Einführung in die Computerlinguistik
Sommersemester 2005
Universität Konstanz
Miriam Butt

Hausaufgabe 1
— POS und Finite State Transducers —

Wer einen Anglistikschein machen möchte, sollte bitte jeweils immer die englischen Sätze analysieren—bitte "English" auf die Hausarbeit schreiben (N.B.: wer möchte, kann alle Aufgaben auf English beantworten.) Wer einen Germanistikschein machen möchte, sollte bitte jeweils immer die deutschen Sätze analysieren (bitte "Deutsch" auf die Hausarbeit schreiben). Ansonsten kann man frei wählen.

1 POS Tagging

In (1) und (2) sind ein deutscher und ein englischer Satz aufgeführt (`http://www.spiegel.de/`, `http://mobile.theonion.com/` oder `http://theonion.com` vom 4.5.05).

(1) Kaum klettern die Temperaturen, erwacht die Lust am Protest: In Freiburg campieren Studenten seit Montag im Uni-Rektorat, in Stuttgart besetzten sie gleich einen ganzen Radiosender.

(2) According to a report by the Bureau of Accessory Statistics, each year the U.S. loses more pairs of sunglasses per person than any other nation.

1.1 Manuelle Annotation

Annotieren Sie den Satz mittels gängiger Part-of-Speech (POS) Tags: STTS für das Deutsche, Penn Treebank Tagset für das Englische.

1.2 Automatische Annotation

Suchen Sie 3 weitere englische oder deutsche Sätze aus Zeitungstexten aus und annotieren Sie sie mittels einer der im Web erhältlichen POS-Tagger (z.B. http://infogistics.com/ oder http://www.ifi.unizh.ch/CL/tagger/). Evaluieren und diskutieren Sie kurz das Resultat.

2 Morphologie Allgemein

- Wieso ist eine morphologische Analyse bei der maschinellen Sprachverarbeitung wichtig (nur eine kurze Antwort wird erwartet)?
- Beschreiben Sie kurz einige Vorteile und Nachteile der "Porter Stemmer" (Beispiele sind z.B. hier zu finden: http://snowball.tartarus.org/).

3 Finite-State Methoden

3.1 Einfache Netzwerke

Entwerfen Sie ein Finite-State-Netzwerk, das die Wörter in (3) annehmen würde, aber nicht die Worte in (4).

(3) graben, grub, grabe, grubst, Hunde, Spatzen, Spatz, Hund

(4) grube, grabst, Hunden, Spatze

3.2 Netzwerke Live

XRCE (Xerox Research Centre Europe) hat einiges an On-Line-Materialien zu Finite-State-Technologien zusammengestellt. Insbesondere kann man on-line reguläre Ausdrücke eingeben, um dadurch finite-state Netzwerke zu erzeugen und so morphologische Analyse zu betreiben.

- Gehen Sie zur Seite
  `www.xrce.xerox.com/competencies/content-analysis/fst/` und klicken Sie von da aus auf *Xerox Finite State Compiler* (`www.xrce.xerox.com/competencies/content-analysis/fsCompiler/fsinput.html`).
- Auf dieser Seite kann man jetzt Finite-State-Netzwerke mittels regulärer Ausdrücke eingeben und dann auch das Netzwerk abfragen. Beispiele, wie das gehen kann, sind unter **Examples** zu finden (`www.xrce.xerox.com/competencies/content-analysis/fsCompiler/fsexamples.html`).
- Wir werden mit den Beispielen 1 und 2 arbeiten.

3.3 Morphologische Analyse

1. Kopieren Sie den regulären Ausdruck in *Figure 1.1* von der *Examples* Seite und geben Sie ihn in den Kompiler ein (das Netzwerk wird kompiliert wenn der *Submit* Knopf gedrückt wird — eine weitere Webseite zeigt die Resultate der Kompilation an).
2. Testen Sie das Netzwerk durch *Upward Application* (d.h., nachsehen, was die morphologische Analyse des Wortes ist) mit den folgenden W¨ortern: *leave, leafs, left, leaft* (man muss nochmal auf "Submit" dücken).
3. Welche nimmt das Netzwerk an und welche nicht? Warum?

3.4 Französisch

1. Lesen Sie sich das Beispiel 2 (Elision and Contraction) genau durch und versuchen Sie, das aufgestellte Regelwerk nachzuvollziehen.
2. Kopieren Sie nun den regulären Ausdruk in *Figure 2.4* in den Kompiler und erstellen Sie ein Netzwerk.
3. Testen Sie die in *Figure 2.5* angegebenen Beispiele aus. Funktioniert alles, wie man es haben wollte? Warum werden die Beispiele *Gare de l'Est et Gare de le Nord, Gare de le Est et Gare du Nord, Gare de le Est et Gare de le Nord* nicht angenommen?