ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Machine Learning Models for Code Search

The goal of the "Code Search" endeavour is to be able to retrieve code fragments from a large code corpus that most closely match developer's intent, which is expressed in natural language.

More precisely, we wish to develop a system where the user, having provided a natural language query, is presented with a list of code fragments that correspond to that query. A "Google for Code" of sorts, if you wish.

The key difficulty lies in retrieving those queries that are related to the input query *semantically* (they have the same meaning), and not just syntactically (i.e. "having the same structure") or lexically (i.e. "using the same words").

We strongly suspect that most approaches to the problem, though functional by their own tailored metrics, come short of understanding the semantics of queries and code.

We therefore wish to do the following three things:

1. Tweak some of the existing benchmarks for code search to better test whether models possess any sort of semantic understanding, say by simple obfuscation of test samples, and evaluate existing models w.r.t. these modified benchmarks.
2. Create a benchmark or an increasingly demanding sequence of benchmarks that properly reflect whether a machine learning model understands this or that semantic concept often appearing in code.
3. Try to improve on the existing models by using representations of input data that better capture there-present semantic relationships of interest.

Candidate Profile. Varies from project to project. *Bachelor's* students will work on a smaller project, focusing mostly on item 1. *Master's* students will work on an larger project, coding and advancing our work on items 1, 2, and perhaps 3.

Generally speaking, a good candidate is a motivated, competent programmer in the language of his/her choice, and is interested in one or more of the following fields: statistical machine learning, deep neural networks, natural language processing, algorithm learning.

Interested? Please contact us to learn more!

Contact (please send an email with the following as recipients)

- Peter Belcak: belcak@ethz.ch, ETZ G61.3
- Florian Grötschla: fgroetschla@ethz.ch, ETZ G93