CORPUS BASED IDENTIFICATION OF TEXT SEGMENTS

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

29.05.2017

Text processing for NLP tasks is mostly word-based. The words are not clearly defined, yet intuitive. Tokenisation is a process when the written text is divided into units with meaning, for instance, words, sentences, this process is quite error-prone, which is why local adjustments are necessary. Is the intuitive ”Word” concept the best way for a computer to segment a text?

The aim of this work is:

- Develop an algorithm, which breaks down a sentence or a text in its ”best” segments
- Determine whether the non-symbolic approach is better than the word-based one
- Chances and risks of the non-symbolic approach

The procedure goes in several steps. First of all, the extraction of N-gramms with length from 1 to 10 from the Wikipedia Corpus takes place. The corpus includes unannotated raw texts and the first 10 000 texts of this corpus (with around 22M characters) are exctracted. Secondly, a frequency list with the N-gramms is created and later the N-gramms are assesed with a quality measure.

Quality measure = n *log(freq) (where ”n” is the N-gramm length and ”freq” - the absolute frequency of the N-gramm. Finally, the sentence is given and broken down into N-gramms with the highest quality measures).

There can be a certain problem with this process - depending on the size of the input, the running time increases exponentially, which is why a heuristic approach might come in handy. The calculation of the highest quality measure is no longer guaranteed, however, the segmentation might be possibly better in comparison to the symbolic approach.

Evaluation of the text segments can prove to be quite a difficult task due to the often disagreement about the granularity of segments. Depending on the application two types of errors may occur: relevant and irrelevant. For instance, in Information Retrieval the correctness of segment boundaries can be neglected, whereas by ”news boundary detection” not. The impact on the end is used as a measure. Word2Vec allows us to get N-gramm embeddings of the letters. The sentiment analysis on the sentence level leads to evaluation. Movie reviews are used as the data and compared with word embeddings. The sigmoid function is applied to the sum of the weighted input values to get the result. Common N-gramms that are bigger that 3 contain function words and 8 contain content words.

Findings and open questions:

- No results of the evaluation are ready yet
- Find another way to extract N-gramms
- Meaningfulness of result of the evaluation