Literature Review: DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence (Zhao et al., 2023)

Elia Mounier-Poulat | 314771 | elia.mounier-poulat@epfl.ch
REAL

1 Summary

This paper introduces ***DiscoScore***, a new parametrized discourse metric for evaluating text coherence.

The authors start by emphasizing the need for an accurate metric to evaluate the logical consistency of discourse-based Natural Language Generation (NLG) systems. They highlight the shortcomings of existing non-discourse NLG evaluation metrics such as ***MoverScore*** and ***BERTScore***, which mainly assess semantic similarity and are ineffective at capturing discourse coherence, leading to weak correlations with human-rated coherence. To address this gap, the authors introduce ***DiscoScore***, a metric that uses **BERT** to compare focus frequency and semantics between hypothesis and reference texts.

Furthermore, the authors propose an innovative approach that integrates contextualized embeddings with graph-based methods for their discourse metric; and present two primary variants of ***DiscoScore***: ***DS-FOCUS*** and ***DS-SENT***. ***DS-FOCUS*** models focus frequency and semantics by comparing differences between hypothesis and reference foci using a bipartite graph and measuring distance based on embeddings. On the other hand, ***DS-SENT*** employs a graph-based approach to model interdependence between sentences, aggregating sentence embeddings with adjacency matrices and calculating cosine similarity between graph-level embeddings, with the added advantage of weighted scoring.

The experimental setup involves testing 16 existing metrics, along with ***DiscoScore***, to provide a solid baseline for comparison. However, the evaluations were only conducted on two types of datasets: summarization (*SummEval* (Fabbri et al., 2021) and *NeR18* (Grusky et al., 2018)) and document-level machine translation (MT) (*WMT20* (Mathur et al., 2020)), which limits the assessment of the proposed metric.

The findings of the study reveal that non-discourse metrics based on **BERT** have limitations in adequately evaluating text coherence. On the other hand, , ***DiscoScore*** demonstrates consistent and robust performance in both single- and multi-reference settings, excelling not only in text coherence but also in various other aspects such as factual accuracy.

This paper introduces a promising method for assessing the coherence of generated text, making a significant contribution to the advancement of discourse metrics in the field. However, we will also explore areas where further improvements could be made.

2 Strengths

This paper presents a valuable contribution to the field of discourse metrics and has several strengths.

Firstly, the paper is well-written and easy to follow, with a clear explanation of the state of the art and history of discourse metrics. The authors make their code available and provide clear documentation for easy installation and use of ***DiscoScore***.

The authors provide a clear motivation for the need for an accurate metric to evaluate the coherence of discourse-based NLG systems. The proposed approach is methodical, detailing how both variants of ***DiscoScore*** work and how they differ from one another. The authors also explain how to interpret the difference in benchmark results for ***DS-SENT*** and ***DS-FOCUS***.

The experimental setup is well-designed and explores a variety of metrics for baseline comparison, encompassing both discourse and non-discourse metrics. These metrics were carefully selected to offer insightful perspectives on the performance of popular evaluation metrics in discourse coherence and their limitations.

In addition, the results are well stated and demonstrate the effectiveness of ***DiscoScore*** in capturing

discourse coherence, outperforming existing metrics in system-level correlation with human ratings. The authors' work outperforms ***BARTScore*** and ***BERT***-based metrics, which exhibit notably weaker correlations with human-rated coherence compared to earlier discourse metrics that were developed over a decade ago. These findings not only highlight the effectiveness of the authors' contributions but also pave the way for future research directions, such as the importance of incorporating discourse-level features in NLG evaluation metrics and the potential benefits of combining contextualized embeddings with graph-based approaches to better capture the interdependence between sentences.

Finally, the authors provide a thorough analysis of the strengths and weaknesses of ***DiscoScore***, acknowledging areas for future research and improvement. This self-critical approach strengthens the credibility of the study.

3 Weaknesses

The paper presents several strengths, but there are also some weaknesses that should be acknowledged.

One weakness is the disproportionate ratio between the introduction and related work sections compared to the results and analysis sections. The former sections are more extensive, which could have been balanced with less detailed information about related works.

Another weakness is that the experimental setup only tests ***DiscoScore*** on two types of datasets: summarization and document-level machine translation. While these are important application areas, this limited scope may not fully demonstrate the potential of the metric, as it could be applied more widely in NLG, for instance, in question answering or storytelling.

In addition, the authors acknowledge that they only investigated two focus choices popular in the discourse community: noun and semantic entity. There could have been more options to explore, such as verb-based focus or even sentence-level focus. The authors also suggest that there are weaknesses in the assessment of the metric, such as discourse connectives and coreference, which could be factored into the assessment of text coherence.

Finally, the new metric presented is not perfect. The paper does not provide a detailed analysis of how ***DiscoScore*** could adapt to other languages, which could limit its applicability to non-English texts. Although the authors briefly mention that it should generalize well, they do not discuss this in depth. Additionally, ***DiscoScore*** requires a gold reference to compare the generated text to, which may not always be available.

4 Conclusion

Despite these weaknesses, the paper presents a valuable contribution to the field of discourse metrics and provides a promising approach for evaluating the coherence of generated text. Future work could address the identified limitations and further improve the metric for broader and more fine-grained applications.

References

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. Discoscore: Evaluating text generation with bert and discourse coherence.