Rhetorical Figure Detection in Political Texts

Zubair Rafiq, Ahmad Bilal Sohail, Eltun Ibrahimov
Faculty of Computer Science and Mathematics
University of Passau, Germany
{rafiq01,sohail01,ibrahi10}@ads.uni-passau.de

1 INTRODUCTION

Rhetorical figures can be described as a word or a phrase used in the non-literal sense so that the speech/writing becomes more persuasive, relatable and vivid. The literal meaning of the phrase or word is not taken into account. For example "He racked his brain so that he could come up for new ideas for his book". Generally, the literal meaning of rack is a device or instrument that in earlier ages people used to torture the servant or to break their limbs apart. But here it is used to make a great effort to think or to remember something. The figures of repetitions are a family of figures. They usually involve repetition of any linguistic element, ranging from sound, as in rhyme, to concept and ideas, as in tautology and pleonasm. A computer can easily detect the repetition of words on the other hand detecting only the ones provoking a rhetorical effect requires much effort and the reason is the presence of many irrelevant and accidental repetitions.

1.1 Motivation

We believe that this is a challenging project which is going to offer us great learning skills in terms of speech exploration and analysis, data pre-processing, data exploration, machine learning and visualization techniques. After successful completion of this project, we will be able to differentiate between a good and a persuasive speech.

1.2 Goals

- A machine learning pipeline will be designed which will detect the potentially rhetorical figures from the presidential speeches corpus.
- We will focus on the repetitive figures family. Epistrophe, Epanaphora, Anadiplosis, Epanalepsis and most probably Chiasmus as well.
  - Epistrophe: repetition of the same word or words at the end (or near the end) of successive phrases, clauses or sentences. Also known as epiphora or antistrophe. (Ali wants pizza, Ahmad wants pizza, in fact, everybody wants pizza)
  - Epanaphora: repetition of the same word or group of words are repeated at the beginning of two or more clauses or sentences. (I can act. I can sing. I can do whatever I want to do)
  - Anadiplosis: the presence of a word or a group of words at the end of a sentence as well as at the beginning of the following sentence. (He had an idea, an idea that changed his life)
  - Epanalepsis: repetition of similar words/phrases at the beginning and end of the same sentence. (A poor can feel the pain of a poor)
  - Chiasmus: grammar from the first phrase is inverted in the second phrase.(Profit makes money and money makes profit) (Dubremetz and Nivre, 2015)
- The main challenge here is to detect only truly provoking rhetorical figures as there will be many false positive cases of repetitions as well which makes it a needle in the haystack.

2 PLAN

(1) A presidential speeches corpus, having millions of words, is to be selected.(Gawryjolek *et al.*, 2009)
(2) At the very next step, the data will be cleaned and pre-processed.
(3) An algorithm will be created for extracting all the occurrences of repetitions.
(4) In order to find true Chiasmus we will have to assign weights to the strings for which we will stick to a predefined linear model to rank.

$$f(r) = \sum_{i=1}^{n} x_i * w_i \tag{1}$$

(5) Decision Tree classifier will be used to train on 80% of data which will then be tested of the rest of 20%.
(6) F1 scores look promising for the evaluation phase.
(7) We will be using pandas, nltk, scikit-learn and matplotlib libraries of python.

REFERENCES

[1] M. Dubremetz and J. Nivre, Rhetorical Figure Detection: the Case of Chiasmus, 2015, pp. 23–31.
[2] J. Gawryjolek, C. DiMarco and R. Harris, An Annotation Tool for Automatically Detecting Rhetorical Figures SYSTEM DEMONSTRATION, 2009.