Joseph Birkner
Mat.Nr. 11393024

**Protokolle zum Colloquium (26.06.2017)**

---

Studentin: Leonie Weißweiler, Betreuer: Alexander Phraser

*„Developing a Stemmer for German based on a Comparative Analysis of Publicly Available Stemmers“*

(Dieser Vortrag bezieht sich auf ein Paper, welches Leonie in der Gruppe von Prof. Phraser bei der GSTL eingereicht hat.)

Vor Allem im Bereich des Information Retrieval (Dokumentensuche) ist es of wichtig für einen bestimmten Suchstring nicht nur Dokumente als relevant zu betrachten, die die gesuchten Wörter buchstäblich enthalten, sondern auch solche, die einem Suchwort semantisch ähnliche oder aus dem Paradigma eines Suchwortes stammende Wörter enthalten.

Ein s.g. “Stemming”-Algorithmus ist hierbei eine Funktion T -> S, die ein Token T, welches sich in einer flektierten Form befindet, auf ein Symbol S abbilden kann, wobei gilt: Gehören zwei Tokens T1 und T2 zu dem gleichen Paradigma, dann und nur dann werden sie auf das gleiche Symbol S abgebildet.

Für die deutsche Sprache gibt es einige Stemmer, die sich in Vorgehensweise, Leistung und Verfügbarkeit in Programmiersprachen unetrscheiden. Ziel von Leonies Arbeit ist es, diese Stemmer vergleichend zu evaluieren, und basierend auf dieser Analyse einen “State-Of-The-Art” Stemmer zu entwickeln und programmiersprachenübergreifend zu implementieren.

Am Beispiel verschiedener Formen des Substantives “Adler” und des Imperativ “adle” wurde die Leistung mehrerer Stemmer vorgestellt:

1. Der aktuell meistverwendete Stemmer für Deutsch ist “Snowball”, eine Variante des Porter Stemmers. Dieser arbeitet auf Basis der Hypothese, dass in jedem Wort eine Region existiert, die für alle Worte eines Paradigmas einzigartig und gemeinsam ist: Er arbeitet daher sehr aggressiv.
2. “Text::German” ist ein System, welches Stemming auf Basis von potenziell entfernbaren Präfixen und Suffixen, sowie Vokaländerungen, und manuellen Zuordnungen von Wörtern zu diesen durchführt. Das System zeigt akzeptable Leistung, ist aber sehr schwer wartbar.
3. “Caumanns” ist ein theoretisches System ohne standardisierte Implementierung. Im Rahmen dieser Arbeit zeigte sich, dass die ursprünglich verwendete Impementierung mangelhaft war, und eine eigene geschreiben werden musste.
4. “UniNe” ist ein Stemmer von der Uni Neuchatel, welcher zwei Modi bietet: Light und Aggressive. Keiner dieser Modi arbeitet jedoch hinreichend zufriedenstellend.

Folgende Stemming-Goldstandards wurden zur Evaluierung der oben ganannten Systeme über dem CELEX2-Korpus verwendet:

1. Stemming auf Basis der Konkatenation von semantischen Morphemen (z.B. “Be-licht-mess-er” wird zu “lichtmess”).
2. Gestemmten Wortformen aus dem CELEX2-Korpus

Basierend auf diesen Goldstandards konnte festgestellt werden, dass der Caumann-Algorithmus mit 93%, bzw. 94% F1-Score die leistungsfähigste Vorgehensweise bietet.

Vor dem Hintergrund dieser Erkenntnis implementierte Leonie den neuen CISTEM-Stemmer. Dieser arbeitet auf Basis einer 4-schrittigen vorgehensweise:

1. Transformiere Wort in Kleinbuchstaben
2. Ersetze ß, ge*, *sch*, [x]+, Umlaute, *ei*, *ie*
3. Solange Länge des Wortes > 3:
    - Entferne em, er, nd, t, e s, n am Ende des Wortes
4. Schritte 1 & 2 Rückgängig machen

Dieses System konnte die Leistung über den definierten Goldstandards gegenüber dem Caumann-Algorithmus sogar auf 93%, bzw. 95% steigern. In weiterfolgender Arbeit soll der Algorithmus in möglichst vielen Programmiersprachen implementiert und zur Verfügung gestellt werden.