Diploma in Computer Science Project Progress Report

Exploration of Spoken Programming Languages

January 29, 2019

Name: James Hargreaves **College:** Gonville and Caius **crsid:** jh2045

Project Supervisor: Paula Buttery

Director of Studies: Graham Titmus

Overseers: Robert Mullins and Pietro Lio

1 Work Completed:

1. Collect Corpus - I wrote a web app which present the respondent with a prompt of a simple algorithm and recorded their response using the microphone on the respondents computer. I got 14 respondents giving me a data set of size 135. This data was then transcribed into a text file using the app at trint.com and then correcting the mistakes by hand. I then defined a pseudocode grammar and for each transcription I typed out the pseudocode which I felt best matched it.
2. Entity recognition - This first thing I did was build python scripts to convert the pseudocode by replacing variables and function names with a variable token e.g. VARABLE0 and FUNCTIONCALL0. and also remove the braketing from the pseudocode since this could be infered in a post processing step.
3. Baseline - I produced a list of mappings of transcript words to functional words using the list of mappings at https://blog.codinghorror.com/ascii-pronunciation-rules-for-programmers/ plus adding some example myself where i thought the corpus was lacking e.g. '*' to 'multiplied by'. I used this to give a list of functional words / phrases a user could use by adding in the key words of the language. And used the transcripts to give a stream of pseudocode tokens using longest prefix match continuously. I trained and used an n-gram model to compute the most likely ordering of the given tokens.
4. Statistical Machine Translation - built word alignment models using IBM models 1 and 2 and use these to build phrase alignement tables. This combined with a n-gram mdoel was used to construct a beam search decoder. I also experimented with and without pruning.
5. Evaluation - Have written a python script which takes a translation attempt of the pseudocode and the actual pseudocode and computes the minimum edit distance.

2 Still To Do:

I still need to construct the translation model using traditional machine translation techniques - using manually written rules.

3 Timetable:

The work I have done covers everything I plan to have done by the 3rd of February other than the tradition machine translation technique. However I already planned in a 2 week overflow time from the 4th of February and so I am not concerned about being behind.

[1]

[1] Word count = 391