# hinoki at SemEval-2024 Task 7: Numeral-Aware Headline Generation (English)

**Hinoki Crum** and **Steven Bethard**
School of Information
University of Arizona
{hinokicrum,bethard}@arizona.edu

## Abstract

Numerical reasoning is challenging even for large pre-trained language models. We show that while T5 models are capable of generating relevant headlines with proper numerical values, they can also make mistakes in reading comprehension and miscalculate numerical values. To overcome these issues, we propose a two-step training process: first train models to read text and generate formal representations of calculations, then train models to read calculations and generate numerical values. On the SemEval 2024 Task 7 headline fill-in-the-blank task, our two-stage Flan-T5-based approach achieved 88% accuracy. On the headline generation task, our T5-based approach achieved RougeL of 0.390, BERT F1 Score of 0.453, and MoverScore of 0.587.

## 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 a text.

Despite the significance of numerical values, much natural language processing work has treated numerical words in the same manner as all other words, without any direct understanding of the values they represent. As a result, numerical reasoning is still challenging for natural language processing models, even the pre-trained language models that have been so successful on other natural language processing tasks.

NumEval (Chen et al., 2024) provides shared tasks that encourage research systems to generate headlines with accurate numeral information. We fine-tuned pre-trained models for two sub-tasks. In the first, models are required to compute the correct number to fill the blank in a news headline given the corresponding news article. In the second, models are required to construct an entire headline (including its numerical information) based on the provided news article.

## 2 Related Work

A Math Word Problem (MWP) consists of a short natural language narrative describing a state of the world and poses a question about some unknown quantities Patel et al. (2021). The MWP task is a type of semantic parsing task where given an MWP the goal is to generate an equation, which can then be evaluated to get the answer. The task is challenging because a machine needs to extract relevant information from natural language text as well as perform mathematical reasoning to solve it. Patel et al. (2021) proved in their paper that the existing models can rely on superficial patterns present in the narrative of the MWP and achieve high accuracy without even looking at the question.

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 some numerical reasoning ability with Exact Match (EM) of 64.56 and numerically-focused F1 score of 67.97 on the test data. However, NumNet is not applicable when an intermediate number has to be derived in the reasoning process such as from arithmetic operation.

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

*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 34–39
June 20-21, 2024 ©2024 Association for Computational Linguistics

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 similarly sized state-of-the-art numerical reasoning over text models.

Petrak et al. (2023) proposed arithmetic-based pre-training that combines contrastive learning to improve the number representation, and a novel inferable number pre-training objective to improve numeracy. Their experiments showed performance improvements due to better numeracy in three different state-of-the-art pre-trained language models, BART, T5, and Flan-T5, across various tasks and domains, including reading comprehension, inference-on-tables, and table-to-text generation.

Peng et al. (2021) proposed a novel pre-trained model, namely MathBERT, which is the first pre-trained model for mathematical formula understanding. MathBERT was jointly trained with mathematical formulas and their corresponding contexts to evaluate three downstreamtasks, including mathematical information retrieval, formula topic classification and formula headline generation. Formula headline generation is a summarization task aiming to generate a concise math headline from a detailed math question which contains math formulas and descriptions. In addition, in order to further capture the semantic-level structural features of formulas, a new pre-training task is designed to predict the masked formula sub-structures extracted from the Operator Tree (OPT), which is the semantic structural representation of formulas.

# 3 Data

## 3.1 Subtask 1: Headline Fill-in-the-Blank

The training dataset (Huang et al., 2023) consists of 21,157 news articles with masked headlines and the validation dataset consists of 2,572 news articles 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 the masked headline are shown in underscores. The calculation column shows the operations required to get to the answers, such as copy, round, paraphrase, convert number words to numbers, and arithmetic operations. The calculation may also be a combination of multiple operations.

The test set consists of 4,921 news articles with masked headlines without the calculation and answer columns.

## 3.2 Subtask 2: Headline Generation

The training dataset consists of 21,157 news articles with unmasked headlines and the validation dataset consists of 2,365 news articles with headlines. The datasets for subtask 2 do not have the calculation column. The test dataset consists of 5,227 news articles.

