E E/Cpr E/S E 491 Weekly Report 8

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 researching ways to make our translations more accurate. The main categories of research were labeling mined data, input generalization, using NLP tools like the Natural Language Toolkit.

Past week accomplishments

John Jago

- Research data mining, labeling, and input generalization
  - Created a document highlighting pros and cons with regards to methods for obtaining Java data sets, methods for labeling those data sets, and methods for generalizing our natural language input

Keaton Johnson

- Researched alternatives for a synonym API. (See Keaton Johnson Research in Sprint 6)
  - Found a public api with very little limitations, also returns confidence scores.
  - This could be used as a tool for assisting with manual inputs
  - Other apis had either limited uses or were costly.

Jon Novak

- Looking into github method labeling

Matthew Orth

- NLTK Research and Basic Implementation:
  - Researched functionality of NLTK and how it could be useful to our project

- - - ■ Determined that these methods could be useful for preprocessing the input natural language statement before passing it to OpenNMT-py
    - ○ Uncovered the limitations of WordNet synsets (synonyms) to words in the programming domain
    - ○ Determined how to use chunking and tags to segment a natural language input statement into the action and parameters
  - Word Embeddings Research:
    - ○ Researched word embeddings as a potential tool to support generalized natural language input
    - ○ Discovered GloVe as a potential tool to implement the word embeddings
  - Dataset Mining, Labeling, and Generalization Research:
    - ○ Dataset Mining: Created pros and cons list for GitHub, StackOverflow, and Java API documentation mining
    - ○ Dataset Labeling: Created pros and cons list for manual, automatic, and OpenNMT-py dataset labeling
    - ○ Input Generalization: Created pros and cons list for manual, NLTK, and word embedding input generalization

Garet Phelps

- Did some research on NLTK, and how we could use it for our project.

Isaac Spanier

- Continued Testing the UI Element for errors.
  - ○ Ran the the project and found some potential run errors for new users.

Individual contributions

Name	Contributions	Hours this week	Hours cumulative
John Jago	● Research data mining, labeling, and input generalization ● Prepared script for demoing the plugin for the video assignment	3	33
Keaton Johnson	● Researched alternatives for a synonym API.	3	22
Jon Novak	● Looking into method labeling	5	24

Matthew Orth	● NLTK Research and Basic Implementation ● Word Embeddings Research ● Dataset Mining, Labeling, and Generalization Research	8	82
Garet Phelps	● Did some research on NLTK, and how we could use it for our project.	1	19
Isaac Spanier	● Did some testing on the IntelliJ component	4	18

Plans for the upcoming week

John Jago

- Review design document and prepare slides for semester presentation
- Make sure the plugin works with n_best > 1 now that support for this feature (https://github.com/OpenNMT/OpenNMT-py/pull/1631) has been merged

Keaton Johnson

- Continue researching APIs
- Look into dataset mining tools

Jon Novak

- Continue looking into github method labeling

Matthew Orth

- Finalize optimization for the Java print dataset
- Design Documentation and Final Presentation Updating and Preparation

Garet Phelps:

- Review the design documentation
- Look into NLTK a bit more

Isaac Spanier

- Document potential errors for the UI component, and review the design documentation.

Summary of weekly client/advisor meeting

No meeting was held this week.