

# Übungsblatt 13

## Präsenzaufgaben

### Aufgabe 1 Evaluationsmetriken

Betrachten Sie die Daten in Tabelle 1. Es handelt sich um ein vereinfachtes Tagging-Schema fürs Chunking, bei dem nur zwischen „Teil einer NP“ (NP) und „nicht Teil einer NP“ (0) unterschieden wird.

Füllen Sie Tabelle 2 aus, indem Sie die geforderten Evaluationsmetriken für binäre Klassifikation berechnen.

Sample	0	1	2	3	4	5	6	7	8	9
Ground Truth	NP	0	NP	0	0	NP	NP	NP	NP	0
Chunker 1	NP	NP	NP	0	NP	0	NP	NP	NP	NP
Chunker 2	NP	0	NP	0	0	0	0	0	NP	0
Chunker 3	0	0	0	0	0	NP	NP	NP	NP	0

Tabelle 1: Daten für Aufgabe 1

	Accuracy	Precision	Recall	F1-Score
Chunker 1				
Chunker 2				
Chunker 3				

Tabelle 2: Evaluationstabelle

### Aufgabe 2 Herunterladen von Ressourcen

Das CoNLL 2000 Korpus ist ein POS- und Chunk-getaggttes Korpus (IOB-Format), das in ein Test- und ein Trainingskorpus aufgeteilt ist. Wir werden es zum Training und zur

Evaluation von Chunk-Parsern verwenden. Laden Sie es sich dafür zunächst über die Ressource `corpora/conll2000` herunter.

Wenn Sie es erfolgreich heruntergeladen haben, können Sie folgendermaßen darauf zugreifen:

```
1 | from nltk.corpus import conll2000
2 | conll2000.chunked_sents('train.txt', chunk_types=['NP'])[99]
```

Das `chunk_types`-Argument dient der Auswahl von Chunk-Typen (in diesem Beispiel Nominalphrasen).

### Aufgabe 3 Chunking mit regulären Ausdrücken

Erstellen Sie einen einfachen `RegexParser`, der für Nominalphrasen charakteristische Tags zu NPs zusammenfasst. Solche charakteristischen Tags sind z. B. Kardinalzahlen (CD), Artikel (DT), Adjektive (JJ, JJR, JJS) und natürlich Substantive (NN, NNS, NNP, NNPS).

Weitere interessante Tags wären PDT (z. B. *both*, *a lot of*), POS ( *'s*), PRP (Personalpronomen), PRP\$ (Possessivpronomen).

Evaluieren Sie Ihren Parser anschließend auf dem CoNLL 2000 Korpus:

```
1 | test_sents = conll2000.chunked_sents('test.txt', chunk_types=['NP'])
2 | cp = nltk.RegexpParser(regex)
3 | print(cp.evaluate(test_sents))
```

### Aufgabe 4 Komplexität und Sprachverarbeitung

(a) Betrachten Sie folgenden Satz:

- (1) My brother opened the window the maid the janitor uncle bill had hired had married had closed.

und beantworten Sie hierzu die folgenden Fragen:

Um was für ein Sprachkonstrukt handelt es sich?

- ☐ center embedding
- ☐ cross-serial dependencies
- ☐ garden-path sentence
- ☐ Keine der anderen Möglichkeiten

Welchem Typ muss eine Grammatik mindestens genügen, um solche Sätze modellieren zu können?

- ☐ kontextfrei
- ☐ kontextsensitiv
- ☐ regulär
- ☐ rekursiv aufzählbar

Welche Rekursionstiefe hat der Satz?

- ☐ 0
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4

(b) Betrachten Sie folgenden Satz:

(2) The complex houses married and single soldiers and their families.

und beantworten Sie hierzu die folgenden Fragen:

Um was für ein Sprachkonstrukt handelt es sich?

- ☐ center embedding
- ☐ cross-serial dependencies
- ☐ garden-path sentence
- ☐ Keine der anderen Möglichkeiten

Ist der Satz ambig?

- ☐ Ja
- ☐ Nein

Sind Teile des Satzes ambig?

- ☐ Ja
- ☐ Nein

## Aufgabe 5 Evaluationsmetriken 2.0

(a) Wie lauten die Formeln für Precision, Recall, Accuracy und F1-Score?

(b) Berechnen Sie nun wie in Aufgabe 1 für die Daten in Tabelle 3 alle diese Metriken.

Sample	0	1	2	3	4	5	6	7	8	9
Ground Truth	NP	0	NP	0	0	NP	NP	NP	NP	0
Chunker 4	0	NP	NP	0	NP	0	NP	NP	NP	NP
Chunker 5	NP	0	NP	NP	0	0	0	0	NP	0
Chunker 6	0	NP	NP	0	0	NP	NP	NP	NP	0

Tabelle 3: Daten für Aufgabe 5

## Aufgabe 6 Chunking

Lesen Sie NLTK-Kapitel 7.2 (<http://www.nltk.org/book/ch07.html#chunking>) und beantworten Sie dazu die folgenden Fragen.

Wo ist der Unterschied zwischen Chunks und Phrasen?

- ☐ Chunks sind oft kürzer als Phrasen
- ☐ Chunks sind oft länger als Phrasen.
- ☐ Es gibt keinen Unterschied.

Wozu sind Tag Pattern ähnlich?

- ☐ XML-Tags
- ☐ part-of-speech tags
- ☐ regular expressions
- ☐ Keine der anderen Möglichkeiten

Was kann mit Chinking erreicht werden?

Welche Klammern stehen für Chinking: {} oder {}{ ?

Wofür stehen die anderen Klammern?

Wodurch kann mit einem RegexpParser eine hierarchische Chunk-Struktur aufgebaut werden (ähnlicher den hierarchischen Analysen mit einer Phrasenstrukturgrammatik)?