# 4 Methodology

## 4.1 Models

We employed several different types of neural network models for these tasks.

**DistilRoBERTa** RoBERTa (Liu et al., 2019) is transformer network trained on 16GB of text with a masked language modeling objective, making it appropriate for fill-in-the-blank tasks like Subtask 1. RoBERTa follows the standard transformer formulation, using self-attention to process an input sequence and generate contextualized representations as the output sequence. DistilRoBERTa (Sanh et al., 2019) is a distilled version of the RoBERTa-base model.

**T5-Headline-Pleban** The Text-to-Text-Transfer-Transformer (T5) model is a transformer network trained on 750GB of text with a language modeling objective where multiple consecutive tokens are masked and the output is a sequence. Because T5 models are designed to produce a sequence, they are suitable for headline generation tasks like Subtask 2. T5-Headline-Pleban (Pleban, 2020) is a T5-base model that was further fine-tuned to predict headlines from articles using a collection of 500k articles.

**T5-Title-Zearing** (Zearing, 2022) is a T5-base model that was further fine-tuned to predict titles from articels using a collection of Medium articles.

**Flan-T5-LaMini** Flan-T5 is an enhanced version of T5 that has been finetuned on a mixture of tasks (Chung et al., 2022). LaMini-Flan-T5-783M is a fine-tuned version of google/flan-t5-large on the LaMini-instruction dataset that

| news | masked headline | calculation | answer |
|---|---|---|---|
| (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

contains 2.58M samples for instruction fine-tuning (Wu et al., 2023).

### 4.2 Subtask 1: Headline Fill-in-the-Blank

We trained three types of models for subtask 1.

#### 4.2.1 DistilRoBERTa

To construct the input for DistilRoBERTa, we concatenated the news text, masked headline, and calculation columns. The underscores we replaced with DistilRoBERTa's mask token, and time stamps were removed. We then trained DistilRoBERTa to predict the answer given this input, using a learning rate of 5e-5. At prediction time, we took the top 20 highest probability vocabulary tokens predicted by the model for the mask token, and returned the first numerical value.

#### 4.2.2 T5 One-Step

To construct the input for our one-step T5 and Flan-T5 models, we replaced the underscores in the masked headline with the token `<extra_id_0>` and concatenated it to the news text. Unlike DistilRoBERTa, we did not include the calculation in the input as we found it deteriorated model performance. We trained the two T5 models with a learning rate of 5e-5, and the Flan-T5 model with a learning rate of 2e-5. At prediction time, we found the index of the extra token in the model output and used that to extract the numerical value.

#### 4.2.3 T5 Two-Step

As Patel et al. (2021) demonstrated, if models rely on shallow heuristics to solve the majority of math problems without word-order information or question text, instead of training the models to have them directly predict numerical values from question texts, it might be more beneficial to train them to predict numerical values from formulas or calculation methods. Accordingly, we propose two-step models in which we constructed two training sets. For step one, we concatenated the news text and the masked headline as input, and used the calculation as output. For step two, we used the calculation as input, and the answer as output. We then trained two models, one on each dataset. At prediction time, we applied the step-one model to the concatenation of the news text and masked headline, then passed the output of the step-one model as the input to the step-two model, which then predicted the final answer. We used the same extra token processing and learning rates as in the T5 One-Step approach. This process is shown diagrammatically in Figure 1.

