Module: Intermediate Methods and Programming in Digital Linguistics
Name: George Y.
Semester: FS 2021
Proposal Draft

Intermediate Methods and Programming in Digital Linguistics – Main Project Proposal Draft:

Main project proposal 2021:

The overall aim of the project is:

After browsing for relevant NLP tasks:
http://nlpprogress.com/english/simplification.html -->

*"Simplification consists of modifying content and structure to make it **easier to read and understand**, but preserving its **main core meaning**. A simplified version of a text could benefit: low literacy readers, English learners and etc. "*[1]

The main goal of this (supposedly and hopefully simple) project mainly involves simplification of texts based on Wikipedia against their simple Wikipedia counterpart.
Evaluation through simple N-gram metrics calculations will be used on the converted "simplified" sentences along with toolkits (Example. **E**asier **A**utomatic **S**entence **S**implification **E**valuation) and other packages available for Python to obtain metrics to compare and contrasts output values. The measure of simplicity and readability or ease to understand will be evaluated through different metrics but will also be judged by the subjects such as an English scientific/military corpus section (with high presence of technical terms and complex named entities) against standard English novel/newspaper section (less complex or more accessible terminology).

Expected input of program:

Text version chunks selected from Wikipedia and Simple Wikipedia articles varying through complexity of topic/subject. Furthermore, another critical metric would be 1-N-Gram evaluation to see the calculation in different word length per simplified sentence. The process will involve breaking down the chunk of text into sentence components. Then simplified with comparison of the N-grams → use packages to obtain scores of interests.

Example taken from: https://github.com/feralvam/easse with (**E**asier **A**utomatic **S**entence **S**implification **E**valuation):

```
from easse.sari import corpus_sari

Original Sentence: ["Regular more complex version of sentence"],
            System version of sentence:["About 95 you now get in.", "Cat on
mat."],
            Reference sentence:[["About 95 species are currently known.",
"The cat sat on the mat."],
                        ["About 95 species are now accepted.", "The cat is on
the mat."],
                        ["95 species are now accepted.", "The cat sat."]])
```

Expected output:

Scores and to a certain degree the alignment accuracy along with some measure of 'simplicity' and difficulty to understand.

```
Output score number [1]: 33.17472563619544
Output score number [2] – which will be another sentence and compared to
previous output score!
```

SARI Score will also be used:

SARI (Xu et al., 2016) is a *lexical simplicity* metric that measures "how good" are the words added, deleted and kept by a simplification model. The metric compares the model's output to *multiple simplification references* and the original sentence. SARI has shown high correlation with human judgements of simplicity gain (Xu et al., 2016). Currently, this is the main metric used for evaluating sentence simplification models.[2]

[1] http://nlpprogress.com/english/simplification.html
[2] http://nlpprogress.com/english/simplification.html