Department of Computer Science
Databases and Information Systems

Universitätsstr. 1 D–40225 Düsseldorf

Heinrich Heine
University Düsseldorf

Paper Summary

N-gram-Based Text Categorization

Zeynep Boztoprak

The paper N-gram-Based Text Categorization written by William B. Cavnar and John M. Trenkle in 1994 is about an N-gram based approach to categorize documents into predefined categories. These categories will be represented by text profiles generated from N-grams where N-grams are defined as contiguous N-character slices of a longer string.

Zipf's Law implies that there is a set of words which dominates most of the other words in terms of frequency of use. Therefore we can assume that comparing two documents from the same category there N-gram frequency distributions should be similar.

Based on this idea the authors presented a system which can categorize electronic data coming from different sources. At the same time their system is able to tolerate some kinds of textual errors like spelling or grammatical errors.

First a training dataset with a set of pre-existing text categories is needed. Categories might be language or subject domains. For each of the categories N-gram frequency profiles will be generated and used as representations for the corresponding categories. This process is called Generate Profile. This is done by seperating sample text into tokens and generating all possible N-grams (N=1 to N=5). The N-grams with the most occurence will be used for their N-gram frequency profiles.

When a new document arrives its N-gram frequency profile will be computed too.

In the so-called Measure Profile Distance phase the new arriving documents N-gram frequency profile will be compared to all the category profiles using the out-of-place distance measure. The out-of-place measure is a simple rank-order statistic which determines how far out of place an N-gram in one profile is from its place in the other profile. The sum of all out-of-place values for all N-grams is the distance measure for the document from the category.
In the final classification step the new document is then classified as belonging to the category with which it has the smallest distance.

Testing their system in language classification on Usenet newsgroup articles where articles are written in different languages resulted in an accuracy of 99.8%. It also seemed to work well for classifying articles according to subjects with an accuracy of 80%.

Observations show that the top 300 N-grams are always highly correlated to the language and reflects the distribution of the letters of the alphabet in the document language or contains very frequent prefixes and suffixes. After that N-grams are more specific to the subject of the document. In larger documents these shift might occur later.

Thus, the main advantage of their system is that due to the decomposition of strings into small parts text errors can only affect a limited number of these parts while the rest remains intact. When using whole word statistics the system becomes more sensitive to this type of errors. Also their system is language independent and requires no further information about the underlying language. Results on the language classification task show that performance is less sensitive to the length of the document as the authors expected.

In our course we work with data in the form of texts that we want to interpret. This paper seems to be relevant for our course because it shows some concepts how to deal with this kind of data and it is also a good introduction to this terminology and basic understanding.