# Improving Faithfulness in Abstractive Text Summarization with EDUs Using BART (Student Abstract)

**Narjes Delpisheh, Yllias Chali**

University of Lethbridge, Alberta, Canada
narjes.delpisheh@uleth.ca,yllias.chali@uleth.ca

## Abstract

Abstractive text summarization uses the summarizer's own words to capture the main information of a source document in a summary. While it is more challenging to automate than extractive text summarization, recent advancements in deep learning approaches and pre-trained language models have improved its performance. However, abstractive text summarization still has issues such as unfaithfulness. To address this problem, we propose a new approach that utilizes important Elementary Discourse Units (EDUs) to guide BART-based text summarization. Our approach showed the improvement in truthfulness and source document coverage in comparison to some previous studies.

## Introduction

Unfaithfulness refers to text that is inaccurate or misleading compared to the original source text. EDUs are foundational elements in discourse analysis, representing independent units of meaning within text. They capture complete thoughts and have found application in NLP and text summarization (i.e., enhancing the cohrence). Originating from Mann and Thompson (1988), EDUs define the smallest pragmatic discourse units. By grouping sentences into EDUs, inter-sentence relations like elaboration and causality can be identified, improving summary quality. This study enhances text summarization faithfulness by effectively integrating crucial EDUs with the source text, guiding BART-based summarization for improved accuracy and informativeness compared to previous methods.

## Methodology

Our inspiration comes from a previous manual evaluation (Narayan, Vlachos et al. 2019) where they discovered that highlighted content can assist human evaluations and reduce variation in results. We conclude that, as EDUs are semantic simpler units than sentences, encoding them is more accurate way and can improve the faithfulness of the generated summaries. Consequently, we propose new guidances including integral EDUs. The GSum (Dou et al. 2020) framework is a general and extensible approach that applies a guidance signal to determine which tokens should be generated.

To improve the guidance mechanism, we utilized a joint-training approach and enhanced the baseline summarizer model (BART) (Lewis et al. 2019). Firstly, we used an EDU extractor model (DISCOBERT) (Xu et al. 2019) to identify crucial EDUs from the source document. We selected the EDUs from each document that were labeled with one, indicating that they were important. Then, we employed the extracted EDUs as constraints to guide the summarizer model using GSUM (Dou et al. 2020) instantiated with BART to generate more faithful summary.

Abstractive text summarization involves encoding a source document $X$, comprising of multiple sentences $x_1, x_2, ..., x_{|X|}$, and generating a summary $y$ word by word. Guidance, denoted as $g$, is additional input to the model during training and inference. The model's parameters $\theta$ are optimized to maximize the likelihood of outputs $y$ given inputs $x$ and $g$. Two guidance methods are explored: automatic prediction using $x$, and oracle extraction using both $x$ and $y$ to enhance model attention during testing. Oracle guidance offers a notable advantage by providing informative signals for improved testing performance.

$$\arg\max_{\theta} \sum_{(x^i, y^i, g^i) \in \langle \mathcal{X}, \mathcal{Y}, \mathcal{G} \rangle} \log p(y^i \mid x^i, g^i; \theta) \qquad (1)$$

## Experimental Findings

In this study we make our experiments with CNN/Daily Mail dataset. Our summarization models are constructed with BART (Lewis et al. 2019) as the baseline summarizer using its default hyperparameter settings. GSum (Dou et al. 2020) instantiated with BART (Lewis et al. 2019) consists of 24 encoding layers, with the top layer initialized with pretrained parameters and separately trained for each encoder. While the first cross-attention block of the decoder is randomly initialized, the second cross-attention block is initialized with pretrained parameters. We provide two forms of guidance: during training, we use the oracle extracted guidance approach. During testing, we use MatchSum (Zhong et al. 2020) to extract the highlighted sentences and DISCOBERT (Xu et al. 2019) to identify important EDUs. We use different guidances in our experiment such as highlighted sentences, important EDUs, and combination of

| Model | R-1 | R-2 | R-L |
|---|---|---|---|
| BART+Sents(G) | 0.45 | 0.22 | 0.42 |
| BART+EDUs(G) | **0.54** | **0.31** | **0.50** |
| BART+Sents,EDUs(G) | 0.52 | 0.29 | 0.49 |
| BART | 0.44 | 0.21 | 0.41 |
| CLIFF | 0.44 | 0.21 | 0.41 |
| SEASON | 0.46 | 0.23 | 0.43 |

