IS 6733 Fall 2024 Group Project: Numeral-aware Headline Generation

October 31, 2024

Overview

Recognizing the significance of understanding numbers can enhance performance in practical Natural Language Processing (NLP) tasks. For instance, the headline "Expecting the stock price to increase by 30%" is different from "Expecting the stock price to increase by 3%" [1, 2], with the former example suggesting a higher level of sentiment than the latter. Furthermore, including numbers in a headline is more informative and engaging than a headline with no numbers [3]. In this project, our objective is to generate "numeral-aware headlines", *i.e.* based on a dataset of news articles, and evaluate our generation models based on text generation metrics.

Tasks

The primary tasks in this project are taken from the SemEval-2024 Task 7 [1]. The tasks involves generating a numeral-aware headlines of news articles, and perform numerical reasoning. Given a news article, your model should output a headline for the news article that contains numerals sourced from the news article, and accurately perform numerical reasoning. Figure 1 shows a sample from this task.

Your solution must use an LLM. Solution should not be just based on "out-of-the-box" application of LLMs (such approaches would also perform poorly). For samples approaches, take a look at [1]. If you want to re-implement one of the sample approaches, that would be perfectly okay as well. You will need to implement the system yourself in any case.

Since the major goal of this task is to generate a headline that contains key numerical information in the news article, the evaluation will be based on not only the presence of these numerical information, but also the correctness of this information.

News:
At least **30** gunmen burst into a drug rehabilitation center in a Mexican border state capital and opened fire, killing **19** men and wounding **four** people, police said. Gunmen also killed **16** people in another drug-plagued northern city. The killings in Chihuahua city and in Ciudad Madero marked one of the bloodiest weeks ever in Mexico and came just weeks after authorities discovered **55** bodies in an abandoned silver mine, presumably victims of the country's drug violence. More than **60** people have died in mass shootings at rehab clinics in a little less than **two** years. Police have said **two** of Mexico's **six** major drug cartels are exploiting the centers to recruit hit men and drug smugglers, ...

Headline (Question): Mexico Gunmen Kill ____

Answer: 35

Annotation: Add(19,16)

Figure 1: Task sample.

Dataset

We will use the Num-HG dataset, available in this paper: [2].

Evaluation

We will evaluate both numerical reasoning capability and headlined generated. We will have two kinds of evaluations:

- Automatic evaluation: Based on metrics, such as Accuracy [1], ROUGE [4], BERTScore [5], and MoverScore [6].
- Peer evaluation: You will exchange a few samples anonymously and each team will rank the best headlines. See Section 4 of the SemEval task [1].

Timeline

The tentative timeline of the group project is depicted in Table 1.

Date	Objective
November 7th	Lecture
November 14th	Presentation: System Description
November 21st	Lecture/Short Project Update
November 28th	Thanksgiving!
December 5th	Presentation: Final Project Presentation and peer evaluation

Table 1: Timeline

References

[1] Chung-Chi Chen, Jian-Tao Huang, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Semeval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1482–1491, 2024.

[2] Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Numhg: A dataset for number-focused headline generation. *arXiv preprint arXiv:2309.01455*, 2023.

[3] Kenza Lamot, Tim Kreutz, and Michaël Opgenhaffen. "we rewrote this title": How news headlines are remediated on facebook and how this affects engagement. *Social Media+ Society*, 8(3):20563051221114827, 2022.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[6] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019.