MEMORANDUM

To: Northeastern University Applied AI Research Team, Natural Language Processing Domain
From: Marcus Roldan, Associate Professor, Applied AI Research Team
Date: 31 October 2023
Re: Automatic Documentation Generation for Source Code: Recommendation for Future Research

---

This memo recommends the continuation of two key research points into automatic documentation generation: enhancement of the DECOM model and creation of an evaluation model of generated documentation adequacy.

RESEARCH PROBLEM

Documentation Generation Models

While Machine Learning techniques were shown effective at method level documentation generation for Java source code, there was a heavy reliance on prefabricated templates [1], limiting the applicability of the technique. Large Language Models are susceptible to error accumulation in this task, as well as being unable to leverage higher-level information to enhance the quality of generated documentation. The development of the DECOM model, using a multi-pass deliberation strategy, improves applicability, reduces error accumulation, and incorporates global insight [2].

Model Evaluation

The evaluation process for these strategies typically utilizes human evaluation and automatic evaluation metrics adapted from computational linguistics. Because of time and resource limitations, human evaluations involve a smaller sample size in terms of the model data and number of human evaluators. Automatic evaluation using adapted metrics allows for larger samples, but often lacks the ability to interpret key dimensions of the documentation [2, Sec. 2.4]. After correlating computational linguistic metrics and human evaluations of source code and generated documentation pairs, evaluation of the adequacy of documentation was found to be lacking in the standard suite of metrics.

Exigence

By extending the most viable strategies for automatic documentation generation, the workflow of software development can be streamlined. Especially as the scale and importance of software increases, reducing the resources required to create meaningful documentation allows for a shift in focus to the more complex and critical aspects of software development. As the capabilities of Large Language Models are ever-increasing in recent years, Northeastern can establish itself at the forefront of a rapidly emerging field, demonstrating the utilities of such models to improve efficiency in a key industry of the Information Age.

METHODOLOGY

To properly examine the current state of automatic documentation generation, techniques at the method level and documentation generation as a translation task were examined. With the multi-faceted nature of this problem and the limited nature of this research process, it is essential to examine multiple recently developed strategies to make determinations about the most promising avenue to continue research. The examination was also extended to include the effectiveness of the standard suite of computational linguistic metrics most used in studies of this nature.

Method Level and Translation Paradigm

A method level tool (*Method Man*) [1] was examined, but the extension of this tool is limited by its fundamental generation strategy. Continuing at the method level, but through the lens of documentation generation as a translation task, utilization of the GPT-3 variant Codex in both translation directions were examined [3], [4]. The tool developed in [2] aims to enhance the strategy demonstrated in [3] and [4].

Model Evaluation Techniques

A commonality among the aforementioned sources was the use of a standard suite of evaluation metrics adapted from computational linguistics used to aid the analysis of different models and tools. [5] conducts a review of these metrics, aiming to correlate them with human evaluations, aiming to determine the overall efficacy of these metrics.

LITERATURE REVIEW

The tool *MethodMan*, a method-level documentation generation tool showed impressive speed but relied heavily upon a well crafted set of templates and machine learning token extraction techniques [1].

Both directions of the translation task paradigm, from Programming Language to Natural Language [3] and Natural Language to Programming Language [4], were possible though the use of the LLM model and GPT-3 variant Codex. This paradigm was found to yield higher quality results than [1].

After correlating human evaluations of generated documentation and source code with computational linguistic metrics commonly utilized across studies [5, Tab. 1], it was found that even the most effective metrics lacked the ability to capture the key aspect of documentation adequacy [2].

