10/28/2019 - 11/03/2019

E E/Cpr E/S E 491 Weekly Report 6

Intelligent Code Editor

Client & Advisor: Ali Jannesari

sdmay20-46

John Jago – Software Lead

Keaton Johnson – Systems Lead

Jon Novak – Machine Learning Lead

Matthew Orth – Meeting Facilitator

Garet Phelps – Report Manager

Isaac Spanier – Test Lead

Weekly summary

This week was focused on adding more language to the dataset, using context to identify variables on the plugin side, and figuring out how to improve the accuracy of the translations.

Past week accomplishments

John Jago

- Context aware translation in plugin
    - Names of initialized variables are saved off
    - The variables are replaced with myVar1, mayVar2, etc. before the string is sent to the translation server
    - After getting the translation back, the original variable names are put back
    - This allows the language model to correctly return statements like the following:
        - int x = 30;
        - print my age is x
        - Result: System.out.println(“my age is “ + x);
            - Previously it would have returned: System.out.println(“my age is x”);

Keaton Johnson

- Updated design document with Isaac
- Looked into creating a powershell script to query GitHub API
    - Returned raw files containing requested search text
    - Looked at the highest starred projects written in java on GitHub

    - Looked into ways around the API limit

Jon Novak

- Continuing to work on making the dataset accurate

Matthew Orth

- Automatic Dataset Creation Research:
    - Researched methods for automatically generating our natural language to code dataset
        - SNIFF (https://link.springer.com/content/pdf/10.1007%2F978-3-642-00593-0_26.pdf)
        - Conala (https://arxiv.org/pdf/1805.08949.pdf)
- Pronto GPU Server Configuration:
    - Completed configuration of the shared Pronto GPU server to run the OpenNMT-py preprocessing, training, and translation models on
    - Ran multiple training models through for the Conala dataset to test speed and functionality including for the Google GNMT model
- Java Print Dataset Training:
    - Ran our manually generated Java Print Dataset through OpenNMT-py
    - Recorded the BLEU score and accuracy of the results

Garet Phelps

- Added a dataset containing the 10,000 most used words in Engligh
- Looked into wordnet
    - Not necessarily useful as a database of words, but could be useful when looking at synonyms of words and different meanings.

Isaac Spanier

- Updating the Design Document

Individual contributions

Name	**Contributions**	**Hours this week**	**Hours cumulative**
John Jago	• Context aware translation in plugin • Design documentation update	5	24

Keaton Johnson	• Looked into creating a powershell script to query GitHub API • Updated design doc with isaac	4	17
Jon Novak	• Working on making the dataset more accurate	5	19
Matthew Orth	• Automatic Dataset Creation Research • Pronto GPU Server Configuration • Java Print Dataset Training	7	68
Garet Phelps	• Added 10,000 words to the dataset •	2	16
Isaac Spanier	• Updating the Design	1	12

Plans for the upcoming week

John Jago

- Add myVar examples to the print dataset
- Adjust plugin logic as necessary when testing with trained model

Keaton Johnson

- Adjust the arithmetic examples that you generated in the dataset to be formatted as: print 10 + 35, System.out.println (10 + 35); instead of System.out.println (350); (do not evaluate the arithmetic in the code translation)

Jon Novak

- Look into methods for labeling the print statements that we mine from GitHub

Matthew Orth

- Integrate generic variables examples into the dataset (myVar, myMethod, myClass) / run the updated dataset through OpenNMT.py

Garet Phelps:

- Add variable examples to the dataset
    - myVar
    - myLiteral?
    - “I want ” + myVar

Isaac Spanier

- Research more about the usages of NLTK for being able to match more generalized inputs (https://www.nltk.org/)

Summary of weekly client/advisor meeting

We met with Hung on 11/1/19 at 4pm. Four members were present. We discussed using placeholders for variables to easy translation on the side of tht NMT. We will add that logic to the front-end and more cases to the database this week. We also discussed using variables and strings in the same print statement, and how to get that to work fine.