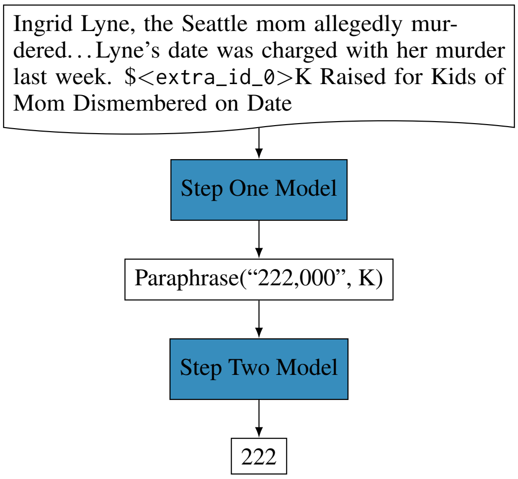
Ingrid Lyne, the Seattle mom allegedly murdered...Lyne's date was charged with her murder last week. $<extra_id_0>K Raised for Kids of Mom Dismembered on Date
Step One Model
Paraphrase("222,000", K)
Step Two Model
222

Figure 1: Two-step prediction for the headline fill-in-the-blank task.

### 4.3 Subtask 2: Headline Generation

We trained T5 models with the news text as input and the headline as output. We prefixed the input

| Data | Model | Before | After |
|---|---|---|---|
| Val | DistilRoBERTa | 6.23 | 3.68 |
| Val | T5-Headline-Pleban | 2.66 | 1.05 |
| Val | T5-Title-Zearing | 2.14 | 1.05 |

Table 2: Perplexity of models on the Headline Fill-in-the-Blank validation data

| Data | Model | 1 Step | 2 Steps |
|---|---|---|---|
| Val | DistilRoBERTa | 0.798 | N/A |
| Val | T5-Headline-Pleban | 0.877 | 0.879 |
| Val | T5-Title-Zearing | 0.878 | 0.881 |
| Val | Flan-T5-LaMini | 0.886 | **0.902** |
| Test | Flan-T5-LaMini | - | 0.88 |
| Test | GPT-3.5 baseline | | 0.74 |
| Test | Best system | | 0.95 |

Table 3: Accuracy of models on the Headline Fill-in-the-Blank validation and test data

with a prompt "headline: " so T5 knows this is a headline generation task. Both T5 models were trained with the learning rate of 5e-5. We also tried Flan-T5, but results were similar to the other T5 models, so we focused our analysis on the headlines generated by the T5 models only.

## 5 Results and Evaluation

### 5.1 Subtask 1: Headline Fill-in-the Blank

One measure of the quality of a model is perplexity, defined as the exponential of the cross-entropy loss over the probabilities the model assigns to the next word in all the sentences of the test set. As shown on Table 2, perplexity decreased significantly for all models after training.

A more direct measure of the models in the headline fill-in-the-blank task is accuracy, counting the fraction of times that the model's prediction of a numeric value exactly matched the expected numeric value in the data. Table 3 shows accuracy of the different models on the validation data. Training in two steps did not improve the performance of T5-Headline-Pleban or T5-Title-Zearing, but did slightly improve performance of Flan-T5-LaMini. The final row of Table 3 shows that the best model, two-step Flan-T5-LaMini, achieved 88% accuracy on the test data.

We manually analyzed the errors of the models on the validation data. Errors often revolve around arithmetic operations, rounding of decimal numbers, and the combination of operations. Table 4 shows examples of such errors.

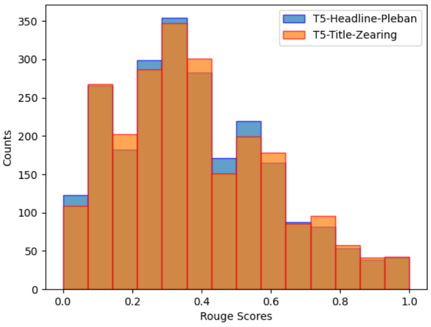

Figure 2: Rouge Scores of models on the Headline Generation validation data

While Patel et al. (2021) achieved about 65% accuracy from their best model, we achieved on the validation dataset the accuracy of 82% on predicting correct formulas while 88% on predicting correct numerical values from those formulas. We also noted that the accuracy on predicting right answers from correctly predicted formulas is 96%. This indicates that the models have no problem with making predictions from simple heuristics, which agrees with the findings by Patel et al.

