Übungsblatt II

Brigitte Krenn, ÖFAI
brigitte@ofai.at
`http://www.oefai.at/~brigitte`

1 Nützliche Infos

1.1 Files

Hier finden Sie die Pfade für die in Übung II zu benutzenden Files.

Inhalt	Pfad
Referenzlexikon	/raid/stud/archive/vu_lm/Tools/l-chunkie/models/negra-pos.lex
Negra Corpus im Exportformat	/raid/stud/archive/vu_lm/Korpora/negra-corpusv2/negra-corpus.export.gz
Minimalistischer Tagger	/raid/stud/archive/vu_lm/Tools/mt/mini.pl
Programm zum Extrahieren der ersten 20 Sätze aus negra-corpus.export	/raid/stud/archive/vu_lm/Tools/mt/get_sents.pl
Tagset (STTS)	/raid/stud/archive/vu_lm/Tools/tnt/doc/tagset
Testsätze	/raid/stud/archive/vu_lm/Testdata/negra-neu
Testsätze	/raid/stud/archive/vu_lm/Testdata/news-test

1.2 Emacs für Anfänger

Der Texteditor wird mittels
emacs Filename
aufgerufen. Bezeichnet Filename ein bereits vorhandenes File wird dieses geladen, ansonsten wird ein neues File namens Filename erzeugt. Das File *tagset* wird zum Beispiel folgendermassen geöffnet:
emacs tagset
(wenn emacs aus dem Directory aufgerufen wird, in dem das File *tagset* steht) oder
emacs /raid/stud/archive/vu_lm/Tools/tnt/doc/tagset
(wenn emacs aus einem anderen Directory aufgerufen wird).

Befinden wir uns im emacs, wird ein File mit CTRL-x CTRL-f Filename geladen. Das geladene File wird mit CTRL-x CTRL-s gespeichert. Mit CTRL-x CTRL-c steigt man aus dem emacs wieder aus.

Den Cursor kann man mit der Maus bewegen oder mit den Pfeiltasten, bzw. mit CTRL-e ans Ende der Zeile, CRTL-a an den Anfang der Zeile, etc.

Seitenweise vor- und zurückgeblättert wird mit CTRL-v bzw. ESC-v.

Einzelne Zeichen können mit CTRL-d gelöscht werden, Wörter bzw. Wortteile (Cursorposition bis Leerzeichen) mit ESC-d, Zeilen bzw. Teilzeilen (Cursorposition bis Zeilenende) mit CTRL-k. Gelöschtes kannn mit CTRL-y wieder eingesetzt werden.

Eine Handlung rückgängig machen kann man mit CTRL-x u.

Emacs kann natürlich viel mehr. Auskunft darüber gibt es mittels
ESC-x info.

2 Minimalistischer Tagger

Wie in der Vorlesung besprochen, ist der einfachste Ansatz für Part-of-Speech Tagging jedem Wort in einem neuen Text das Tag zuzuweisen, welches dem entsprechenden Wort in einem Referenzlexikon am häufigsten zugeordnet wird.

Als Referenzlexikon verwenden wir ein Lexikon, das aus dem Negra-Korpus extrahiert wurde, siehe *negra-pos.lex*.

2.1 Untersuchung des Referenzlexikons

Aufgabe: Wieviele Wörter enthält *negra-pos.lex*? (Benutzen Sie **egrep -v** und **wc -l.**

Aufgabe: Wieviele Wörter in *negra-pos.lex* haben nur ein einziges PoS-Tag? (Benutzen Sie **mawk '{if (NF == 4) print $_}'** und **wc -l.**)

2.2 Taggen von bekanntem Material

Für einen ersten Test des minimalistischen Taggers benutzen wir das Negra Korpus (*negra-corpus.export.gz*). Das heißt, wir verwenden bereits bekanntes Material, was somit ein Fall von (normalerweise unerwünschtem) "testing on the training set" ist.

Aufgabe: Taggen Sie die ersten 20 Sätze aus *negra-corpus.export.gz* mit dem minimalistischen Tagger **mini.pl**. Vergleichen Sie das Ergebnis mit der ursprünglichen Annotation in *negra-corpus.export.gz*.

Dazu sind folgende Teilaufgaben zu lösen.

1. Extrahieren Sie Sätze 1 bis 20 aus dem Korpusfile *negra-corpus.export.gz*. Benutzen Sie **zcat** und **get_sents.pl**. Schreiben Sie das Ergebnis in ein File namens *negra-orig*.

2. Extrahieren Sie die Wörter aus *negra-orig* mittels **mawk**, und wenden Sie **mini.pl** auf die Wortliste an (**mini.pl** erwartet ein Wort pro Zeile als Eingabe). Schreiben Sie das Ergebnis von **mini.pl** in ein File namens *negra-mini*.

   mini.pl _Referenzlexikon_ _Text_ > _taggedText_

3. Vergleichen Sie die Files *negra-orig* und *negra-mini*. Verwenden Sie **diff** für den Vergleich. Zur Verwendung von **diff** siehe die man pages (der Befehl lautet **man diff**).

4. Evaluieren Sie den minimalistischen Tagger: Wieviele Tags wurden korrekt annotiert? Welche Fehler sind aufgetreten?

2.3 Taggen von unbekanntem Material

Damit wir eine bessere Vorstellung von der Qualität des minimalistischen Taggers bekommen, wenden wir den Tagger auf neues Material an.

- Das File *negra-neu* beinhaltet Sätze aus dem Frankfurter Rundschau Korpus, die jedoch nicht im Negra Korpus enthalten sind. D.h.: Referenzlexikon und Testtext kommen aus dem selben Grundkorpus bzw. der selben Domäne, der Testtext ist jedoch nicht direkt in das Referenzlexikon eingeflossen.

- Das File *news-test* beinhaltet Sätze aus einem Korpus von Newsgruppenbeiträgen. Hier unterscheiden sich Referenzlexikon und Testtext grundlegend.

3 Allgemeines

Aufgabe: Überlegen Sie sich eine Behandlung für unbekannte Wörter.

Aufgabe: Nennen Sie Vor- und Nachteile des minimalistischen Tagginganssatzes.

4 Tagging mit TnT

Aufgabe: Benutzen Sie TnT zum Taggen der Sätze in den Files
/raid/stud/archive/vu_lm/Testdata/negra-neu und
/raid/stud/archive/vu_lm/Testdata/news-test. Benutzen Sie einen der drei in TnT zur Verfügung stehenden Modi zur Behandlung des Sparse-Data-Problems. Schreiben Sie die Ergebnisse jeweils in ein File, und vergleichen Sie diese mit den entsprechenden Ergebnissen von mini.pl.

Zusatzinformation: Die Benutzeranleitung zu TnT finden Sie unter
/raid/stud/archive/vu_lm/Tools/tnt/doc/manual.ps. Ps-Files lassen sich mit **ghostview** <filename> anschauen. Tagging mit TnT wird in Abschnitt 4.2 beschrieben. Der Tagger liegt unter
/raid/stud/archive/vu_lm/Tools/l-chunkie/tnt.
Zum Tagging von deutschem Text soll das Model negra-pos siehe
/raid/stud/archive/vu_lm/Tools/l-chunkie/models/ verwendet werden.

4.1 Besprechung von TnT

Aufgabe: Was sind die Vorteile von TnT gegenüber unserem minimalistischen Tagger? (Benutzen Sie die Ergebnisse der vorherigen Aufgabe.)

Aufgabe: Diskutieren Sie anhand von Abschnitt 6 in *manual.ps* die wesentlichen Punkte für die Evaluierung eines statistischen PoS-Taggers.