Project Title: SHSAT Scrambled Paragraphs - Sunny Side Up

Rebecca Poch and Zhengqi Xi

ECE-467 NLP 2016

Introduction - description and motivation. What are you going to do, and why is it interesting and/or important?

The SHSAT is taken by NYC middle schoolers in hopes of getting into the top high schools in NYC, including Stuyvesant High School, Bronx Science, and Brooklyn Tech. There is one section that gives the students a scrambled paragraph of 5 sentences that must put into the correct order. We aim to create a system that will determine algorithmically the right order of sentences.

http://www.kaptest.com/blog/admission-possible/2014/05/10/raiders-of-the-scrambled-paragraph/

http://scrambledparas.blogspot.com/

http://www.test-preparation.ca/scrambled-paragraphs-practice-questions/

Background research

There are multi-document summarization sentence ordering techniques described in the following articles. These articles use sentence cohesion to improve their results.

http://nlp.hivefire.com/articles/share/160/ - Sentence ordering in multi document summarization

http://www.icmlc.org/icmlc2012/006_icmlc2012.pdf - An Improved Approach to Sentence Ordering For Multi-document Summarization

This article discusses pronoun resolution techniques (Ch 21 in the textbook) with a system that achieved 86% success.

http://www.aclweb.org/anthology/J94-4002 - An Algorithm for Pronominal Anaphora Resolution

Implementation

Possible approaches

- Machine learning on set of 1st sentences, 2nd sentences, etc
    - Features
        - POS (using nltk) - number of occurrences and/or order
            - http://www.nltk.org/book/ch05.html
        - Parsing with CFG (nltk, Ch 13) - examine likely sentence structures
        - Parsing with PCFGs (nltk, Ch 14)
            - http://www.nltk.org/howto/grammar.html
- Pronoun resolution on all possible sentence permutations (5! = 120 permutations) (Ch 21.3) - pick permutation with the best overall pronoun resolution
    - An Algorithm for Pronominal Anaphora Resolution
- Cohesion and cue words (e.g. however, then, after that, etc) (Ch 21.2)
- Sentence ordering algorithms from document/ multi-document summarization techniques
    - An Improved Approach to Sentence Ordering For Multi-Document Summarization
    - Sentence ordering in multi document summarization
- Combination of above methods (e.g. ML to pick top 10 orderings, then pick set with the best pronoun resolution.)

Input (from internet sources)

- Training set of documents with scrambled sentences and their correct ordering
- Test set of documents with scrambled sentences
- Make our own test set of 5 sentence paragraphs

Output

- Sentence order guesses alongside correct answers
- Overall statistics (micro (per # sentence e.g. every 1st sentence was correct) and macro (overall) averaging)

Evaluation

The evaluation of our project will be based on statistics - precision/accuracy/recall/F1. Since we do not expect a very high success rate, we will aim for results that are somewhat better than chance. We at least would like to be able to identify, for example, the first sentence of each of the scrambled paragraphs.

Demo product:

The demo product will show results of algorithm on test set and examples of successful and unsuccessful results. We would also like to show results of system applied to sets of 5 sentences that we come up with.

Submission

The final project submission will consist of:

- Source code
- Training set files
- Test set(s) files
- Results of
- Writeup describing the project itself, including how to run the program and all the methods we tried