Master Thesis Topic

Code Generation from Natural Language Documentation

Motivation and Background

Code generation has been one of the most popular applications of natural language processing (NLP) over the last few years. Previous studies [1, 2, 3, 4] viewed code generation as a probabilistic problem of maximizing the conditional probability of generating code *y* based on a given document *c*, where neural network is usually leveraged to support an encoder-decoder approach. Even though many of the state-of-art approaches have achieved good performance with respect to BLEU (bilingual evaluation understudy) [5], the accuracy of the generated code is still low and the functional correctness remains a main issue.

Goals

The student should develop a prototype to generate general-purpose code (i.e. in programming languages) from a given documentation. This could be achieved by optimizing previous techniques such as Bi-RNN and AST modeling. The goal is to generate functional correct code with higher accuracy, when given a functional description of the target code.

Description of the Task

A detailed description of the task and the underlying techniques will be given personally on interest.

Research Type

Theoretical Aspects:	*****
Industrial Relevance:	*****
Implementation	*****

Prerequisite

The student should be enrolled in the master of computer science program, and has completed the required course modules to start a master thesis.

Skills required

Programming skills in (preferably) Java and Python, Understanding of, or willingness to learn, the architectural and statistical foundations needed for the project.

References

[1] Yin, Pengcheng and Graham Neubig. “A Syntactic Neural Model for General-Purpose Code Generation.” *ACL* (2017).

[2] Ling, Wang et al. “Latent Predictor Networks for Code Generation.” *CoRR*abs/1603.06744 (2016): n. pag.

[3] Li Dong, Mirella Lapata. “Language to Logical Form with Neural Attention.” ACL (1) 2016.

[4] Miltiadis Allamanis, Daniel Tarlow, Andrew D. Gordon, and Yi Wei. 2015. Bimodal modelling of source code and natural language. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15), Francis Bach and David Blei (Eds.), Vol. 37. JMLR.org 2123-2132.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318. DOI: https://doi.org/10.3115/1073083.1073135

Contacts

Minxing Tang, Humboldt-Universität zu Berlin, Institut für Informatik, Lehrstuhl Software Engineering, Unter den Linden 6, 10099 Berlin, Germany

Application

Please contact during office hours or write an email with the title: “CG-NLD” to se-career@informatik.hu-berlin.de