



Lehrstuhl Angewandte Informatik IV
Datenbanken und Informationssysteme
Prof. Dr.-Ing. Stefan Jablonski

Institut für Angewandte Informatik
Fakultät für Mathematik, Physik und Informatik
Universität Bayreuth

Bachelorarbeit

Klaus Freiburger

Juni 14, 2019

Version: Final

Universität Bayreuth

Fakultät Mathematik, Physik, Informatik

Institut für Informatik

Lehrstuhl für Angewandte Informatik IV

Rapid Miner NLP Tools: Syntax Parsing

Bachelorarbeit

Klaus Freiberger

- | | |
|--------------------|---|
| <i>1. Reviewer</i> | Dr. Lars Ackermann
Fakultät Mathematik, Physik, Informatik
Universität Bayreuth |
| <i>2. Reviewer</i> | Prof. Dr.-Ing. Stefan Jablonski
Fakultät Mathematik, Physik, Informatik
Universität Bayreuth |
| <i>Supervisors</i> | Lars Ackermann and Stefan Jablonski |

Juni 14, 2019

Klaus Freiberger

Bachelorarbeit

Rapid Miner NLP Tools: Syntax Parsing, Juni 14, 2019

Reviewers: Dr. Lars Ackermann and Prof. Dr.-Ing. Stefan Jablonski

Supervisors: Lars Ackermann and Stefan Jablonski

Universität Bayreuth

Lehrstuhl für Angewandte Informatik IV

Institut für Informatik

Fakultät Mathematik, Physik, Informatik

Universitätsstrasse 30

95447 Bayreuth

Germany

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Abstract (different language)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

1	Einleitung	1
1.1	Motivation and Problem Statement	1
1.1.1	Some References	1
1.2	Thesis Structure	1
2	Natural Language Processing	3
2.1	Grundlagen	3
2.2	Syntaktisches Parsen	7
2.2.1	Ambiguität	8
2.2.2	Dynamische Programmierung	9
2.2.3	Cocke-Kasami-Younger Algorithmus	9
2.3	Statistisches Parsen	12
2.3.1	Probabilistische Kontextfreie Grammatiken	12
2.3.2	Probabilistische Lexikalisierte Kontextfreie Grammatiken	14
3	Konzept	17
3.1	Input Text	17
3.2	Parser Output	18
3.3	Parser Modell	18
3.4	Goldstandard	19
3.5	Parser	19
3.5.1	Stanford-Parser	20
3.5.2	Berkeley-Parser	20
3.5.3	OpenNLP-Parser	20
3.6	Evaluiierung	20
4	Implementierung	23
4.1	RapidMiner Studio	23
4.2	Verwendete Operatoren	23
4.3	Eigene Operatoren	24
4.3.1	Berkeley-Parser	25
4.3.2	OpenNLP-Parser	26
4.3.3	Stanford-Parser	26
4.3.4	Compare Results	27

4.3.5 Show Results	29
5 Evaluierung	31
6 Conclusion	33
6.1 System Section 1	33
6.2 System Section 2	34
6.3 Future Work	35

Einleitung

1.1 Motivation and Problem Statement

1.1.1 Some References

[WEB:GNU:GPL:2010; WEB:Miede:2011]

1.2 Thesis Structure

Natural Language Processing

In diesem Kapitel wird der theoretische Hintergrund, die Methoden des Natural-Language-Processings, eingeführt. Nach den Grundlagen werden zwei verschiedene Methoden zum Parsen vorgestellt.

2.1 Grundlagen

Betrachtet man einen Satz in einer natürlichen Sprache, die im Rahmen dieser Arbeit auf Englisch festgelegt ist, so kann ein Algorithmus dessen Semantik nicht ohne Weiteres herauslesen. Hierfür bedarf es verschiedener Hilfsstrukturen und zusätzlicher Informationen. Als ersten Schritt bietet es sich an, den einzelnen Wörtern eines Satzes ihre Wortart zuzuordnen. Mit Wortart, alternativ auch Wortklasse oder im Englischen part-of-speech (POS), ist gemeint, wie ein Wort im Satz auftritt. Beispiele für Wortarten sind Nomen, Verb und Adjektiv. In dieser Arbeit wird das Tagset, also die Menge an Wortarten, aus der Penn Treebank verwendet. Siehe hierfür Tabelle 2.1. Der Satz

My dog also likes eating sausage.

würde mit diesem Tagset also folgendermaßen annotiert werden:

My/PRP\$ dog/NN also/RB likes/VBZ eating/VBG sausage/NN ./.

