Cell2Doc Supplementary Material

August 18, 2023

1 Cell2Doc Human Evaluation with CodeBERT

We have done the human evaluation of Cell2Doc using CodeT5 and CodeBERT as the CoDoc model. As CodeBERT is an encoder-only model, we have used a 6-layer transformer decoder with it to build the CoDoc model. We have only used CodeT5 in the main paper because it has performed better than CodeBERT overall. Figure 1 summarizes the human evaluation results using CodeBERT in CoDoc of Cell2Doc. Like CodeT5 [1], we can see that Cell2Doc also improves CodeBERT's effectiveness across all three dimensions.

Index	Model	Correctness	Informativeness	Readability
1	CodeBERT (CM)	$\mu = 3.00, \sigma = 1.34$	$\mu = 2.8, \sigma = 1.24$	$\mu = 3.65, \sigma = 1.18$
2	CodeBERT (CSM)	$\mu = 2.74, \sigma = 1.39$	$\mu = 2.73, \sigma = 1.29$	$\mu = 3.62, \sigma = 1.14$
3	CodeBERT (ECSM)	$\mu = 2.57, \sigma = 1.31$	$\mu = 2.57, \sigma = 1.28$	$\mu = 3.62, \sigma = 1.11$
4	CodeBERT (SCSCM)	$\mu = 3.10, \sigma = 1.30$	$\mu = 3.04, \sigma = 1.19$	$\mu = 3.72, \sigma = 1.10$
5	CodeBERT (Cell2Doc)	$\mu = \mathbf{3.81}, \sigma = 1.13$	$\mu = \mathbf{4.00}, \sigma = 1.11$	$\mu = \mathbf{4.22}, \sigma = 0.90$

Figure 1: Results of the human evaluation using CodeBERT in CoDoc of Cell2Doc

2 Other Input Representations

Inline code comments can provide additional information about the code, which can be utilized to create better documentation. We tested this input representation idea with the CodeBERT model in CoDoc, which can take both code and natural language as input. Code and comments present in a code cell can be represented as $[code\_snippet_1, comment_1, code\_snippet_2, comment_2, \cdots code\_snippet_n]$ where $code\_snippet_i$ are code segments (present between two consecutive comments) and $comment_i$ are the comments. We use it as input to CodeBERT by using a delimiter between code and comment. As target documentation, we use only summarized markdown in these cases. The BLEU [2] and ROUGE [3] (only F1 for ROUGE-1, ROUGE-2 and ROUGE-L) scores for this experiment are **20.37** and (**25.29**, **11.60**, and **28.24**) respectively. Compared with the automated evaluation where CodeBERT is used in CoDoc, we can see that the numbers are better when comments are included in the input with respect to other baselines. It does not beat the SCSCM representation, though, and the reason for it might be that SCSCM has more data points in the training set as it uses comments as separate labels, and comments serve as good quality labels as well.

References

[1] Y. Wang, W. Wang, S. Joty, and S. C. H. Hoi, "Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation," 2021. [Online]. Available: https://arxiv.org/abs/2109.00859

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[3] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013