### 5.2 Subtask 2: Headline Generation

We evaluated headline generation models based on how well their generated headlines matched the headlines in the data. We used two metrics, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and BERTScore. Both of these metrics measure the similarity between the predicted headlines and actual headlines, with the former relying on word n-grams and the latter relying on cosine similarity over contextualized embeddings derived from BERT (Mansuy, 2023). Figures 2 and 3 show the distribution of scores of the different T5 models over the validation data. The models are similar in terms of ROUGE score, but T5-Headline-Pleban performs slightly better than T5-Title-Zearing in terms of BERTScore.

We also used the official scoring script, producing the results shown in the first two rows of Table 5, where we see that T5-Title-Zearing is slightly better than T5-Headline-Pleban on the validation data for most measures. We thus submitted T5-Title-Zearing on the test set. The last row of Table 5 shows that it acheived 62.3% numerical accuracy

| 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: Examples of incorrect calculations generated by Flan-T5-LaMini on the Headline Fill-in-the-Blank data

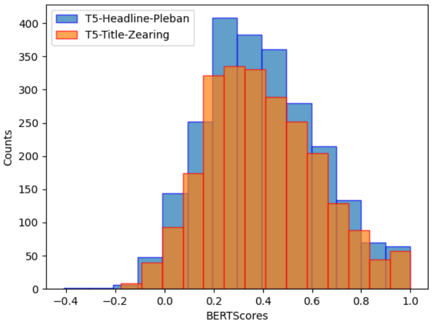

Figure 3: BERTScores of models on the Headline Generation validation data

on the test set with F1 scores of R1 of 43.1, R2 of 19.7 and RL of 40.0. MathBERT which trained with source texts, formulas and OPTs, achieved F1 scores of R1 of 61.25, R2 of 48.06 and RL of 57.72 on formula headline generation, which indicates that training the models with OPTs as inputs help improving the results.

We manually analyzed some of the errors of the models on the validation data. Table 6 shows examples of the headlines generated by T5 models. Items 1 and 2 show that both models properly included the numerical values and captured the meanings, but the expressions of the numerical values and the wordings are different. Several headlines were perfectly generated by T5-Headline-Pleban but not by T5-Title-Zearing, as in item 3, and vice versa, as in item 4. Item 5 is an example of perfect generations by both models. In item 6, a woman who offered a $25K reward for information on her husband’s killer was arrested as the killer after 13 years. T5-Headline-Pleban properly captured the $25K reward, but failed to mention that she was the one who got arrested, while T5-Title-Zearing did the opposite. The predictions for item 7 made by both models are close to the actual headline, but the actual headline is designed to better draw attention and drive 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, we 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, we plan to try larger pre-trained models, which might improve performance. Also, the training datasets that we used are relatively small. If we increase the data size by data augmentation, we may be able to obtain better results.

## References

Chung-Chi Chen, Jian-Tao Huang, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2024. Semeval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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.

| | | Numerical Accuracy | | | ROUGE | | | BERTScore | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | Model | Overall | Copy | Reasoning | 1 | 2 | L | P | R | F1 | MoverScore |
| Val | T5-Headline-Pleban | 48.8 | | | 43.7 | 20.1 | 39.5 | 46.0 | 44.8 | 45.4 | 59.0 |
| Val | T5-Title-Zearing | 51.2 | | | 44.1 | 20.3 | 39.9 | 47.3 | 44.7 | 46.0 | 59.2 |
| Test | T5-Title-Zearing | 62.3 | 66.3 | 55.1 | 43.1 | 19.7 | 40.0 | 47.2 | 43.4 | 45.3 | 58.7 |
| Test | Best system | 73.5 | 82.2 | 67.3 | 48.9 | 24.7 | 44.2 | 51.6 | 50.1 | 50.4 | 60.6 |

Table 5: Performance of models on the Headline Generation data

| | Actual | T5-Headline-Pleban | T5-Title-Zearing |
|---|---|---|---|
| 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 6: Examples of headlines predicted by models on the Headline Generation data

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.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? *arXiv preprint arXiv:2103.07191*.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding. *arXiv preprint arXiv:2105.00377*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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

Caleb Zearing. 2022. article-title-generator.