — https://www.forbes.com/sites/jaysondemers/2016/08/08/59-percent-of-you-will-share-this-article-without-even-reading-it/#71991fa2a648.

[4] Gavrilov, D. Self-attentive model for headline generation / Daniil Gavrilov, Pavel Kalaidin, Valentin Malykh // Proceedings of the 41st European Conference on Information Retrieval. — 2019.

[5] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries / Chin-Yew Lin // Proc. ACL workshop on Text Summarization Branches Out. — 2004. — P. 10. — http://research.microsoft.com/~cyl/download/papers/WAS2004.pdf.

[6] Mihalcea, R. Graph-based ranking algorithms for sentence extraction, applied to text summarization / Rada Mihalcea. — 2004. — 01. — Vol. 170-173.

[7] Page, L. The pagerank citation ranking: Bringing order to the web. — 1998.

[8] Putra, J. W. G. Incorporating topic sentence on neural news headline generation / Jan Wira Gotama Putra, Hayato Kobayashi, Nobuyuki Shimizu. — 2018.

[9] Řehůřek, R. Software Framework for Topic Modelling with Large Corpora / Radim Řehůřek, Petr Sojka // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. — Valletta, Malta: ELRA, 2010. — May. — P. 45–50. — http://is.muni.cz/publication/884893/en.

[10] Sennrich, R. Neural machine translation of rare words with subword units / Rico Sennrich, Barry Haddow, Alexandra Birch // CoRR. — 2015. — Vol. abs/1508.07909.

[11] tldr arxiv. tldr arxiv. — 2019. — feb. — https://t.me/tldr_arxiv.

[12] Variations of the similarity function of textrank for automated summarization / Federico Barrios, Federico López, Luis Argerich, Rosa Wachenchauzer // CoRR. — 2016. — Vol. abs/1602.03606.

[13] vkontakte. vkontakte. — 2019. — feb. — https://vk.com/feed.

[14] Yandex. Yandex.toloka. — 2019. — feb. — https://toloka.yandex.ru.