Jordan Williams and Jason Deutsch

Project Report

We learned quite a lot from this project. For starters, we never thought about using natural language processing techniques as a tool/application to grade papers. We learned how parse trees can help classify if the sentence is a complete one or not. We also learned how useful the POS taggers can be in order to classify incorrect verb usage and/or pronoun usage. We also never knew how complicated it can be to develop an NLP system. There are all sorts of ambiguities to deal with, rare exceptions, and issues in trying to capture at least some of the meanings behind a body of text.

There are many things that worked and were a bit more straightforward. The grammar checking was fairly accurate, but it didn’t include some forms of punctuation so we had to add those in. For the most part, getting the sentence length worked fairly well, since we also included how much variety the writer used by exploiting the POS tagger. Looking at agreement between verb usage with number as well as tense, worked fairly well in capturing a lot of the basic verb usage errors, but for some more complicated structures it can be difficult to just use a POS tagger to look for errors.

Certain things were complicated and didn’t work so well. As mentioned earlier, using just the POS tagger to look for errors on more complicated structures can be difficult. The NLTK tagger should probably include a more diverse set of tags in order to improve the calculation of verb usage errors. Of course one issue the NLTK tagger has is when a sentence’s verb usage is incorrect the tagger doesn’t perform as accurate as it should, since the tagger is trained on formal English more or less. Similar issues can occur with the Stanford parser. The sentence “It is unless.” for example returns a parse tree in which there is nothing wrong with the sentence.

We also tried to use Wordnet’s similarity score between various nouns in the essays such as car and pollution for example and to iterate this type of score similarity throughout the whole essay (for each noun). The hope was that we could build a metric based on the similarities to address how much the essay remained on topic but unfortunately this approach doesn’t work very well. Not only does similarity score not return very accurate values, but it is also very hard to capture different styles of the way each writer writes using this approach.

By improving the accuracy of the Stanford Parser and NLTK tagger, certainly we can achieve a higher level of performance on our software. We could also increase the performance by adding a spell corrector function which would replace the misspelled words. This would (potentially) prevent double penalizing in the rest of the grading. We could also check for errors using the POS tagger, but instead of just limiting it to verbs and nouns we could check for incorrect usages (i.e. a transition from PRP to TO). We could also add another category that checks for proper use of punctuation i.e. some essays didn’t have any periods, didn’t use periods etc.