Wie ein Satz maschinell mit POS-Tags versehen werden kann, wird in dieser Arbeit nicht weiter behandelt. Über diese zusätzliche Notation hinaus kann man erkennen, dass sich in der englischen Sprache oftmals mehrere Wörter als Gruppe oder als eine Komponente innerhalb des Satzes verhalten. So eine Gruppe wäre zum Beispiel die Nominalphrase *My Dog* oder die Verbalphrase *likes eating sausage*. Auch hier wird wieder die Annotation der Penn Treebank verwendet, abgebildet in Tabelle 2.2. Zu diesem Satz an Tags gibt es noch die Erweiterung um relationale Tags. Diese liefern eine Zusatzinformation und werden an die eben vorgestellten angehängt. Zum Beispiel bekommt das Subjekt eines Satzes das Suffix *-SBJ*. Das heißt aus *NP* wird *NP-SBJ*. Dieser Erweiterungssatz wird im Rahmen dieser Arbeit herausgelassen, aber der Vollständigkeit halber erwähnt. In 2.1 wird der Satz mit der syntaktischen

Penn Treebank Part-of-Speech Tags		
Tag	Beschreibung	Beispiel
CC	Koordinierende Konjunktion	<i>and</i>
CD	Kardinalzahl	<i>third</i>
DT	Artikel	<i>the</i>
EX	Existentielles <i>there</i>	<i>there is</i>
FW	Fremdword	<i>les</i>
IN	Präposition, unterordnende Konjunktion	<i>in</i>
JJ	Adjektiv	<i>green</i>
JJR	Adjektiv, Komparativ	<i>greener</i>
JJS	Adjektiv, Superlativ	<i>greenest</i>
LS	Listenelement Markierung	<i>1)</i>
MD	Modal	<i>could</i>
NN	Nomen, singular oder Masse	<i>table</i>
NNS	Nomen, plural	<i>tables</i>
NNP	Eigennamen, singular	<i>Germany</i>
NNPS	Eigennamen, plural	<i>Vikings</i>
PDT	Predeterminer	<i>both his children</i>
POS	possessive Endung	<i>'s</i>
PRP	Personalpronomen	<i>me</i>
PRP\$	Possesivpronomen	<i>my</i>
RB	Adverb	<i>extremely</i>
RBR	Adverb, Komparativ	<i>better</i>
RBS	Adverb, Superlativ	<i>best</i>
RP	Partikel	<i>about</i>
SYM	Symbol	<i>%</i>
TO	to	<i>what to do</i>
UH	Ausruf	<i>oops</i>
VB	Verb, Grundform	<i>be</i>
VBD	Verb, Vergangenheitsform	<i>was</i>
VBG	Verb, Gerund \Partizip Präsens	<i>being</i>
VCN	Verb, Partizip Perfekt	<i>been</i>
VBP	Verb, Präsens, nicht 3.Person Singular	<i>am</i>
VBZ	Verb, Präsens, 3.Person Singular	<i>is</i>
WDT	<i>wh</i> -Artikel	<i>which</i>
WP	<i>wh</i> -Pronomen	<i>who</i>
WP\$	<i>wh</i> -Possesivpronomen	<i>whose</i>
WRB	<i>wh</i> -Adverb	<i>be</i>

