Text - Summarization using Conceptual Statistical Extraction

Vivek Patani
School of Informatics and Computing
Indiana University
Bloomington, Indiana – 47405
Email: vpatani@umail.iu.edu

Jinsu Kim
School of Informatics and Computing
Indiana University
Bloomington, Indiana – 47405
Email: vpatani@umail.iu.edu

***Abstract*—There have been increasing number of researches conducted in the domain of text summarization using different types of contexts using different methods. In this paper, we attempt to summarize text with important facts eliminating redundant or irrelevant information using extractive method. In this paper we focus mostly on extracting entities and concepts which occur more commonly and are meaningful. Our approach is a statistical one, wherein we weigh each concept in terms of relations and how frequently they occur.**

I. Introduction

With increasing amount of textual information available, it becomes difficult to find and read important text. Thus, text summarization would be very useful in that it produces important and concise summary and help readers to save time without having to read entire text.

Summary can be generated by **extractive** and **abstractive** methods. Extractive methods create short summary with each words, POS from whole sentences without modification of words. Abstractive summarization is more complex way, which makes new sentences summarize by paraphrasing sentences of source documents.

The paper talks about the selective picking up of entities in the text. We, in this paper weigh **entities** heavily over other concepts. An entity may be a person or any proper noun or moreover any noun. We along with this also collect the verbs associated to them. However, a deep dive into the subject tells us that not all verbs represent an important concept related to the paragraph, hence the term selective.
We divide the approach into **5** sections:

- Sentencifying
- Anaphora Resolution
- Dependency Parsing
- Statistical Inference
- Entity Recognition

We also use a lot of other ideas, but this briefly describes a good overview of the subject we are presenting here. Our focus is on English for different languages hold different semantic syntax, such like languages like Japanese and Chinese have unambiguous sentence ending markers.

II. Approach and Methodology

A. Sentencifying

To begin with we approach the data with splitting of the set of text into simple sentences. We do this to make relations and dependencies much easier and flexible to detect. The idea here is to break the sentence down at periods with a few reservations such as places like *Dr., Mr., St., etc.* 47% of the articles written in the Wall street journal represent abbreveations[2]. This problem in NLP is also referred to as Sentence Boundary Disambiguation (SBD), which is the science of detection of the begining and the ending of a sentence.

B.

III. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.
[2] Texting Mine Online.