Challenge at SemEval 2024 Task 7: Contrastive Learning Approach on Numeral-Aware Language Generation

Ali Zhunis
University of Tübingen
Tübingen, Germany
ali.zhunis@student.uni-tuebingen.de

Hao-Yun Chuang
National Chengchi University
Taipei, Taiwan
110555010@nccu.edu.tw

Abstract

Although Large Language Model (LLM) excels on generating headline on ROUGE evaluation, it still fails to reason number and generate news article headline with an accurate number. Attending SemEval-2024 Task 7 subtask 3, our team Challenges aims on using contrastive loss to increase the understanding of the number from their different expression, and knows to identify between different number and its respective expression. This system description paper uses T5 and BART as the baseline models, comparing its result with and without the constrative loss. The result shows that BART with contrastive loss have surpassed all the models, and its performance on the number accuracy has the highest performance among all.

1 Introduction

This paper is a description of the methods we have applied for our implementation on this year's SemEval Task 7, NumEval: Numeral-Aware Language Understanding and Generation. SemEval is an annual workshop which is consisted of various natural language processing shared tasks. Teams that join the tasks is required to design systems that could enhance the understanding or improve results on various kinds of semantic evaluation challenge. The task we decide to join was task 7, NumEval: Numeral-Aware Language Understanding and Generation (Huang et al., 2023). Specifically, we focus on the second subtask from the third task in NumEval, which centers on generating proper news headline based on the provided news articles. Different from article summary, a headline must condense the essence from the full length article. Although the encoder-decoder language models nowadays has excelled on generation of the text based on the ROUGE metric, it still fails on providing precise numeral generation in headlines owing to the fact that the representation of the number may differ in various kinds of forms. Therefore, the goal of this task is to enhance the accuracy of the model in the generation of the number from the headline of a news article.

While the numeral expression in the article consists of text and numbers, our system aims to use the technique of contrastive learning. With this technique, it is possible to help the model enclose the similarity between the number and its text expression, and enlarge the difference between different numbers (and its respective expression).

The evaluation of the performance is divided into two parts, one is to evaluate the accuracy of the predicted number, the other is to evaluate the accuracy of the word prediction, which uses ROUGE, BERTScore, and MoverScore to evaluate the results.

Two datasets are included in the task. One is the dry-run dataset that is provided on the official site. It contains a total of 100 instances; the other is the official training set that is provided after the registration of the task, which includes a total of 21157 instances. All of them have the same data structure. Each instance contains both the news article that includes the date of release, and its respective headline. Each team is expected to generate a precise news title from its respective news article.

Numbers are one of the most important element among medical, business, and legal article, and it could be dangerous if the large language model has misunderstood the content of the article. The finding of this paper could further discover a better way for large language model to detect and reason the numbers from the respective article, and thus generate the accurate headline with correct number.

Our system description is divided into three main section. First, we present the main approach that we used as the final submission result in detail, including the preparation of the dataset, its augmentation, and the structure of the model in contrastive learning that we designed in order to solve the task.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1659–1662
June 20-21, 2024 ©2024 Association for Computational Linguistics

Hence, we talk about the adjustment of the model, including parameter optimization on the model. Finally, we present our experimental result. In the end, we discuss about the possible future work, and conclude with a brief summary of our system.

2 Related work

Large Language Model (LLM) like ChatGPT has been long commented with its brittleness on the ability of numerical reasoning. When the questions presented in the varying textual form (comprising words and numbers), LLM would result in inconsistent performance (Ahn et al., 2024). In (Huang et al., 2023)'s work, it shows that although large language model excelled based on ROUGE metrics, it still fails to generate precise numeral in headline.

Researchers have applied contrastive learning in natural language processing. To generate headline with different author style, (Liu et al., 2022) has applied contrastive learning to integrate the stylistic feature of the author into the model. This research hence inspires us to use contrastive learning on integrating the numeral features in different text form to the model, in order to let model identify the correct number from the news article.

