

Übungsaufgaben 2

- 2.1 Lesen Sie im Jurafsky/Martin die Abschnitte 17.3 (Information Retrieval) und 17.4. Lesen Sie im Carstensen et al. den Abschnitt 5.5 (Information Extraction).
- 2.2
- 2.2.1 Natürlich-sprachliche Systeme verwenden entweder Wortformenlexika, oder sie verwenden eine Flexionsmorphologie (Lemmatisierer), und brauchen dann nur die Wortstämme explizit als Einträge aufzuführen (Stammllexika). Die Ersparnis durch eine Flexionsmorphologie ist von Sprache zu Sprache stark verschieden: Bei der flexionsarmen englischen Sprache ist das Verhältnis Wortformen: Stämme und damit der Einsparungsfaktor < 2 . Bei Sprachen wie dem Türkischen und Finnischen liegt er eher in der Gegend von 50 oder 100. Beim Deutschen liegt er irgendwo zwischen den Extremen. Aufgabe: Versuchen Sie, die Relation Wortformen:Stämme für das Deutsche abzuschätzen, und beschreiben Sie Ihr Vorgehen. Hinweis: Wichtig ist, daß Sie nachvollziehbar darstellen, wie Sie zu Ihrem Resultat gekommen sind. Es ist nicht erforderlich, daß Sie bei einer genauen Ziffer ankommen. Eine Schätzung der Größenordnung reicht. Wenn Sie Schwierigkeiten oder Probleme für die Schätzung sehen, stellen Sie sie bitte ebenfalls dar.
- 2.2.2 Testen Sie vier bekannte Internet-Suchmaschinen und berichten Sie, ob sie irgendeine Art von Morphologie verwenden (z.B. Suche nach Stammform beim Verb, Singular/Plural beim Nomen, etc.). Denken Sie daran, daß die Suchmaschinen für unterschiedliche Sprachen über verschieden viel Wissen verfügen können.
- 2.3 Testen Sie das QA -System auf <http://answerbus.coli.uni-sb.de>.
- 2.3.1 Geben Sie Beispiele für jeweils zwei Anfragen, für die brauchbare bzw. unbrauchbare Ergebnisse gefunden werden. Beispielfragen finden Sie z.B. auf der TREC-Website (http://trec.nist.gov/data/topics_eng/). Versuchen Sie, **kurz** zu charakterisieren, was schwere Fragen schwer macht.
- 2.3.2 Woher bezieht Answerbus die Dokumente, die es zur Extraktion der Antworten benutzt?
- 2.3.3 Ändert sich die Brauchbarkeit der Antworten, wenn man Fragen umformuliert? Wenn ja: welche Formulierungen funktionieren besser? Begründen Sie anhand von Beispielen.
- 2.4 Auf der Folie 5 der aktuellen Vorlesung („Fragen über Fragen“) finden Sie drei Fragen. Geben Sie für jede der Fragen an, welche Methode (IE/IR/QA) Sie zur Beantwortung dieser Frage am besten einsetzen würden und begründen Sie.