Gündüz Vehbi Demirci

21102970

Emir Gülümser

20403203

Turkish Text Summarization

a) Description of the Problem

In this project, we plan to do summarization method that generates meaningful summaries for the texts. We plan to rank the sentences according to their importance and display the results to the users. In this project, we only focus on Turkish texts.

b) Motivation / Importance

In past, retrieving information from a subject was hard due to lack of information or difficulty of finding relevant resources [1]. With the widespread usage of the internet, documents and resources are brought to online which become accessible for everyone. From the information retrieval perspective, getting relevant information from the huge amount of data become more important and popular nowadays. Today available text search engines return too much documents for a person to identify which one are relevant to his/her needs. Therefor there is need for technologies that help people on this purpose. Presenting document summaries will help people to find their desired documents easily.

b) Methodology

We consider using Java programming language. In order to do summarization, we start with words. First, we would clean the text using stop word list and stemming. We plan to use Zemberek [2] for stemming of the Turkish words. Later, we plan to evaluate the words for candidate sentences using feature calculation techniques such as TFxIDF calculation and discretization. We can use training data in order to be used in the extraction stage. Finally, in the extraction stage using training data and the feature values, candidate words will be extracted. We plan to combine these words which have ranks, we would construct the sentences. However, at the end we may have similar sentences, so we would eliminate these similar sentences as a final step and display the summary to the user. We would evaluate out sentences using ROUGE evaluation technique which is used to compare summarization outputs with human generated summaries [3].

d) Expected Results

Firstly we expect to see our summaries match with test data. But there is possibility to see some of important sentences are not included and vice versa. Actually the effectiveness of the developed code and algorithm will determine the results.

e) References

[1] Mücahid Kutlu, Celal Cığır, and Ilyas Cicekli, Generic Text Summarization for Turkish, The Computer Journal, Vol. 53, No. 8, (2010), pp:1315-1323.

[2] Zemberek website : http://code.google.com/p/zemberek/.

[3] Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of HLT-NAACL-2003, Edmenton, Canada (2003).