The DECOM model already outperforms other state-of-the-art models, reducing the possibility of error accumulation negatively influencing the produced documentation and leveraging global level source code information to produce higher quality documentation [2]. Prompt engineering was also identified as a key component, similarly to [3] and [4].

```
public CategoricalTable copy() {
    Map<Value, Double> newTable = new HashMap<Value, Double>();
    if(variable == null){
        variable = 1;
    }
    if(table.isEmpty()){
        return new CategoricalTable(variable);
    }
    for(Value v : table.keySet()){
        newTable.put(v, table.get(v));
    }
    return new CategoricalTable(variable, newTable);
}
```

Ground Truth:	returns a copy of the probability table
Initial Draft (DECOM):	constructs a new multivariate table from a univariate table
First-pass (DECOM):	creates a new copy of the given distribution
Second-pass (DECOM):	returns a copy of the table from this table
Third-pass (DECOM):	returns a copy of the probability table
Re^{2}com Production:	returns a copy of the bytes written to this stream
Rencos Production:	prunes all table values that have a probability lower than the threshold
Editsum Production:	creates a new copy of the given table of the table

Figure 1: an example of DECOM performance over multiple deliberations, along with the ground truth from the source documentation and a production from another state-of-the-art model, Re^{2}com [2, Fig. 1].

RECOMMENDATIONS FOR FURTHER RESEARCH

DECOM Enhancement: Reinforcement Leaning

As identified by the authors of [2], the introduction of reinforcement learning techniques into the transformer-encoder framework could result in an improvement of quality of generated documentation. This enhancement to the DECOM model should then be tested against the same metrics (BLEU, ROGUE-L, METEOR, CIDEr) and dataset from [2] to establish the existence and magnitude of improvements to produced documentation. The structure of the DECOM model is such that the addition of reinforcement learning is a sensible and viable extension of the model.

Evaluation Model Enhancement: Adequacy Metric

The development of a new evaluation metric which quantifies the adequacy of generated documentation, shown by [5] to be the next step in advancing the effectiveness of automatic evaluation metrics will be key in the evaluation of any enhancements made to the DECOM model. Adequacy encompasses both the usefulness and understandability of documentation, two critical components to facilitate understanding of source code. The authors of [5] outline the premise behind such a metric to be the quantification of the rate of overlap between the generated documentation and key tokens from the source code (APIs, identifiers, etc.). This metric should be evaluated against the same dataset used in [5] and compared in the same manner to the evaluation metrics used in [5] (BLEU, ROGUE-L, METEOR, CIDEr, and SPICE).

DECOM Testing

After reinforcement training has been incorporated into the DECOM model and an adequacy metric has been established, it is imperative that these new advancements be integrated and tested rigorously. Using the same datasets as [2] and [5], the enhanced DECOM model can be directly compared using five evaluation metrics (BLEU, METEOR, ROUGE-L, CIDEr, and SPICE), allowing for one-to-one comparison of the two model versions. Additionally, the newly developed adequacy metric will evaluate a novel dataset (from [2]), a key step in the development of this metric.

Adequacy Metric Testing

The newly developed adequacy metric can then be tested against the dataset in [5], following the same methodologies used to calculate the correlation between it and human evaluation. After the two new developments have been tested separately, integrated testing of both versions of DECOM using the adequacy metric will provide valuable data and insight into the effectiveness of the enhancements made to the model, as well as the added perspective the adequacy metric provides.

