Linguistic Features Usage in Single-Document Extractive Summarization

A. Dlikman, M. Last

Extractive Summarization is outputting the most relevant sentences of a document(s) without doing any changes [1].

Extractive summarizing can be divided into several steps. Preprocessing is a first of them and it usually includes: sentence splitting, stop words removal, stemming etc. In processing step, sentence features are calculated and then weights are assigned to these features using machine learning or heuristic methods. Those features are numerical characteristics of each sentence (e.g. sentence location, length, number of keywords, title similarity) which indicate the importance of sentence. In the end of processing stage final score of each sentence is calculated and highest score sentences are included in the summary. [1]

The purpose of our study is to improve performance and summary quality of existing extractive summarization techniques by usage of additional linguistic features. We investigate contribution of features, which make use in additional linguistic information, such as multi-words expressions (MWE), named entities (NE) and part-of-speech (POS).

During our study and following experiments we make use of Java-based parallel implementation of MUSE summarizer, which was introduced by Litvak and Last [2]. For training and testing we have used two English texts corpus DUC-2002 and MultiLing-2013, containing human summarized documents.

In our study we present 24 original MUSE features (out of total 31) which are extended by adding multiword expressions. In our opinion, extension of “bag of words” assumption by adding multi-words to the vector, additional to single words can improve performance of features are made usage in TF, TF-IDF, Cosine similarity and other similar approaches.

In addition, we introduce 15 new linguistic features based on named entities. Those features take into account four different types of NEs: location (geographical names like countries, cities, rivers etc.), person (person names), date-time (date or time mentions including days of week,
months etc.) and quantitative NE (any numeric information: numbers, percentage, amounts of money).

The POS grammatical data can indicate to an extent the presence or absence of information content in texts [3]. We suggest 17 new POS based features which include POS ratio (calculate the ratio of specific POS in a sentence), POS filtering (original MUSE features are applied to particular POS and the rest of words are discarded) and POS patterns features (POS 2, 3 and 4-grams considered).

We evaluated all our experiments in terms of ROUGE recall metrics [4]. POS features can significantly improve extractive summaries quality. The feature set containing POS features outperformed the baseline algorithm with p-value of 0.009 and 0.04 in the DUC-2002 and MultiLing-2013 data respectively (Wilcoxon test). The feature set containing NE and MWE features together slightly outperform the baseline (about 1% in terms of ROUGE-1), but the result isn’t statistically significant.

References


