6. Übung zur Vorlesung "NLP – Analyse des Wissensrohstoffes Text" im Sommersemester 2008 – mit MusterlÃ¶sungen –

Dr. Andreas Hotho, Dipl.-Inform. Dominik Benz, Wi.-Inf. Beate Krause

25. Juni 2008

1 Parse Tree Derivations

- Verschiedene Derivationen eines Parsebaumes machen die Berechnung der Wahrscheinlichkeit eines Parsebaumes unnötig kompliziert. Erklären Sie den Begriff Derivation in diesem Kontext. Was wird angewandt, um die unnötige Komplexität zu reduzieren?

LÖSUNGSVORSCHLAG:

Unter einer Derivation eines Parsebaumes versteht man verschiedene Reihenfolgen der Anwendung von Ableitungsregeln, die zu ein und demselben Parsebaum führen. Die Berücksichtigung aller Derivationen wäre nur unnötig kompliziert; deshalb stellt man Regeln auf, um eine einzige *kanonische Derivation* für einen Baum festzulegen.

2 Evaluation

- Das Beurteilen der Ähnlichkeit zweier Parsebäume stellt eine fundamentale Aufgabe dar. Welche Bewertungsmaße kennen Sie dafür?
- Betrachten Sie die verschiedenen Parsebäume auf Seite 410, Abb. 12.2. Parsebaum (a) sei der "Gold-Standard". Berechnen Sie die Ähnlichkeit der Parsebäume (b) und (c) mit (a) mittels mindestens 3 verschiedener Maße.

LÖSUNGSVORSCHLAG:

1. Tree Accuracy / Exact Match
2. PARSEVAL measures:
   - (Labeled) Precision
   - (Labeled) Recall

- Crossing Brackets

Für die Parsebäume aus dem Buch gilt:

Klammerung im Gold-Standard (a):
`(0,7), VP(2,7), NP(3,6)`

Klammerung in (b):
`S(0,7), VP(2,7), VP(2,3), VP(5,7)`

Klammerung in (c):
`S(0,7), VP(2,7), NP(3,7), NP(5,7)`

Damit ergeben Sich:

$precision(a,b) = \frac{2}{4} = 0.5$

$recall(a,b) = \frac{2}{3} \approx 0.66$

$labeled\_precision(a,b) = \frac{2}{4} = 0.5$

$labeled\_recall(a,b) = \frac{2}{3} \approx 0.66$

$precision(a,c) = \frac{2}{4} = 0.5$

$recall(a,c) = \frac{2}{3} \approx 0.66$

$labeled\_precision(a,c) = \frac{2}{4} = 0.5$

$labeled\_recall(a,c) = \frac{2}{3} \approx 0.66$

Für das Ausrechnen der crossing brackets ist Tabelle 2 hilfreich:

Daraus folgt:

$crossing\_brackets(a,b) = 1$

$crossing\_brackets(a,c) = 2$

3 Independence Assumption

- Worin besteht die Independence Assumption bei PCFGs?
- Erklären Sie einige Phänomene natürlicher Sprache, die der Independence Assumption widersprechen. Kennen Sie Ansätze, die diesen Problemen begegnen?

Die Unabhängigkeitsannahme besteht darin, dass die Anwendung jeder Ableitungsregel gleichwahrscheinlich ist ungeachtet ihres Kontextes, in dem sie angewandt wird. So hat z.B. die Regel `VP -> V NP NP` die gleiche Wahrscheinlichkeit, ungeachtet um welches Verb es sich handelt. Diese Unabhängigkeitsannahme bildet aber z.B. das Phänomen nicht ab, dass genau diese Regel bei ditransitiven Verben wie z.B. *give* diese Regel sehr viel wahrscheinlicher ist wie bei nicht-ditransitiven Verben. Ein Ansatz, die Kontextfreiheit der Grammatik zu bewahren und "über die Hintertür" trotzdem solche

Tabelle 1: Crossing Brackets für die Parsebäume

Parse (a)								
	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7
S	(							)
VP			(					)
NP				(			$)*_{ab}, *_{ac}, *_{ac}$	
Parse (b)								
S	(							)
VP			(					)
VP			(	)				
VP					(			$)*_{ab}$
Parse (c)								
S	(							)
VP			(					)
NP				(				$)*_{ac}$
NP					(			$)*_{ac}$

Phänomene abzubilden, ist in jeder Regel das "Kopfwort" der Ableitung mitzuführen wie `VP`$_{give}$ `->` `V`$_{give}$ `NP NP`.