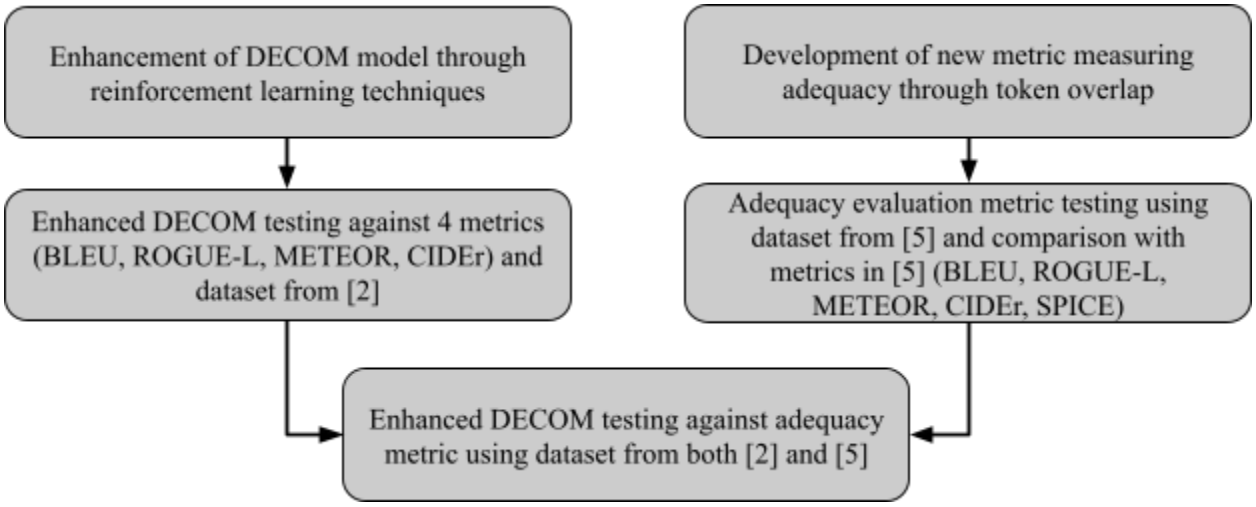

Figure 2: Outline of recommended research process, consisting of parallel development and testing phases and a final integration testing phase.

CONCLUSION

Because of the promise shown by the DECOM model, it serves as the most advantageous model upon which enhancements can be made, extending the effectiveness of a novel strategy shown to have merit. Being a novel strategy, standing in contrast to the documentation generation as translation paradigm, the first iteration was able to outperform other techniques, demonstrated high quality generation during qualitative spot checks, and has key improvement areas available, underscoring its viability as a candidate for further research.

To complement the enhancements of DECOM, evaluation metrics tailored to the documentation generation task will be key in increasing the granularity of feedback provided by these metrics over a large dataset, as the adapted computational linguistic metrics, while having provided useful insight, have been identified as lacking incorporation of key documentation aspects, reducing the informativeness of their produced results. To allow for efficiency in the workflow and analysis of future metrics, the adequacy metric must be developed and tested against the enhancements to the DECOM model.

By conducting the above proposed research, our Applied Artificial Intelligence research team and Northeastern can greatly expand the field. Creating viable and reliable models with tangible benefits to a colossal industry such as software development will not only justify the continuation of Applied Artificial Intelligence research but also increase our team's credentials. Through the review of literature, it is clear that the DECOM model possesses both the best potential improvements from enhancements and the best platform for enhancements.

REFERENCES

[1] C. D. Newman *et al*., "Automatically generating natural language documentation for methods," *2018 IEEE Third Int. Workshop on Dynamic Software Documentation* (DySDoc3), Sep. 2018. doi:10.1109/dysdoc3.2018.00007

[2] F. Mu, X. Chen, L. Shi, S. Wang, and Q. Wang, "Automatic comment generation via multi-pass deliberation," *Proc. of the 37th IEEE/ACM Int. Conf. on Automated Software Eng.*, 2022. doi:10.1145/3551349.3556917

[3] J. Y. Khan and G. Uddin, "Automatic code documentation generation using GPT-3," *Proc. of the 37th IEEE/ACM Int. Conf. on Automated Software Eng.*, 2022. doi:10.1145/3551349.3559548

[4] Y. Su, C. Wan, U. Sethi, S. Lu, M. Musuvathi, and S. K. Nath, "HOTGPT: How to make software documentation more useful with a large language model?," *Proc. of the 19th Workshop on Hot Topics in Operating Syst.*, 2023. doi:10.1145/3593856.3595910

[5] X. Hu *et al*., "Correlating automated and human evaluation of code documentation generation quality," *ACM Trans. on Software Eng. and Methodology*, vol. 31, no. 4, pp. 1–28, 2022. doi:10.1145/3502853