KyuHyunChoi at SemEval-2025 Task 10: Narrative Extraction Using a Summarization-Specific Pretrained Model

Kyu-Hyun Choi
Jeonbuk National University
South Korea
chlrbgus321@naver.com

Seung-Hoon Na
UNIST
South Korea
nash@unist.ac.kr

Abstract

Task 10 of SemEval 2025 was proposed to develop supporting information for analyzing the risks of misinformation and propaganda in news articles. In this study, we selected Subtask 3—which involves generating evidence explaining why a particular dominant narrative is labeled in an article—and fine-tuned PEGASUS for this purpose, achieving the best performance in the competition.

1 Introduction

Task 10 of Semeval 2025 (Piskorski et al., 2025; Stefanovitch et al., 2025) was proposed to support the research and development of new analytical functions aimed at analyzing the news ecosystem and characterizing manipulation attempts, in recognition that internet consumers are at risk of exposure to deceptive content and manipulation attempts, and that major crisis situations are also susceptible to the spread of harmful misinformation and propaganda.

The task focused on climate change and the Ukraine–Russia conflict as its main topics and provided three subtasks related to news articles. Subtask 1 involves assigning roles to each named entity mention in a news article using a predefined fine-grained role classification scheme. Subtask 2 requires assigning all appropriate subordinate narrative labels to a given article based on a two-stage narrative labeling system specific to a domain. Subtask 3 entails generating a free-text explanation, limited to 80 words, that provides evidence supporting the selection of the dominant narrative in an article.

In this study, we chose Subtask 3 and selected PEGASUS large (Zhang et al., 2020a) as the model best suited for this task. We fine-tuned PEGASUS large using the provided training dataset for this challenge. Section 2 explains the dataset and task for sub-task 3, Section 3 describes the model used and fine-tuning process, Section 4 details the hyperparameter settings and evaluation methods, Section 5 compares the results from Pegasus large with the baseline, and Section 6 concludes.

2 Background

2.1 Dataset

The training dataset provided for SemEval Task 11 Subtask 3 comprises news article text file titles, dominant narratives, subdominant narratives, and the target text to be generated. As our team employed the sequence-to-sequence model PEGASUS, we configured the input and output formats to align with the model's architecture. The input is constructed by listing the dominant narrative and the subdominant narrative separated by a space, followed by the news article prefixed with "Context:" at the end. The output is the target text to be generated. The input format is as follows:

$$Input = \{D; S; Prefix; Article\} \quad (1)$$

where D denotes the dominant narrative, S denotes the subdominant narrative, $Prefix$ is "Context:", and $Article$ represents the news article.

2.2 Task

Subtask 3 is a task that, given a news article along with its dominant and subdominant narratives, requires generating an explanation that provides evidence for why these narratives were selected. This task was designed with reference to the following studies: (Da San Martino et al., 2020; Piskorski et al., 2023, 2024). Since generating an explanation within 80 words based on the article and its narratives essentially amounts to a summarization task, our team selected PEGASUS large—a model well-suited for summarization—for this challenge.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2262–2264

3 System Overview

3.1 Model

PEGASUS is a model that employs the full transformer architecture and has been pre-trained specifically for the downstream task of summarization. Under the assumption that performing pre-training on tasks similar to the downstream task leads to better performance on that task, it was pre-trained using the Gap Sentence Generation (GSG) method, which is analogous to summarization. In this study, only PEGASUS large was used; unlike the base model, PEGASUS large was trained solely with GSG based on experimental results indicating that MLM is ineffective.

3.2 fine tuning

No special methods were used for fine tuning the model. The training followed the procedure described in Section 2.1, where the inputs from the training set were fed into the encoder and outputs were generated via the decoder. The model was saved only when the highest BERTScore F1 score (Zhang et al., 2020b) was achieved during training over several epochs.

4 Experimental Setup

4.1 hyper parameter setting

The model 'google/pegasus-large' was downloaded from Hugging Face, with the maximum input length set to 1024 and the maximum output length set to 128. The maximum number of epochs was set to 5, the batch size to 8, the learning rate to 3e-4, and the warmup rate to 0.00. AdamW was used as the optimizer, and the warm-up scheduler was implemented using the transformers' "get linear schedule with warmup" function. During validation, the BERTScore was used for evaluation, with roberta-large serving as the BERTScore model.

4.2 BERTScore

Subtask 3 of Task 11 is evaluated using the BERTScore. In this task, Precision is calculated to measure the similarity between the tokens of the generated sentence and those of the gold sentence, while Recall is computed to measure the similarity between the tokens of the gold sentence and those of the generated sentence. The performance of Subtask 3 of Task 11 is assessed using the F1 score, which is the harmonic mean of Precision and Recall. Let x represent the tokens of the generated sentence and y represent the tokens of the gold sentence. The similarity between the tokens of the generated sentence and those of the gold sentence is calculated as shown in Equation (2), with Precision computed as in Equation (3) and Recall as in Equation (4). Here, $\frac{1}{|C|}$ represents the total number of tokens in the generated sentence, and $\frac{1}{|R|}$ represents the total number of tokens in the gold sentence. The harmonic mean is calculated as shown in Equation (5).

model	Precision	Recall	F1 macro
ours	0.7669	0.7352	0.7504
baseline	0.6514	0.6834	0.6669

Table 1: Results of BERTScore.

$$S_{i,j} = \frac{\mathbf{x}_i \cdot \mathbf{y}_j}{\|\mathbf{x}_i\| \|\mathbf{y}_j\|} \tag{2}$$

$$P = \frac{1}{|C|} \sum_{x_i \in C} \max_{y_j \in R} S_{i,j} \tag{3}$$

$$R = \frac{1}{|R|} \sum_{y_j \in R} \max_{x_i \in C} S_{i,j} \tag{4}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{5}$$

5 Results

Our team conducted only Subtask 3 in English. With just a single training run, we secured first place, and the scores are as shown in the table 1. Compared to the baseline, our model achieved an increase of 0.11 points in precision, 0.05 points in recall, and 0.09 points in F1 macro score. It is evident that the improvement in precision was significant, and the F1 macro score benefited considerably from this enhancement. Instead of employing the latest decoder-only large language models, our team utilized PEGASUS-large—an encoder–decoder model pre-trained for summarization that was introduced in 2019—and even after five years since its release, it still demonstrates the best performance in this task.

6 Conclusion

Although it has been five years since the introduction of PEGASUS, our experiments have confirmed that it continues to exhibit robust performance, and the results of this study may offer valuable insights

to the organizers of SemEval Task 11. While the three subtasks of Task 11 have distinct characteristics, the third subtask can be effectively addressed by employing a summarization-specialized model such as PEGASUS.

References

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Jakub Piskorski, Alípio Jorge, Maria da Purificação Silvano, Nuno Guimarães, Ana Filipa Pacheco, and Nana Yu. 2024. Overview of the clef-2024 checkthat! lab task 3 on persuasion techniques.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.