SemEval-2024 Task 7: NumEval Task 3: Numeral-Aware Headline Generation (English)

Anonymous ACL submission

Abstract

Numerical reasoning is a challenging task even with large pre-trained language models. In this task, I show that some T5 models are capable of generating relevant headlines based on given news including proper numerical values. However, models sometimes have poor reading comprehension and miscalculate numerical values. To overcome those issues, I split the training process into two steps. In the first step, I trained the models to generate the calculation methods and in the second step, with those methods as an input, I trained them to produce numerical values to fill the blank in news headlines. Flan-T5 produced appropriate numbers with the accuracy of 90.2%.

1 Introduction

Comprehension of numerical values can significantly enhance performance in certain tasks as numbers provide important information in words. Numerical values are particularly important in accounting and finance fields as the majority of data is in monetary terms. While words can be ambiguous, numbers provide clear and precise information. They not only represent exact numerical values, but can also indicate a magnitude of the subject matter, which can be critical to fully understand the context of the texts.

Despite the significance, analyzing words within a text in Natural Language Processing has given a little consideration to numerical data in the past. This may be because all the texts are converted to numerical values in Natural Language Processing and it is hard to distinguish numerical values converted from words with the numbers in the original texts. Numerical reasoning is particularly difficult even for pre-trained language models.

The task of headline generation takes the form of text summarization, but accurate numeral generation in news headlines is still a challenge. In this task, which consists of two subtasks, I finetuned pretrained models to predict numerical values in news headlines. In the first subtask, which focused on numerical reasoning, models are required to compute the correct number to fill the blank in a news headline while the second subtask requires to construct an entire headline based on the provided news.

2 Related Work

2.1 NumNet

Ran et al. Ran et al. (2019) proposed a numerical Machine Reading Comprehension model named NumNet, which utilizes a numerically-aware graph neural network to make numerical comparison and performs numerical reasoning over numbers in the question and passage. Their NumNet model achieved the numerical reasoning ability with Exact Match (EM) of 64.56 and numerically-focused F1 score of 67.97 on the test data. However, it has a major limitation that NumNet is not applicable to the case where an intermediate number has to be derived such as from arithmetic operation in the reasoning process.

2.2 GENBERT

Geva et al. Geva et al. (2020) proposed a general method for injecting additional skills into Language Models, assuming automatic data generation is possible. They applied their approach to the task of numerical reasoning over text, using a general-purpose model called GENBERT, and a simple framework for generating large amounts of synthetic examples. Their experiments demonstrated the effectiveness of their method, showing that GENBERT successfully learns the numerical skills, and performs on par with state-of- the-art Numerical Reasoning Over Text models of the same size.

2.3 Arithmetic-Based Pretraining

Petrak et al. Petrak et al. (2023) proposed a new extended pretraining approach called Arithmetic-Based Pretraining that jointly addresses both in one extended pretraining step without requiring architectural changes or pretraining from scratch. Arithmetic-Based Pretraining combined contrastive learning to improve the number representation, and a novel extended pretraining objective called Inferable Number Prediction Task to improve numeracy. Their experiments showed performance improvements due to better numeracy in three different state-of-the-art pretrained language models, BART, T5, and Flan-T5, across various tasks and domains, including reading comprehension, inference-on-tables, and table-to-text generation.

3 Data

3.1 Subtask 1: Fill the Blank In News Headline

The training dataset consists of 21,157 news articles with masked headlines and the validation dataset consists of 2,572 news with masked headlines. Both the training and validation datasets have four columns consisting of "news", "masked "headline", "calculation" and "answer" as shown in Table 1. The numerical values which should be predicted in masked headline are shown in underscores. The calculation column shows the operations how to get to the answers including copy, round, paraphrase, convert number words to numbers and arithmetic operations, or some are a combination of multiple operations.

3.2 Subtask 2: Headline Generation

The training dataset consists of 21,157 news articles with headlines and the validation dataset consists of 2,365 news articles with headlines. The datasets for subtask 2 do not have the calculation column.

4 Methodology

4.1 Models

4.1.1 Masked Language Model

Masked language models such as BERT can predict a masked token to fill the blank in news headline. RoBERTa is a variant of BERT, which is a transformer-based language model that uses self-attention to process input sequences and generate contextualized representations of words in a sentence (pawangfg, 2023). But RoBERTa was trained on nearly 10 times more data than the original BERT. Also RoBERTa uses a dynamic masking technique during training that helps the model learn more robust and generalizable representations of words. DistilRoBERTa is a distilled version of the RoBERTa-base model.

