NLP Project Proposal - Sentence Splitter

https://cs.nyu.edu/courses/spring22/CSCI-UA.0480-057/homework6.html

Team Members

Hoaran Xu - Responsible for software development and analyzing theoretical issues
Sreeya Narra - Responsible for drafting the final paper and aiding in software development
Zachary Waxman - Responsible for software development and evaluation of the project from time to time.

Problem Statement

Many tokenizers do not take into consideration the concept of sentence boundary, which makes it difficult to do certain algorithms that require defining relationships between sentences. Similarly, many POS taggers and semantic role label programs require sentence separation. It is also shown that in machine translations, incorrect sentence boundaries have a significant impact on the quality of the translation. In addition, the problem of sentence boundary is a relatively less researched topic and hopefully our project can provide some more insight into this area.

Our program is designed to be used as pre-processing for algorithms. It will split a given English text into a list of sentences after appropriate training. We want our algorithm to be as general as possible, being able to achieve good accuracy and f-score across multiple domains.

Evaluation Plan

Our system will take in a Raw English text and output a list of split sentences. The result will be compared with a baseline system and another widely used system. Since sentence boundaries can be reduced to a tagging problem (Talked about in Strategy for Solving the problem), we can use accuracy, precision/recall and f-score to determine the performance of our splitter. All the dataset used will be split into a training and testing dataset, where the training data set is used for development purpose and testing dataset used for cross validation.

The baseline system used is probably some simple ruleset using manual rules. We might add another widely used sentence boundary labeler for comparison.

Academic Articles related to Project Scope

https://aclanthology.org/C12-2096/
This paper offers some insight on the most prominent sentence boundary labelers and how they compare to each other. It serves as a good source to determine competing algorithms and scoring systems.

https://aclanthology.org/W19-2204/
This paper describes the problem of sentence boundaries in legal text which our system should hopefully overcome with training. It serves as a provider of dataset which we can use and compare with their system.

https://aclanthology.org/W15-5938/
Serves as a reference paper to see how other systems do on social media texts and common mistakes in other models.

https://aclanthology.org/P05-1056/
Reference paper for implementing Conditional Random Field if we get the chance to.

https://aclanthology.org/J06-4003/
Reference paper for the implementation and performance of an unsupervised machine learning sentence boundary labeler. Use this to compare our algorithm.

Strategy for Solving the problem

We are preparing to use a supervised machine learning algorithm called Ada Boosting to determine the sentence boundaries. The algorithm will iterate over possible sentence boundaries (this will be determined through manual rules) and the decision tree will determine whether or not this is a sentence boundary.

We make the assumption that a sentence starts from the immediate end of another sentence. The definition of a sentence might sometimes be controversial, we will refer to our training dataset for the definition instead of making one ourselves.

The only issue for ada boosting is the features to be selected for determining. We are going to estimate POS tags for context information and include other relevant information concluded through human observation. For certain words such as OOV words, we can also use certain manual rules to determine further information.

If time permits, we also want to try and incorporate other ML algorithms, specifically CRF.

Collaboration Plan

Haoran Xu: I’m going to be mainly responsible for implementing ada boosting and integrating input to feature of the algorithm

Zachary Waxman: I’m going to be responsible for implementing ada boosting and testing our program to measure progress.

Sreeya Narra: I’m going to be responsible for drafting the research paper and handling the output of our algorithm.