2. Übungsblatt - Abgabe: 10.11.2015

Aufgabe 2.1 Morphologie

Natürlich-sprachliche Systeme verwenden entweder Wortformenlexika, oder sie verwenden eine Flexionsmorphologie (Lemmatisierer), und brauchen dann nur die Wortstämme explizit als Einträge aufzuführen (Stammlexika). 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.

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. Falls Sie in Ihrer Abschätzung Angaben aus Büchern oder aus dem Internet nutzen, geben Sie bitte die vollständige Quelle an.

Aufgabe 2.2 Endliche Automaten und Morphologie

(a) Das Adjektiv-Endungs-Diagramm auf der Vorlesungsfolie akzeptiert das Wort *st*. Die Superlativendung *kleinst* kommt aber für sich allein gar nicht vor. Korrigieren Sie den NEA mit möglichst wenig Aufwand so, dass *st* für sich genommen nicht mehr akzeptiert wird (alle übrigen Endungen dagegen wie bisher).

(b) Entwerfen Sie einen möglichst einfachen NEA, der die einfachen flektierten Formen des Verbs *reden* spezifiziert (Präsens und Präteritum, ohne Partizipialendungen!).

Aufgabe 2.3 grep und Reguläre Ausdrücke

In der folgenden Aufgabe (und auf manchen weiteren Übungsblättern) werden Sie mit dem Tiger-Corpus arbeiten. Das Tiger-Corpus enthält ca. 50000 deutsche Sätze, die mit morphologischen und syntaktischen Informationen wie zum Beispiel Wortarten annotiert sind. Auf der Vorlesungshomepage ist ein Ausschnitt aus dem Tigercorpus verlinkt (tiger.txt), bei dem in jeder Zeile ein Wort steht.

Ein regulärer Ausdruck ist eine Zeichenkette, die eine Sprache beschreibt. Sie sind äquivalent zu NEAs und DEAs. Dabei gelten folgende Notationen:

.	irgendein beliebiges Zeichen
.*	kein oder beliebig viele Zeichen
[a-e]	a, b, c, d, e
[^a-e]	alle Zeichen außer a, b, c, d, e
(maus|hund)	Zeichenfolge maus oder hund
(ab)*	kein oder beliebig viele ab
(ab)+	mindestens ein oder beliebig viele ab
(ab)?	kein oder ein ab
^	Anfang einer Zeile
$	Ende einer Zeile

Mit dem Kommando *grep* können Sie mit regulären Ausdrücken in Dateien suchen. Die Anfrage *grep -e "abc" datei.txt* gibt Ihnen alle Zeilen einer Datei, die auf den Ausdruck *abc* matchen.

(a) Es gibt verschiedene Ausdrücke, mit denen diese Aufgabe äquivalent gelöst werden kann. Finden Sie mithilfe von man page und Internetsuche heraus, wodurch sich die Kommandos *grep -e* und *grep -E* unterscheiden. Richten Sie dabei besonderes Augenmerk auf die Behandlung von Sonderzeichen.

(b) Versuchen Sie einen regulären Ausdruck zu finden, mit dem Sie alle Wortformen des Wortes *reden* im Tigercorpus finden und geben Sie ihn an. Benutzen Sie *grep -e*.

(c) Wie muss der Befehl stattdessen aussehen, wenn Sie *grep -E* benutzen?

(d) Benutzen Sie *wc*, um herauszufinden, wieviele einzelne Vorkommen es sind.
Hinweis: Sie können mit | den Output eines Befehl in einen zweiten umleiten. Mit > können Sie den Output eines Befehls in eine Datei umleiten, z.B. *wc Beispielt.txt > output.txt*.

Aufgabe 2.4 Endliche Automaten und Syntax

(a) In der Vorlesung haben wir einen NEA betrachtet, der einige zulässige Wortartketten für Nominalausdrücke spezifiziert. Der NEA spezifiziert Nominalausdrücke mit Artikel, Gattungssubstantiv, und einem potentiellen pränominalen adjektivischen Attribut, möglicherweise mit Gradpartikeln (GPRT). Erweitern bzw. modifizieren Sie den Automaten so, dass es auch Nominalausdrücken mit post-nominalen Präpositionalausdrücken umfasst. Also z.B.:

ART NN (*das Auto*)
ART ADJA ADJA NN (*das neue schnelle Auto*)
ART ADJA NN APPR ART NN (*das grüne Auto auf dem Parkplatz*)
ART NN APPR ART NN APPR ART ADJA NN (*das Auto auf dem Parkplatz bei dem neuen Institutsgebäude*)

(b) Versuchen Sie, Nominalausdrücke mit Personalpronomen (PPER) und Eigennamen (NE) im Diagramm mit zu berücksichtigen und geben Sie jeweils ein Beispiel an, das von Ihrem Automaten akzeptiert wird.

(c) Geben Sie zwei unterschiedliche Beispiele für Nominalausdrücke, die durch das Zustandsdiagramm nicht abgedeckt sind. – Erweitern Sie das Diagramm so, dass es auch diese Fälle akzeptiert.

Aufgabe 2.5 Formale Schreibweise von Automaten

Gegeben sei der Automat $A = \langle K, \Sigma, \Delta, s, F \rangle$ mit

$$
\begin{aligned}
K &= \{1, 2, 3, 4\} \\
\Sigma &= \{a, h, !\} \\
s &= 1 \\
F &= \{1, 4\} \\
\Delta &= \{\langle 1, h, 2\rangle, \langle 2, a, 3\rangle, \langle 3, h, 2\rangle, \langle 3, !, 4\rangle\}
\end{aligned}
$$

(a) Beschreiben Sie informell die Sprache, die der Automat akzeptiert. Wie lang ist das kürzeste, wie lang das längste akzeptierte Wort?

(b) Geben Sie drei verschiedene Möglichkeiten an, A so zu verändern, dass auch Wörter erkannt werden, die kein Ausrufezeichen am Ende haben, (aber ansonsten genauso wie in a) angegeben aufgebaut sind). Zulässige Veränderungen dabei sind: Das Einfügen einer neuen Transition in Δ, sowie das Hinzufügen eines Zustandes zu F.

(c) Ändern sie A so, dass Wörter mit beliebig vielen a's statt einem a hinter dem h (z.B. haahaaa!) erkannt werden. Ändern Sie dabei nur Δ und lassen Sie den Automaten ansonsten unverändert.

(d) Benutzen Sie wieder den Befehl *grep* mit einem regulären Ausdruck, um alle Vorkommen von nichtleeren Strings, die der ursprüngliche Automat in a) erkennt, in der Datei haha.txt (verlinkt unter diesem Übungsblatt) zu finden. Informieren Sie sich, was die Option *-o* in *grep* macht und wenden Sie sie an.

Hinweise: In b) und c) sollen Sie keine Zustandsdiagramme zeichnen, sondern die neuen Automaten formal angeben. In d) müssen Sie im regulären Ausdruck auch das Ausrufezeichen escapen.

Abgabe in Gruppen von bis zu drei Studierenden am **10.11.2015** vor der Vorlesung.