4.1.2 T5 Language Model

T5 is Text-to-Text-Transfer-Transformer model. It takes input texts for various NLP tasks and outputs text for that respective task. T5 is slightly different from Masked Language Model such as BERT. Masked Language Models are Bidirectional models which use Mask token for each word while T5 replaces multiple consecutive tokens with a single Mask keyword (Mishra, 2020). Since the final objective is to have trained a model that inputs text and outputs text, the targets were designed to produce a sequence, unlike BERT, that tries to output one word. Therefore, T5 is suitable for text summarization and headline generation.

Michal Pleban trained the T5-base on a collection of 500k articles with headings named T5-base-en-generate-headline (Pleban, 2020). Its purpose is to create a one-line heading suitable for the given article. I further trained Michal's model on my training dataset.

Caleb Zearing trained T5 using a large collection of Medium articles (Zearing, 2022), and its objective is to generate article titles. I also trained Caleb's model on my training dataset.

4.1.3 Flan-T5 Model

Flan-T5 is an enhanced version of T5 that has been finetuned on a mixture of tasks (Chung et al., 2022). I particularly used LaMini-Flan-T5-783M, which is a fine-tuned version of google/flan-t5-large on LaMini-instruction dataset that contains 2.58M samples for instruction fine-tuning (Wu et al., 2023).

4.2 Subtask 1: Fill the Blank In News Headline

4.2.1 DistilRoBERTa

I initially trained DistilRoBERTa on the training dataset so the model can predict numerical values. I first converted the underscores to a mask token in the masked headline. I removed the time stamps in the news column and combined the news, masked headline and calculation columns to create a new column, which I then used as an input. I trained it on the dataset with the answer column as a target and based on the learning rate of 5e-5. The first numerical values from top 20 highest probability vocabulary tokens are extracted to fill the blank in the news headlines.

4.2.2 T5 & Flan-T5 Models - Train in One Step

I further trained two T5-based models along with Flan-T5 model on my training set. I first converted the underscores in the masked headline with an extra token and combined it with the news column for an input. Unlike DistilRoBERTa, the calculation column was not included in the inputs since it deteriorates the model performance. Two T5 models were trained with the learning rate of 5e-5 while Flan-T5 was trained with the learning rate of 2e-5. To extract a numerical value to fill in blank, an index of the extra token was searched from each masked headline and used to locate in the predicted sequence.

4.2.3 T5 & Flan-T5 Models - Train Twice in Two Steps

In the first step, I used the news and the masked headline columns as an input and the calculation column as a label to train the models. In the second step, I set the calculation column as an input and the answer column as an output to train the models. I then made predictions in 2 steps as shown in Figure 1. I used the model in the first step to predict calculations and with those predicted calculations as an input, I used the model in the second step to predict answers to fill the blank in the news headlines.

4.3 Subtask 2: Headline Generation

I trained the T5 models on my training set with the news as inputs and the headlines as labels. I prefixed the input with a prompt so T5 knows this is a headline generation task. Both models were trained with the learning rate of 5e-5. I also tried Flan-T5, but since I could not observe

news	masked headline	calculation	ans
(Apr 18, 2016 1:02 PM CDT) Ingrid Lyne, the Seattle mom allegedly murdered while on a date, left behind three daughters—and a GoFundMe campaign set up to help the girls has raised more than $222,000 so far, Us reports. A friend of the family set up the campaign, and says that all the money raised will go into a trust for the girls, who are ages 12, 10, and 7. Lyne's date was charged with her murder last week.	$____K Raised for Kids of Mom Dismembered on Date	Paraphrase(222,000,K)	222

Table 1: Sample Data for Subtask 1

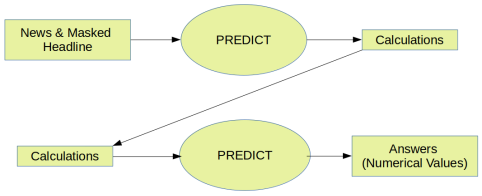

Figure 1: Train and Predict in 2 Steps for Subtask 1

Accuracy	One Step	Two Steps
DistilRoBERTa	0.798	N/A
T5-Headline-Michal	0.877	0.879
T5-Title-Czearing	0.878	0.881
LaMini-Flan-T5-783M	0.886	**0.902**

