2022

News Headlines Generation

Дарья Измалкова, Мария Непомнящая, Дарья Родионова

Цель и задачи проекта

01 Цель

- Попытаться улучшить генерацию заголовков на материалах русского языка

02 Задачи

- Определить, влияет ли добавление разметки на качество
- Испытать методы, которые для русского еще не использовались

Мотивация/актуальность

- Генерация заголовков напрямую полезна для агрегаторов новостей, газет и т. п.
- Может служить основой для классификации текстов (напр. по жанрам), агрегации текстов.
- Наверное, может быть полезно для анализа большого количества информации и вне области СМИ (вместо того, чтобы читать огромное количество текстов целиком можно оценить их актуальность по заголовкам).
- По сравнению с задачей генерации summary данные (текст + заголовок) более доступны и обширны.

Команда и роли

Дарья Измалкова

- Предобработка данных
- Реализация модели на базе GPT-2

Мария Непомнящая

- Сбор данных
- Реализация метода на основе UniLM

Дарья Родионова

- Реализация бейслайна
- Реализация метода на основе BertSumAbs

Данные

	Датасет новостей с Lenta.ru	Датасет новостей «Россия Сегодня»
Количество новостей	800K+	1M+
Период	сентябрь 1999 — декабрь 2019	январь 2010 — декабрь 2014

- Мы планируем также скачать новости с Lenta.ru за 2020–2022 гг.
- Оба датасета будут использоваться для обучения моделей с нуля. В итоге получится примерно 2 миллиона новостей.
- Для файтюнинга уже предобученных моделей будет использован датасет с Lenta.ru

Бейзлайн

На Диалоге 2019 в соревновании по предсказанию новостных заголовков лучшие результаты показал метод с использованием CopyNet, кроме того, для обучения и тестирования были использованы 3 датасета, из которых 2 используем мы, а третий засекречен. Из-за этого было решено взять именно этот метод в качестве бейзлайна.

Метрики

Для оценки качества генерации суммаризации принято использовать набор метрик под названием ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin 2004].

ROUGE-1 и ROUGE-2 основаны на вычислении юниграмм и биграмм соответственно.

ROUGE-L определяется как F-мера, основанная на самой длинной подстроке (Longest Common Subsequence).

BLEU — изначально создана для оценки автоматического перевода, однако также часто используется и для суммаризации. Утверждается, что она немного лучше коррелирует с человеческой оценкой, чем ROUGE. [Аишева 2021]

Lin C. Y. Rouge: A package for automatic evaluation of summaries // *Text summarization branches out*. – 2004. – С. 74-81.
Аишева Д.А. Модификация нейронной сети Transformer для генерации новостных заголовков на русском языке. – 2021. – С. 28.

План действий

1. Предобработка текстов: поэкспериментировать и выяснить нужно ли лемматизировать, убрать пунктуацию, оставить только начала текстов.
2. Реализовать бейзлайн (CopyNet) и поэкспериментировать с признаками: добавить BIO- и POS-разметку, а также тональность. Посмотреть, влияет ли разметка на результат.
3. Аналогичные эксперименты с признаками для BertSumAbs.
4. Реализовать другие архитектуры, которые не были до этого сделаны для русского, а именно:
   - Модель на базе GPT-2 из статьи [Li, Yu, Chen, Guo 2021]
   - UniLM

Li P., Yu J., Chen J., Guo B. HG-News: News Headline Generation Based on a Generative Pre-Training Model. – 2021.

Литература

Malykh V. A., Kalaidin P. S. Headline Generation Shared Task on Dialogue'2019 // *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*. – 2019. – №. 18. – С. 93-100. — Результаты Диалога 2019 по задаче генерации заголовков (выбор модели-победителя для использования в бейзлайне)

Gusev I. Importance of copying mechanism for news headline generation // *arXiv preprint arXiv:1904.11475.* – 2019. — Реализация CopyNet на Диалоге 2019

Bukhtiyarov A., Gusev I. Advances of Transformer-Based Models for News Headline Generation // *Conference on Artificial Intelligence and Natural Language*. – Springer, Cham, 2020. – С. 54-61. — Улучшение результатов Диалога 2019 с помощью mBART и BertSumAbs

Li P. et al. HG-News: News Headline Generation Based on a Generative Pre-Training Model // *IEEE Access*. – 2021. – Т. 9. – С. 110039-110046. — Реализация модели на базе GPT-2

Dong L. et al. Unified language model pre-training for natural language understanding and generation // *arXiv preprint arXiv:1905.03197*. – 2019. — Реализация UniLM

Аишева Д. А. и др. Модификация нейронной сети Transformer для генерации новостных заголовков на русском языке. – 2021. — Обзор существующих методов, метрик и реализация метода на основе mT5

Ссылки