Erol ÖZKAN

n18245609

CMP 670 STATISTICAL NATURAL LANGUAGE PROCESSING

PROJECT PROPOSAL

TEXT SUMMARIZATION

Nowadays plenty of documents, articles are stored in the Internet in an unstructured way. The topics in these documents varies from news to articles and most of these documents are reasonably long. Furthermore, there are lots of documents from lots of sources. Users are facing difficulty to find appropriate information from this mast amount of data.

Text summarization task is the process of identifying the most important and meaningful information in a document or set of related documents and compressing them into a shorter version while preserving their meaning [1]. Text Summarization methods can be classified into extractive and abstractive summarization [2]. In extractive summarization; important sentences, paragraphs etc. are selected from the original document. In abstractive summarization, text is interpreted and a new, shorter text is generated by using linguistic methods [2].

Many text summarization methods have been proposed so far. A feature that was used in this methods was the term frequency - inverse document frequency (tf-idf) feature [3]. Other features included; position of sentence in paragraph, resemblance of sentence to the title, centrality of sentence, path of sentence, relative length of sentence, etc… [2].

In this project I will design and develop an extractive summarization system for English. I will consider features proposed in older works. This features will probably include tf-idf, position of sentence, etc… Furthermore, I will search for datasets and evaluation metrics. Datasets in [4] will be considered and best one will be selected. For evaluation metrics, I will use ROUGE metric [1]. I will use python programming language.

PLANNING

Week 4-7: Domain Research

Between week 4 and 7, I will try to get a deeper understanding of the task, datasets and evaluation metrics. In order to do this, I will read more related works. I will try to understand what researchers proposed, what their main contribution is and what techniques they used in their research. I will write my foundings in my final report.

Week 4-8: Design and Implementation

During week 4-8, I will discover the design alternatives. I will choose some related works and I will try to implement_them.

Week 8-11: Implementation and Synthesis

I will try to contribute by synthesizing something new. I will add something on top of other's research in order to get better results.

Week 11-12: Evaluation

I will evaluate my results on the dataset I found.

Week 12-14: Writing the Final Report (or paper)

I will write the final report (or paper).

REFERENCES

[1] Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E.D., Gutierrez J.B., Kochut K., Text Summarization Techniques: A Brief Survey, *International Journal of Advanced Computer Science and Applications(IJACSA),* 8(10), **2017.**

[2] Gupta, V., A Survey of Text Summarization Extractive Techniques, *Journal of Emerging Technologies in Web Intelligence,* Vol. 2, No. 3, **2010.**

[3] Savyanavar, A., Mehta, B., Marathe, V., Shewale, P., Shewale, M., Multi-Document Summarization Using TF-IDF Algorithm, *International Journal Of Engineering And Computer Science,* **2016.**

[4] Text Summarization Datasets, https://github.com/mathsyouth/awesome-text-summarization, (Mart 2019)