Table 3: Accuracy for Subtask 1

Actual	Predicted
Round(Divide(268,30),0)	Copy(9)
Round(1.29,0)	Span(a trillion)
Subtract(Sep 5,July 8)	Subtract(30,7)
Add(22,Trans(four))	Add(Trans(four),22)
Subtract(2014,1974)	Subtract(2018,1974)
Multiply(Trans(one-quarter),100)	Multiply(Divide(Trans(one-quarter),100)

Table 4: Sample of Wrong Calculations Generated by LaMini-Flan-T5-783M

contrasting results, I made an analysis of the headlines generated by the T5 models only.

5 Results and Evaluation

5.1 Subtask 1: Fill the Blank In News Headline

To evaluate the quality of the model, we need to calculate the probabilities it assigns to the next word in all the sentences of the test set. Perplexity, defined as the exponential of the cross-entropy loss, is a metric used to evaluate the model. As shown on Table 2 below, perplexity decreased significantly after training all models.

Finetuned	Before	After
DistilRoBERTa	6.23	3.68
T5-Base-Michau	2.66	1.05
T5-Czearing	2.14	1.05

Table 2: Perplexity for Subtask 1

The results for subtask 1 are evaluated by the accuracy and are shown in Table 3. Training in two steps did not improve the performance with two T5 models, but with Flan-T5 model, it slightly improved the results. When trained in a single step, two T5 models worked well, but when trained in two steps, Flan-T5 model worked the best and achieved the accuracy of 90.2%.

I further analyzed if there are certain patterns that are easy for the models to make mistakes with. I noted that mistakes are often made with the arithmetic operations, rounding the decimal numbers and the combination of various operations as shown in Table 4 below.

5.2 Subtask 2: Headline Generation

To evaluate the models, I used the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score to calculate the similarity between the predicted headlines and actual headlines.

As shown on Figure 2, T5 model by Caleb slightly worked better than T5 model by Michal. Of 2,365 validation data, with Caleb's model, there are 700 predictions with rougeL of 0.5 and over, and there are 36 perfect matches. Even with lower ROUGE scores, I noted that the models successfully included the numerical values in most of the generated headlines. For the predictions with low ROUGE scores and without proper numerical values, I manually analyzed to see if I can identify some patterns. I noted that the models often captured the meanings but just used different wordings with similar meanings as shown in Table 5, which is not taken into account with ROUGE scores.

BERTScore also measures the similarity between sentences, but it computes the cosine similarity between the contextualized embeddings of the words in the reference sentences (Mansuy, 2023). Thus it can capture semantic similarity between sentences. Figure 3 shows the scaled BERTScores for two T5 models, which have similar trends as the Rouge Scores.

	Actual	T5 Michal	T5 Caleb
1.	25% of Freed Gitmo Detainees Returned to Terror	1 in 4 Ex-Gatoramo Detainees Linked to Terrorists	1 in 4 Guantánamo Detainees Linked to Terror
2.	3rd Victim Dead in Quarry Shooting; Manhunt Still On	3 Killed in California Quarry Shooting Spree	3 Dead in California Quarry Shooting Spree
3.	Cop Finds Driver Playing Pokemon Go on 8 Phones	Cop Finds Driver Playing Pokemon Go on 8 Phones	Driver Playing Pokemon Go on 8 Phones: Trooper
4.	Tucson Cops Search for Missing Girl, 6	6-Year-Old Missing in Tucson	Tucson Cops Search for Missing Girl, 6
5.	NBC Paid Chelsea Clinton $600K a Year	NBC Paid Chelsea Clinton $600K a Year	NBC Paid Chelsea Clinton $600K a Year
6.	She Offered $25K to Catch a Killer. Now, She's Charged	Woman Offers $25K Reward in Husband's 2006 Murder	13 Years Later, Cops Say She Killed Her Husband
7.	Ex-Congressman Caught With $90K in Freezer Is Guilty	Ex-La. Rep Convicted of Taking $90K in Bribes	Former Louisiana Rep Convicted of Stealing $90K in Bribes
8.	We Drink and Drive an Estimated 121M Times a Year	1 in 5 Adults Admit Driving While Under the Influence	1.8% of US Adults Admit Driving While Impaired
9.	Their Film Ran in 14 Theaters. Then Robert Pattinson Called	Robert Pattinson Leads the 30-something Brothers in Good Time	Robert Pattinson's 'Good Time' Is Just the First 5 Minutes
10.	Alec Baldwin Collects $1.4K Every Time He Plays Trump	Alec Baldwin's Trump Impersonation Is 'Puffs'	Alec Baldwin's Trump Impersonation Is a SNL Sting

