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

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 [6], для популярных технических статей создают краткие описания описывающие их достижения и основные моменты [14, 2], 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 [16], приводим

оценки людей нашей системы и приводим сравнение с наивными решениями.

2.1 Постановка задачи

В данной работе мы решили остановиться на двух основных современных видах медиа: тексте и изображениях. В данной работе мы не рассматриваем обработку видео, но есть предположения, что предложенные идеи насчет изображений можно было бы распространить на видеоинформацию.

Для текстовых ресурсов задача суммаризации была разбита на две подзадачи: 1) извлечение ключевых слов, присущих данному источнику информации и 2) автоматическое создание заголовков.

Для изображений – это сбор похожих изображений в кластера и показ некоторых одних изображений, иллюстрирующих каждую группу.

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

2.2 Обзор литературы

Рассказать про литературу, которая рассматривает задачи выше.

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 [15, 12], и моделью трансформера [3], обученной на датасете РИА новостей [7]. Для предобработки данных модели трансформера мы использовали byte pair encoding [13]. 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 [7], который трудно побить как экстрактивной, так и абстрактивной суммаризацией.

3.2 Суммаризация изображений

Для суммаризации изображений мы реализовали алгоритм, описанный в статье [8]. Основная идея состоит в том, что из изображений извлекаются признаки, инвариантные к поворотам [10], эти признаки кластеризуют используя k-means [1] и индексы кластеров используют как признаки для латентного размещения Дирихле [4, 12].

Помимо этого, мы попробовали на нашей задаче обучению метрике между изображениями [11, 5].

3.3 Оценки качества

Для оценки качества текстовых моделей мы использовали метрику ROUGE-L F1 [9], при этом мы считали ее на датасете РИА новостей [7].

Помимо этого, как для текстовых данных, так и для изображений, мы использовали Яндекс.Толоку [17], 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 автоматическим 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. Мы также планируем показать результаты в "одноклассниках"и рассказать об их мнении насчет полученного решения, поскольку год назад предлагалось совместное сотрудничество над данной проблемой.

Список литературы

[1] Arthur, D. K-means++: The advantages of careful seeding / David Arthur, Sergei Vassilvitskii // Proceedings of the Eighteenth An-

nual ACM-SIAM Symposium on Discrete Algorithms. — SODA '07. — Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007. — P. 1027–1035. — http://dl.acm.org/citation.cfm?id=1283383.1283494.

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

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

[4] Blei, D. M. Latent dirichlet allocation / David M. Blei, Andrew Y. Ng, Michael I. Jordan // J. Mach. Learn. Res. — 2003. — Mar. — Vol. 3. — P. 993–1022. — http://dl.acm.org/citation.cfm?id=944919.944937.

[5] Deep metric learning with hierarchical triplet loss / Weifeng Ge, Weilin Huang, Dengke Dong, Matthew R. Scott // CoRR. — 2018. — Vol. abs/1810.06951.

[6] 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.

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

[8] Image summarization using topic modelling / Vasu Sharma, Akshay Kumar, Nishant Agrawal et al. // ICSIPA. — IEEE, 2015. — P. 226–231.

[9] 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.

[10] Lowe, D. G. Distinctive image features from scale-invariant keypoints / David G. Lowe // Int. J. Comput. Vision. — 2004. — Nov. — Vol. 60, no. 2. — P. 91–110. — https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[11] Mining on manifolds: Metric learning without labels / Ahmet Iscen, Giorgos Tolias, Yannis S. Avrithis, Ondrej Chum // CoRR. — 2018. — Vol. abs/1803.11095.

[12] 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.

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

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

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

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

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