Table 1: Comparison of summarization using ROUGE (with 95% confidence interval)

highlighted sentences with important EDUs. After determining the guidance on the CNN/DM dataset, we fine-tune our model on BART with oracle-extracted guidance. We then predict the guidance during testing.

We compare our novel abstractive text summarization approach, with a previous study that uses highlighted sentences for guidance (Dou et al. 2020). We also benchmark against two other leading models, CLIFF (Cao and Wang 2021), which employs contrastive learning, and SEASON (Wang et al. 2022), which uses flexible salience-based guidance to overcome constraints of extractive text guidance.

We used automatic evaluation metrics, including ROUGE (Lin 2004), measuring important information retention; BERTScore (Zhang et al. 2019), capturing deep dependencies and paraphrases; Textual Entailment (Dagan, Glickman, and Magnini 2006), measuring factual consistency via document-to-sentence; QAGS (Wang, Cho, and Lewis 2020), using question-answer alignment to gauge summary-source consistency; and SentSim (Song, Zhao, and Specia 2021), comparing sentence similarity to evaluate information conveyance.

As shown in Table 1 and 2 guiding the BART model with EDUs improves its performance in both ROUGE and faithfulness evaluations, resulting in summaries that capture salient information and exhibit higher semantic coherence and similarity to reference summaries. In the tables, Sents represents sentences and G signifies guidance.

## Conclusion

This study improves BART-based summarization faithfulness by using guided content units (EDUs). Results show this approach outperforms models like CLIFF and SEASON, promising better summarization quality in an era of expanding content. Although pre-trained models have boosted summarization, this work offers a promising avenue to improve summarization quality and faithfulness. Its implications extend to research and practical applications in an era of expanding written content.

| Model | BERT | Entail | QAGS | SentSim |
|---|---|---|---|---|
| BART+Sents(G) | 0.89 | 0.925 | 0.067 | 0.85 |
| BART+EDUs(G) | **0.91** | 0.916 | 0.084 | 0.738 |
| BART+Sents,EDUs(G) | 0.90 | **0.929** | **0.093** | **0.854** |
| BART | 0.88 | 0.916 | 0.049 | 0.578 |
| CLIFF | 0.89 | 0.884 | 0.051 | 0.634 |
| SEASON | 0.89 | 0.914 | 0.083 | 0.761 |

Table 2: Comparison of summarization using different evaluation metrics (with 95% confidence interval)

## References


Cao, S.; and Wang, L. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2109.09209*.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, 177–190. Springer.

Dou, Z.-Y.; Liu, P.; Hayashi, H.; Jiang, Z.; and Neubig, G. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Mann, W. C.; and Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3): 243–281.

Narayan, S.; Vlachos, A.; et al. 2019. HighRES: Highlight-based reference-less evaluation of summarization. *arXiv preprint arXiv:1906.01361*.

Song, Y.; Zhao, J.; and Specia, L. 2021. SentSim: Crosslingual Semantic Evaluation of Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3143–3156. Online: Association for Computational Linguistics.

Wang, A.; Cho, K.; and Lewis, M. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5008–5020. Online: Association for Computational Linguistics.

Wang, F.; Song, K.; Zhang, H.; Jin, L.; Cho, S.; Yao, W.; Wang, X.; Chen, M.; and Yu, D. 2022. Salience allocation as guidance for abstractive summarization. *arXiv preprint arXiv:2210.12330*.

Xu, J.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhong, M.; Liu, P.; Chen, Y.; Wang, D.; Qiu, X.; and Huang, X. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.