Table 5: Sample of Headline Generation by T5 Models

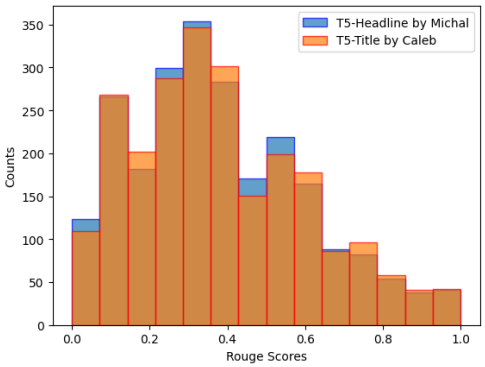

Figure 2: Rouge Scores of Headlines by T5 Models for Subtask 2

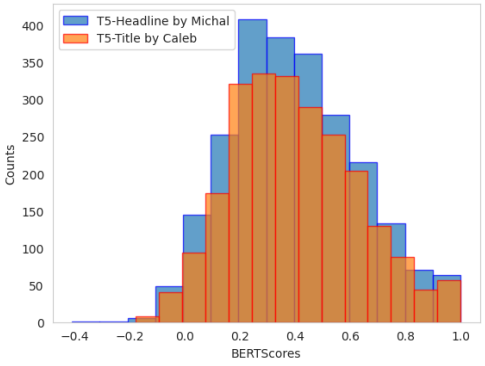

Figure 3: BERTScores of Headlines by T5 Models for Subtask 2

Table 5 shows samples of the headlines generated by T5 models, which were selected based on the criteria discussed below. As items 1 and 2 of Table 5 show, both models properly included the numerical values and captured the meanings, but the expressions of the numerical values and the wordings are different. There are several headlines perfectly generated by T5-Michal but not by T5-Caleb like item 3, whereas there are various headlines perfectly predicted by T5-Caleb but not by T5-Michal such as item 4. Item 5 is an example of the perfect generations by both models. Item 6 is an interesting example. Based on the news, the woman who offered a $25K reward for information on her husband's killer was arrested after 13 years since she was the killer. T5-Michal properly captured $25K reward, but failed to mention the fact that she was the one who got arrested. T5-Caleb properly captured this information, but it did not include the $25K reward part. The predictions for item 7 made by both models are very close to the actual headline, but the actual headline definitely draws people's attention and drives curiosity. For items 8 and 9, both T5 models failed to capture the appropriate numerical values. Item 10 is an example that both models failed to include any numerical value in the headlines.

6 Conclusion

T5 language models seem capable of generating meaningful headlines including appropriate numerical values. Although the models can reasonably compute the correct numbers from the provided news to fill the blank in headlines, they sometimes failed reading comprehension and arithmetic operations. In hope of overcoming those limitations, I trained them to generate the calculation methods first and then trained again with those calculations as inputs to predict the numerical values to fill the blank in the news headlines, but it did not significantly improve the results. In the future, I plan

to try larger pre-trained models, which might improve performance. Also, the training datasets that I used are relatively small. If I increase the data size by data augmentation, I may be able to obtain better results.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex an Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Raphael Mansuy. 2023. Evaluating nlp models: A comprehensive guide to rouge, bleu, meteor, and bertscore metrics.

Prakhar Mishra. 2020. Understanding t5 model : Text to text transfer transformer model.

pawangfg. 2023. Overview of roberta model.

Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. 2023. Arithmetic-based pretraining improving numeracy of pretrained language models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 477–493, Toronto, Canada. Association for Computational Linguistics.

Michal Pleban. 2020. t5-base-en-generate-headline.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine reading comprehension with numerical reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2474–2484, Hong Kong, China. Association for Computational Linguistics.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-flan-t5-783m.

Caleb Zearing. 2022. article-title-generator.