3 Methodology

In this section, we propose the methodology of our system. It mainly consists of three parts. First, we will talk about the augmentation of the data. Secondly, the model training and fine-tuning, and finally, parameter optimization.

3.1 Data pre-processing and augmentation

A total of two vectors are used in our experiment. The first vector enhance the model understanding of different expression on number. One is to change all the text expression of number into numeral expression, and vice versa. While it encloses the similarity between the different expression of the same number, we call it positive distorted sample. With this change, model can better learn the different expression from the same number. We manually annotated the dataset to change the number into different kind of form. For example, if the number is *1000*, then it would be transferred into *1K*. The purpose of this is to increase the range of the understanding of the number in all forms.

The other vector, on the other hand, serves as the role that teaches the model to identify different numbers. It helps to enlarge the difference between the different numeric expressions, we call it negative distorted sample. Therefore, we also apply ChatGPT with the prompt[1] to change the number from every news article.

Positively Distorted
`30K Walmart Part-Timers to Lose Health Insurance.`
`Thirty thousand Walmart part-timers to Lose Health Insurance`
Negatively Distorted
`Dax Shepard: Wedding to Kristen Bell Cost $142.`
`Dax Shepard: Wedding to Kristen Bell Cost eight hundred`

Table 1: Examples of positively and negatively distorted headlines

3.2 Encoder

Transformer Seq2Seq Model (Vaswani et al., 2017) revolutionized the field of sequence-to-sequence learning. The implementation of the self-attention mechanism allowed weighting of the importance of different input tokens during the generation of each output token. The creation of multi-head attention enhanced the ability of the model to capture the diverse relationship between tokens. Based on transformer model advantages, the pre-trained BART-base model (Lewis et al., 2020) was selected as an encoder for headline representation creation. For the comparison, the T5 (Raffel et al., 2020) model was also utilized.

3.3 Models

In this section, we aim to delineate the types of models we trained and the portion of data utilized for training. The first model, referred to as **BART(sub)**, represents the outcomes of a model submitted for evaluation. In this model, the hyperparameters of the pre-trained BART model with Contrastive Learning (CL) were fine-tuned. Following the submission, our focus shifted towards improving and adjusting the Contrastive Learning approach. The subsequent model, named **BART with CL**, was trained using improved contrastive learning techniques. However, due to resource constraints, it was trained solely on **1000** instances of data.

3.4 Contrastive Loss

Our proposed Contrastive Learning enhanced model was implemented on End-to-End Seq2Seq

[1] You are the examiner. Examine the text. If the number in any form appears in the text, change the number into another number. Return the revised text only.

generation model. For the implementation of CL, positive and negative samples of headlines were explicitly constructed. Given news and headline, we trained End-to-End Seq2Seq models to generate headlines based on ground truth headlines. For the CL part, the model uses news, 2 positive and 2 negative headline samples. Given that our main task is specifically numerical aware headline generation, the sampling method was chosen to put more attention to numbers. That is why during the model training, the samples for the batch were exclusively formed from the 2 positive and 2 negative distortions of the same headline. This encourages the model to preserve the semantic content of the headline while allowing variations in numerical values. By explicitly focusing on numerical distortions in the loss function, the model learns to generate headlines that are robust to variations in numerical values. The loss function of the positive pair of examples (i, j) is defined as:

$$\mathrm{L}_s = -\lambda \log \frac{\exp(\tau^{-1} sim(z_i, z_j))}{\sum_s I(s \neq i) \exp(\tau^{-1} sim(z_i, z_j))})$$

where: $I(\cdot)$ is an indicator function such that $I(s \neq i) = 1$ and $I(s = i) = 0$, and τ is a temperature parameter and (s) is an index variable representing the current sample being considered during the training process.

