Second Milestone Report: Non-parametric Language Models for Natural Language to Code Generation

Alex Xie
https://axie66.github.io/07400-project/

Mentored by Vincent Hellendoorn
Institute for Software Research

February 2, 2022

1 Major Changes

The focus of the project has shifted from incorporating unannotated data into the nearest neighbor datastore to improving non-parametric models for code in general. This is because initial results have shown that the vanilla kNN-MT formulation [2] is less effective than expected on the CoNaLa dataset. I hypothesize that this is because the base parametric model yields subpar context representations, limiting the accuracy of the retrieval mechanism. Given this issue, I do not believe that naively adding unannotated data will necessarily improve performance; instead, additional architectural adjustments will likely be necessary.

2 Progress Report

2.1 Accomplishments

	w/o kNN	w/ kNN	Δ
BERT-TAE	33.41	33.50	+0.09
CodeT5	36.39	36.74	+0.35

Table 1: BLEU scores on CoNaLa test set.

To improve datastore representations, I re-ran the CoNaLa experiments using CodeT5 [4] instead of BERT-TAE [3] as the base parametric model, the results of which are shown in Table 1. CodeT5 without kNN already yields state of the art performance on CoNaLa, achieving around 35 BLEU. Adding in kNN yields a greater increase in performance compared to BERT-TAE; however, the size of the increase is still perhaps less than desired and does not fully demonstrate the benefit of the kNN approach.

2.2 Previous Milestone Goals

While I have changed the direction of the project, I have met the goals I set for this milestone as they mainly concerned set-up work for the project, such as implementing the base kNN model and obtaining results on CoNaLa.

2.3 Surprises

The only surprise, as discussed earlier, is the (under)performance of the standard kNN-MT architecture.

3 Next Steps

3.1 Looking Ahead

By the next milestone, I plan to obtain results on the Concode dataset [1], a larger dataset that may better highlight the strengths of kNN-MT; code for this is already written, and all that remains is running the experiments.

In addition, I plan to try out a couple modified kNN approaches (this might take longer than a single milestone). Specifically, the approaches of interest are an adaptive kNN-MT variant that dynamically adjusts the number of neighbors k [6] and a semi-parametric model that attends to retrieved neighbors to generate a kNN context vector [5].

3.2 Revisions to Future Milestones

For the foreseeable future, my milestones will likely revolve around various improvements to the kNN architecture, including but not limited to the ones described in the previous section. Another direction may be incorporating some form of representation learning to improve context representations for kNN retrieval. Finally, if time permits, I may revisit my previous project goals of incorporating unannotated data into the datastore.

4 Resources Needed

At this point, I have all the resources necessary for my 07-400 project.

References

[1] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. Mapping language to code in programmatic context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1652, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Sajad Norouzi, Keyi Tang, and Yanshuai Cao. Code generation from natural language with less prior knowledge and more monolingual data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 776–785, Online, August 2021. Association for Computational Linguistics.

[4] Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of*

the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, 2021.

[5] Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. Adaptive Semiparametric Language Models. *Transactions of the Association for Computational Linguistics*, 9:362–373, 04 2021.

[6] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 368–374, Online, August 2021. Association for Computational Linguistics.