NLP Assignment 2: Subtask C

Besher Hassan
MBZUAI
Leaderboard Team Name : Besher
Besher.Hassan@mbzuai.ac.ae

Abstract

This report presents my approach to Subtask C of SemEval-2025 Task 10 (Organizers, 2025), which focuses on generating concise and evidence-supported explanations for dominant narratives in English news articles. I experimented with multiple transformer models, including GPT-2 (small, medium, large) (Radford et al., 2019) and BART (base, large) (Lewis et al., 2019), and ultimately selected **BART-base** for its balance of performance and efficiency. By integrating contextual sentence extraction with "all-MiniLM-L6-v2" model is based on MiniLM (Wang et al., 2020) and fine-tuning BART-base, the model achieved a BERTScore (Zhang et al., 2020) **F1 of 0.8981** on the **(Test)** and **0.74662** on **(Dev)** .

1 Introduction

Narratives play a central role in shaping public opinion, especially in online news, where framing can influence perception. Understanding and explaining these narratives is critical for combating disinformation and aiding decision-makers. Subtask C focuses on generating concise explanations for dominant narratives in news articles using evidence from the text. This work develops a robust pipeline combining contextual sentence extraction and structured input formatting.

2 Dataset

The dataset comprises annotated articles from two domains: **Ukraine-Russia War (URW)** and **Climate Change (CC)**. Each entry contains:

- **article_id**: A unique identifier for the article.
- **dominant_narrative**: The overarching narrative conveyed by the article.
- **dominant_subnarrative**: A finer-grained narrative under the dominant narrative (or "none" if unavailable).
- **article_text**: The full text of the news article, with an average sentence length of 22 words.
- **explanation**: A human-written explanation (gold label) justifying the narrative assignment and the average sentence length is 20 words.

For the **Training Set** it contain 88 articles, split into 90% training and 10% validation and test for internal evaluation. And for the **Development Set** there are 30 articles without human-annotated explanations, and this set is not used for training.

3 Methodology

3.1 Key Features for Best Performance

The model incorporates several features that significantly enhanced performance:

1. **Contextual Sentence Extraction**: Using 'all-MiniLM-L6-v2', I extract the three most relevant sentences from each article based on cosine similarity with a query formed from the dominant narrative and subnarrative.
2. **Input Formatting**: The input to the model is structured in a fixed format, combining dominant narrative, subnarrative, and contextual sentences in a compact representation:

```
<N> Dominant Narrative </N>

<Sub-N> Subnarrative </Sub-N>

<Art> Contextual sentences </Art>
```

For example:

```
<N> Criticism of Climate Movement </N> <Sub-N>
Industrial Progress Opposition </Sub-N> <Art>
Activists oppose industrial development
without evidence.  </Art>
```

This compact representation ensures efficient processing and compliance with input constraints.

3. **Fine-Tuned BART-base**: Selected for its balance of performance and efficiency compared to alternatives like GPT-2 and BART-large.

4. **Custom Dataset Class**: Handles tokenization, truncation (512 tokens for inputs, 80 words for outputs), and batching.

5. **Training Configuration**:
   - Learning rate: 5×10^{-5}
   - Batch size: 8
   - Epochs: 10
   - Scheduler: Linear decay without warmup.

4 Experiments and Results

4.1 Quantitative Results

Multiple models were evaluated using BERTScore metrics (**Precision**, **Recall**, and **F1**). The baseline model is **GPT-2 small**, achieved an **F1 score of 0.7956**. Table 1 compares their performance, with **BART-base** achieving the highest **F1 score of 0.8981** on the **test set** and **0.74662** on the **dev set**.

Model	Size (M Params)	Precision	Recall	F1
GPT-2 Small	124	0.8032	0.7891	0.7956
GPT-2 Medium	355	0.8456	0.8302	0.8378
GPT-2 Large	774	0.8621	0.8434	0.8526
BART-Large	400	0.9042	0.8870	0.8955
BART-base (Best)	**140**	**0.9085**	**0.8879**	**0.8981**

Table 1: Performance of different models on the test set.

4.2 Qualitative Results

Example from the CC Domain:

- **Dominant Narrative**: Criticism of Institutions and Authorities.
- **Dominant Subnarrative**: Criticism of national governments
- **Explanation (gold label)**: The article talks about the resistance movement against Climate Con Programme initiated by the UK government. The article does not really specify what the residents are specifically resisting, as the article does not give insights into this programme.
- **Generated Explanation**: The text criticises the UK government for not doing enough to fight climate change in the UK. It criticises it for failing to do enough to address the problems of the UK's cities.

Example from the URW Domain:

- **Dominant Narrative**: Discrediting the West, Diplomacy
- **Dominant Subnarrative**: West is tired of Ukraine
- **Explanation (gold label)**: The text presents several paragraphs in which Ukraine's allies are depicted as weak or reluctant to provide further military help. These claims are justified by the upcoming US elections as well as the fact there are no enough funds.
- **Generated Explanation**: The text presents several paragraphs in which Western countries are accused of being the aggressors in the Ukraine conflict. The text criticises Western countries for not providing enough military support to Ukraine.

5 Discussion

The fine-tuned BART-base model demonstrated robust performance, effectively capturing dominant narratives, such as institutional criticism and geopolitical hesitancy. with an average sentence generation length of 18 words since the best maximum length was set to 80 and the gold labels average is 20. However, some limitations were observed in nuanced cases For the first example The explanation aligned with the dominant narrative but missed essential details about the Climate Con Programme. And for the second example The model misinterpreted the narrative, labeling the West as aggressors contrary to the gold standard, though it correctly identified ally hesitance. So to improve the performance we can use GPT-4o API to apply Data Augmentation, to generate alternative explanations for underrepresented cases and generate extra samples by paraphrasing the available samples. And to improve the Context Handling we could Incorporate contextual embeddings or refine preprocessing to improve sentence-level understanding.

6 Conclusion

This approach demonstrated an effective framework for explaining dominant narratives. The combination of contextual sentence extraction and fine-tuned BART-base supported high performance. Applying data augmentation and contextual embeddings can improve the performance of the model.

References

Mike Lewis, Yinhan Liu, Naman Goyal, and et al. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

SemEval Task Organizers. 2025. Semeval-2025 task 10: Explaining dominant narratives in multilingual news articles.

Alec Radford, Jeffrey Wu, Rewon Child, and et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.