This loss function penalizes the model if the distance between the news and positive headline embeddings is not closer than the distance between the news and negative headline embeddings by a certain margin. The final loss function for the numeric headline generation task will consist of a combination of model loss L_{model} and contrastive loss L_{CL} where the β is hyperparameter. Model Loss is the loss calculated from the model's forward pass using the ground truth labels.

$$Loss = L_{model} + \beta \times L_{CL}$$

3.5 Evaluation Metrics

For the evaluation of trained models, the automatic evaluation metric that the task organizer proposed was utilized. It consists of the ROUGE metric, incorporating ROUGE-1, ROUGE-2, and sentence-level ROUGE-L. For the BERTScore it incorporates BERT Precision, BERT Recall, and BERT F1. Also, the overall, copy and reasoning numerical accuracies were calculated.

3.6 Implementation and Hyperparameters

Due to resource constraints, all our models were trained only on **1000** instances of training data. For each model, we established the maximum length for both the article and the target headline, with values set at 512 and 16 respectively. In our trials, we adapt BART-large, T5-large and our newly introduced CL-augmented model. When tackling the headline generation task, we utilize beam search with a beam size of 8 and configure our batch size to 4. For the contrastive loss, the margin was set as 0.5, and β of 0.5 was selected. We apply the Adam optimizer with a learning rate of 5×10^{-6}. The models undergo training for 10 epochs, with the validation set used to assess performance. All experiments were conducted on the Kaggle T4 GPU.

4 Experimental Results

Table 1 shows the performance from different headline generation models evaluated by ROUGE score. BART model trained with contrastive loss has achieved 35.74, which is higher than other baseline models, showing its effectiveness on headline generation. Comparing BART with and without contrastive loss (CL), we observe a notable improvement in ROUGE scores when contrastive loss is incorporated during training. BART with CL achieves the highest ROUGE-1 at 40.91 and ROUGE-2 at 17.49 scores. Results indicate that contrastive loss regularization enhances the model's ability to generate headlines with higher lexical overlap and coverage.

Table 2 is the BERTScore for each headline generation model. Its BERT F1 score of 41.77 reflects strong semantic similarity to reference summaries, indicating robust performance across both lexical and semantic dimensions.

Model	Rouge-1	Rouge-2	Rouge-L
BART	38.48	15.18	33.35
T5 with CL	36.72	14.31	32.58
BART(sub)	31.22	12.23	26.86
BART with CL	**40.91**	**17.49**	**35.74**

Table 2: ROUGE scores of Headline Generation Models

Model	P	R	F1
BART	33.4	45.61	33.46
T5 with CL	35.50	39.90	37.72
BART(sub)	19.53	**47.56**	33.13
BART with CL	**36.91**	46.67	**41.77**

Table 3: BERT scores of Headline Generation Models

Model	Overall	Copy	Reasoning
`BART with CL`	72.956	82.170	56.176

Table 4: Numerical accuracy evaluation results

Additionally, BART with CL exhibits a BERT Precision of 36.91 and a BERT Recall of 46.67, further emphasizing its balanced performance in capturing semantic content accurately.

5 Conclusion

Resolving Task 7 at SemEval-2024 as team Challenges, we applied contrastive learning techniques on several models, in order to see which obtain the model with highest performance. In the final submission we obtained the highest number accuracy in the COPY category, up to 82.170, and got the second place in overall score, also up to 72.956. In our human evaluation process, our headline generation model achieved the third-highest level of numerical accuracy by reaching 1.70 score. It thus proves that our approaches can help train the model in numerical reasoning and numerical headline generation. However, this model is merely trained on the **small part of dataset**. Therefore, in future work, it is suggested that more instances might help all the conventions. If we enlarge the data size, it is possible that the performance may get higher. Augmenting positive and negative headline samples artificially may enhance the effectiveness of numeric-based headline generation. In our future work, we are planning to generate positive and negative samples for the whole dataset, and train models.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Numhg: A dataset for number-focused headline generation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Hui Liu, Weidong Guo, Yige Chen, and Xiangyang Li. 2022. Contrastive learning enhanced author-style headline generation.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.