DEVELOPING A STEMMER FOR GERMAN BASED ON A COMPARATIVE ANALYSIS OF PUBLICLY AVAILABLE STEMMERS BY LEONIE WEISSWEILER

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

26.06.2017

One of the important tasks in Information Retrieval is not only return the documents matching the exact query of words, but also return semantically related words or different morphological forms of these words in the original query. Reducing words with the same root to a common form is called *stemming*. The objective of this work consists in three parts: present first comparative evaluation of existing stemmers in German, present new state-of-the-art stemmer based of evaluation results, make official implementation available in a range of programming languages. This is due to the fact that there is a small choice of stemmers for German language. There are some existing stemmers: Snowball, Text::German, Caumanns, UniNe (Light or Agressive).

Firstly, the comparative evaulation of the existing stemmers in German is done by stemming the CELEX2 corpus. Then the precision and recall are computed by matching the best found stem in the stemmed corpus with each of the gold standards. When comparing both stems and results, some are more accurate than the others. Some stemmers, like Snowball, showed only one root of multiple words, whereas some, such as Text::German represented 3 different roots. When looking at the results, the least accurate of all the available was UniNe Light, after which comes Snowball, then UniNe Agressive, following by Text::German and lastly Caumanns. After determining the best-performing stemmer, it was decided to come up with "CISTEM", a stemmer with higher accuracy and recall, based on the Caumanns stemmer.

The development of this stemmer is following: firstly, the transformation to lowercase takes place, then the replacement rules are applied (for instance, "ü" to "u", "ge" is deleted if the length of the words is more than 6 letters, "xx" to "x*"), later on, if the length of the word is more than 3 letters, some stripping rules are applied, else replacement rules are applied again, which are vice-versa of the ones in the beginning (for instance, "u" to "ü", "x*" to "xx"), in the very end of this chain the result is achieved. By following these simple rules, CISTEM has achieved the highest results, reaching an F-score of 93 percent for the first golden standard and 95 percent for the second one.

The important goals for future work are releasing official CISTEM implementations for a number of pgramming languages, merging two gold standards into one definitive gold standard, implementing a rule-based system for learning the optimal stemmer.

In conclusion, CISTEM has a number of advantages:

- Has achieved the highest F1 measure score
- Quickest stemmer (when implemented in PERL)
- Easy to understand and use
- A case insensitive version is available
- Availability of official implementations among a range of programming languages