01/28/2020 - 02/10/2020

E E/Cpr E/S E 492 Weekly Report 2

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

During this sprint, our team focused on creating the Verb-Noun preprocessing script, custom part of speech tagger, and starting on the Java method invocation dataset mining. The Verb-Noun preprocessing script and custom part of speech tagger were completed successfully and tested on several programming-domain sentences. The framework behind Java method invocation dataset mining was determined using the Octokit C# framework. We will focus on integrating the Verb-Noun script and part of speech tagger into the rest of the system and completing the mining process for the Java method invocation dataset during the next sprint.

Past week accomplishments

John Jago

- Created a set of programming domain verbs and nouns
    - Defined boundaries on what we consider to be the “verb”
    - Confirmed that it doesn’t really matter whether the noun/verb is labeled correctly as noun/verb. It just has to be one of the two so it doesn’t get filtered out.
    - However, Garet and I will label the programming domain nouns and verbs correctly in case it’s useful in the future
- Custom POS tagger
    - Worked with Garet to extend the default NLTK POS tagger in Python
    - Merge request open in GitLab
- Researched potential ways to speed up feedback cycle
    - For example, it took us three days last semester to train a model for print statements

    - The following paper contains concrete steps for examining loss functions in order to adjust hyper-parameters more effectively
    - Once we have our dataset for method invocations, we can look see if the methods in this paper allow us to speed up training times
    - https://arxiv.org/abs/1803.09820

Keaton Johnson

- Continued work on data mining program
    - Authenticated program using a OAuth token using github's authentication
    - Started utilizing octokit to search repos
    - Automatically obtaining java files from each repo.

Jon Novak

-

Matthew Orth

- OpenNMT-py <unk> token issue research
    - Confirmed that OpenNMT-py is not installed for all users, so the <unk> token issue is not related to configuration
        - Tried a fresh installation of OpenNMT-py on another user account and still received the same issue
    - Determined to wait until we create a new dataset before uninstalling and reinstalling to ensure it is not related to a dataset issue
- NLTK Preprocessing Implementation:
    - Documented potentially useful NLTK functionality with experimentation results
    - Created NLTK preprocessing implementation that includes only the Verbs, Nouns, Digits, and Adjectives while splitting up method names into multiple words and converting all verbs to present tense
- Verb-Noun Implementation Testing:
    - Tested the Verb-Noun preprocessing on around 100 natural language statements pulled from AnyCode
        - Best results were achieved when all types of verbs, nouns, digits, and adjectives were matched
    - Tested implementation on Java Documentation descriptions
        - Determined that some processing will need to be done when using Java Documentation descriptions as natural language labels in the dataset
            - Documented a method for this that involves manually converting the sentences

Garet Phelps

- Helped create the modified POS tagger to handle technology words
    - Researched how to do it
    - Decided on extending the default POS tagger
    - Made a tagset of a bunch of programming tags.

Isaac Spanier

- Created a overview for the system diagram to visualise the breakdown of our system in a step by step manner
    - Made Visualization
    - Broke down each block into potential issues and places for improvement
- Revisited the NLTK script and looked at points for improvement

Individual contributions

Name	**Contributions**	**Hours this week**	**Hours cumulative**
John Jago	● Created a set of programming domain verbs and nouns ● Custom POS tagger ● Researched potential ways to speed up feedback cycle	3	6
Keaton Johnson	● Continued working on data mining program	2	4
Jon Novak	● Continuing to work with Keaton on mining methods	2	4
Matthew Orth	● OpenNMT-py <unk> token issue research ● NLTK Preprocessing Implementation ● Verb-Noun Implementation Testing	10	25
Garet Phelps	● Created a set of programming-related nouns and verbs that normal pos taggers don't do correctly	5	7
Isaac Spanier	● NTLK Script Review ● System Diagram	4	6

Plans for the upcoming week

John Jago

- Create user interface parameter type mapping and integrate Verb-Noun preprocessing script into IDE

Keaton Johnson

- Continue mining the 5,000 Java method invocation dataset from GitHub

Jon Novak

- Continue mining the 5,000 Java method invocation dataset from GitHub

Matthew Orth

- Create detailed block diagram of project with challenges of each step, list of tools we are using, cite all research used, and list possible next steps for the project

Garet Phelps:

- Create user interface parameter type mapping and integrate Verb-Noun preprocessing script into IDE

Isaac Spanier

- Create detailed block diagram of project with challenges of each step, list of tools we are using, cite all research used, and list possible next steps for the project

Summary of weekly client/advisor meeting

Meeting with Hung Phan on 2019-02-03, 4:30 pm

We discussed the progress on our three major tasks for this sprint: coming up with a standard format for preprocessing the natural language statements, creating a test data set, and creating a custom part-of-speech tagger for use in preprocessing. After working on these tasks for a week, we cleared up any outstanding questions that each person had. We also sketched a rough diagram representing the overall system design (Figure 1).

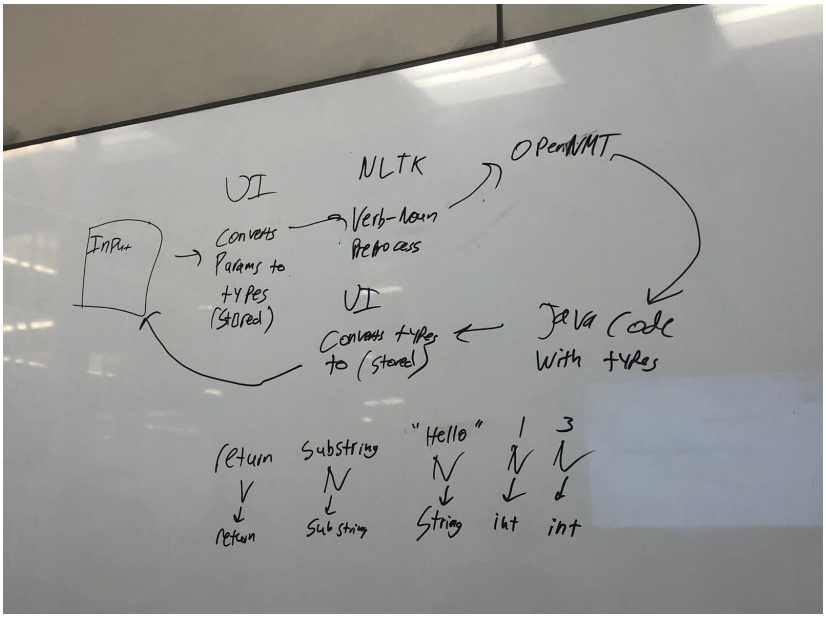

Figure 1: A sketch of the system design

Meeting with Professor Jannesari on 2019-02-10, 4:30 pm

During this meeting, we discussed the progress we made on the Verb-Noun preprocessing script, part of speech tagger, and dataset mining. We also discussed our proposed dataset labeling method and high level project design overview and future additions that could be made to the project. To conclude the meeting, we determined that our focus for the next sprint will be on project documentation, user interface integration, and Java method invocation dataset mining and labeling method.