

# Musterlösungen zum 5. Übungsblatt

Sebastian Padó

24. Januar 2005

## Aufgabe 1a

Eine Baumbank (treebank) ist ein Korpus (d.h. eine Sammlung natürlich vorkommenden Textes, dessen Annotation in Baumform vorliegt. Solche Annotation kann zu verschiedenen sprachlichen Ebenen gehören (Syntax, Semantik, Diskurs, etc.); im engeren Sinn bezeichnet man aber vor allem Korpora mit syntaktischer Annotation als Baumbänke.

## Aufgabe 1b

1. Kontextfreie Grammatiken. Dazu listet man einfach alle Regeln  $LS \rightarrow RS$  auf, die in den Parsebäumen der Baumbank vorkommen.
2. Probabilistische kontextfreie Grammatiken. Dazu listet man ebenfalls alle Regeln auf; ausserdem merkt man sich die Frequenzen  $f(LS)$  (Frequenz der linken Seite) und  $f(LS \rightarrow RS)$  (Frequenz der Regel).

Nach dem Maximum-Likelihood-Prinzip ist dann die Wahrscheinlichkeit, dass das Non-terminal  $LS$  zu  $RS$  expandiert,  $P(RS|LS) = f(LS \rightarrow RS)/f(LS)$

## Aufgabe 2a

Probabilistische Grammatiken können über die **Wahrscheinlichkeit** von Analysen reden. Da Baumbank-Grammatiken sehr stark ambig sind, ist das wichtig, um wahrscheinlichere von weniger wahrscheinlichen Analysen abzusetzen.

## Aufgabe 2b

Diese Grammatik würde exakt das beschreiben, was sie gesehen hat, aber keine neuen Bäume. Sie macht keine Vorhersagen, ist insofern also ein schlechtes Modell. Stattdessen liest man Wahrscheinlichkeiten für die einzelnen Regeln ab.

## Aufgabe 2c

Ein Baum ist eine Abfolge von Regeln. Da man die Wahrscheinlichkeiten der einzelnen Regeln kennt, modelliert man die Wahrscheinlichkeit eines Baumes als gemeinsame Wahrscheinlichkeit.

Das zentrale Problem besteht darin, dass man die gemeinsame Wahrscheinlichkeit als Produkt der Regelwahrscheinlichkeiten berechnen muss, weil andere Methoden zu kompliziert wären. Dies entspricht der impliziten Annahme, dass die Anwendungen der einzelnen Regeln unabhängig voneinander wären,

Dies entspricht aber nicht den tatsächlichen Daten; zum Beispiel sind die Wahrscheinlichkeiten für die Expansion einer Nominalphrase verschieden, je nachdem, ob sie sich in Subjekt- oder in Objektposition befindet (siehe Vorlesungsfolien). Dies widerspricht der Annahme, dass die Wahrscheinlichkeit für die Regeln  $NP \rightarrow \dots$  unabhängig wäre von der Regel “obendrüber”, nämlich entweder  $S \rightarrow NP VP$  oder  $VP \rightarrow V NP$ .

## Aufgabe 2d

Damit kann man die Wahrscheinlichkeit für eine Expansion nicht nur von ihrer linken Seite (ihrer Vaterkategorie), sondern auch von ihrer Grossvaterkategorie abhängig machen. Damit kann man exakt das Problem aus (2c) lösen, indem man sagt:  $P(Det\ N|NP, S) \neq P(Det\ N|NP, VP)$ . In Worten: Die Wahrscheinlichkeit, dass  $NP$  zu  $Det\ N$  expandiert, ist unter  $S$  (in der Subjektposition) verschieden von unter  $VP$  (in der Objektposition).

## Aufgabe 2e

Man zählt in der Baumbank nicht nur lokale Bäume, sondern Paare aus lokalen Bäumen  $LS \rightarrow RS$  und Grossvaterkategorien  $G$ . Dann gilt, wieder mit dem Maximum Likelihood-Prinzip,  $P(RS|LS, G) = f(G, LS, RS)/f(G, LS)$ . Beispielsweise also für  $NPs$  in der Subjektposition:  $P(Det\ N|NP, S) = f(S, NP, Det\ N)/f(NP, S)$ . Der letzte Term lässt sich in Worten paraphrasieren als “der Anteil aller Subjekte ( $NP$  unter  $S$ ), die zu  $Det\ N$  expandieren”.

## Aufgabe 2f

Das zentrale Problem ist, dass sich die Anzahl der Regeln erhöht. Das hat drei Auswirkungen:

- Die Verarbeitung wird langsamer
- Die Regeln werden spezieller: das führt dazu, dass die Generalisierung der Grammatik schlechter wird (z.B. sinkt tendentiell die Abdeckung auf neuen Daten)
- Die Abschätzung der Regelwahrscheinlichkeiten wird schwieriger, weil es mehr Regeln gibt, auf die sich die Trainingsdaten verteilen (die einzelnen Regeln werden seltener gesehen); das führt zu “sparse data”.

Im Allgemeinen muss man daher “backing off” einsetzen, d.h. wenn das Modell mit Geschichte sich nicht sicher ist bzw. eine bestimmte Kombination noch nie gesehen hat, sollte man auf das Modell ohne Geschichte zurückgreifen.

### Aufgabe 3a

- Aufgabe, einzelne Bäume zu annotieren, einfacher als Entwicklung kompletter Grammatik, bei der Regeln interagieren
- Enthält Analysen in deklarativer Form: Kann auch für sonstige linguistische Zwecke eingesetzt werden
- Möglichkeit, für verschiedene Zwecke optimierte Grammatiken zu lernen, indem man verschiedene Lernalgorithmen einsetzt
- Hoffnung, dass mit besseren Lernalgorithmen angemessenere Grammatiken gelernt werden können
- Empirisch angemessenere Grammatiken als durch Introspektion
- Automatische Modellierung von Wahrscheinlichkeiten durch Häufigkeiten
- Modellierung des “ist”, nicht des “soll”: bessere Generalisierung auf echte Daten

### Aufgabe 3b

- Baumbank-Grammatiken i.A. linguistisch wenig relevant (viele Regeln, auch unsinnige Sätze erhalten Analysen)
- Wenig Kontrolle über die gelernten Regeln: wenig Möglichkeit, linguistisches Wissen über “sinnvolle” Regeln zu integrieren
- Baumbanken enthalten immer auch Fehler, was zu unsauberer Grammatiken führt
- Indirektionsstufe durch Lernalgorithmus führt Fehler ein
- Verbreitete Methode der Wahrscheinlichkeitsberechnung (Produkt der Regelwahrscheinlichkeiten) unangemessen; Ausnahmen müssen “von Hand” in das Modell kodiert werden (Geschichte, Lexikalisierung)

### Aufgabe 3c

Prinzipiell ist das möglich, da ja die gelernten Grammatiken genau die Phänomene abdecken, die auch in der Baumbank vorkommen. Die Frage ist, was “rekonstruieren” heisst.

- Normale kontextfreie Baumbank-Grammatiken werden für jeden Baumbank-Satz viele Analysen generieren, unter denen auch die richtige ist, aber sie bieten keine Methode an, diese Analyse zu ermitteln.

- Baumbank-PCFGs erlauben das durch ihre Wahrscheinlichkeitsverteilung. Allerdings hängt es von der Güte des Modells ab, wie gut die Original-Bäume rekonstruiert werden können: wenn die Regeln nicht lexikalisiert bzw. mit Geschichte erweitert sind, werden Bäume mit seltenen Konstruktionen vermutlich falsch analysiert.