Tab. 2.1: Penn Treebank POS Tags

Penn Treebank Syntactic Tags		
Tag	Beschreibung	Beispiel
S	einfacher deklarativer Satz	<i>There we go.</i>
SBAR	Satz beginnend mit unterordnender Konjunktion	<i>feels like we have to move</i>
SBARQ	Direkte Frage beginnend mit <i>wh</i> -Wort oder <i>wh</i> -Phrase	<i>So what's that about?</i>
SINV	Invertierter deklarativer Satz	<i>neither am I a pessimist.</i>
SQ	Invertierte Ja/Nein Frage oder Hauptsatz einer <i>wh</i> -Frage	<i>Will they move on?</i>
ADJP	Adjektivphrase	<i>relatively cheap</i>
ADVP	Adverbphrase	<i>down here</i>
CONJP	Konjunkionalphrase	<i>but also for tissues</i>
FRAG	Fragment	<i>if not today, ...</i>
INTJ	Zwischenruf	<i>Well</i>
LST	Listenmarkierung	<i>1</i>
NAC	Keine Komponente	<i>via the Freedom of Information</i>
NP	Nominalphrase	<i>the sun</i>
NX	Markiert Kopf in komplexen NP	<i>fresh apples and cinnamon</i>
PP	Präpositionalphrase	<i>in some way</i>
PRN	Nebenläufige Phrase	<i>..., bless his heart, ...</i>
PRT	Partikel	<i>up</i>
QP	Quantifizierende Phrase	<i>or two a day</i>
RRC	Reduzierter Relativsatz	<i>titles not presently in the collection</i>
UCP	Ungleich Koordinierte Phrasen	<i>She flew yesterday and on July 4th.</i>
VP	Verbalphrase	<i>this is my dog.</i>
WHADJP	<i>Wh</i> -Adjektivphrase	<i>how great you are</i>
WHADVP	<i>Wh</i> -Adverbphrase	<i>When I see it</i>
WHNP	<i>Wh</i> -Nominalphrase	<i>What they've done</i>
WHPP	<i>Wh</i> -Präpositional	<i>At which</i>
X	Unbekannt	<i>The more ..., the less ...</i>

Tab. 2.2: Penn Treebank Syntaktische Tags

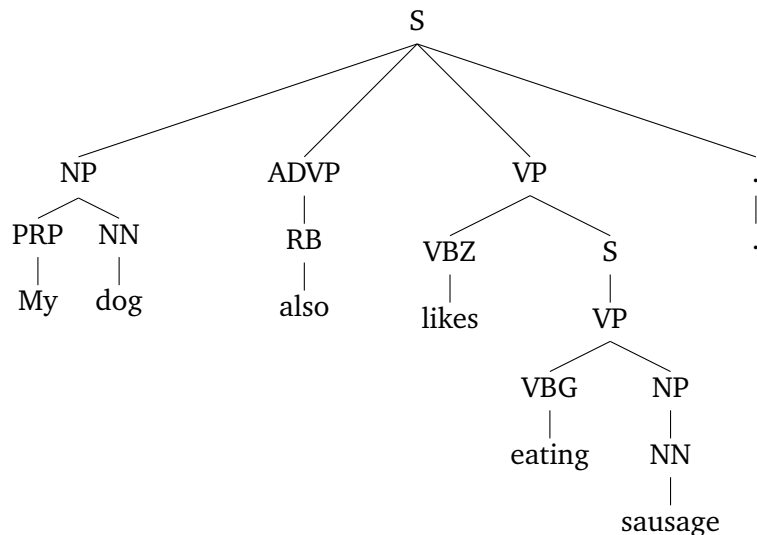


Fig. 2.1: Syntaxbaum zum Satz *My dog also likes eating sausage*.

Annotation dargestellt.

$$\begin{aligned}
 &(S \\
 &\quad (NP (PRP\$ My) (NN dog)) \\
 &\quad (ADVP (RB also)) \\
 &\quad (VP (VBZ likes) \\
 &\quad\quad (S \\
 &\quad\quad\quad (VP (VBGeating) \\
 &\quad\quad\quad\quad (NP (NN sausage)))))) \\
 &\quad (. .))
 \end{aligned} \tag{2.1}$$

Alternativ zur geklammerten Lösung kann man dieses Ergebnis auch als Syntaxbaum zeichnen, siehe hierfür Abbildung 2.1.

Wie man am Beispiel erkennen kann, sind, im Gegensatz zu den POS-Tags, auch diverse Verschachtelungen dieser Komponenten möglich. Um diese Anordnungsstruktur innerhalb einer Sprache zu beschreiben, bieten sich kontextfreie Grammatiken an. Diese Art der Grammatik zeichnet sich dadurch aus, dass sich auf der linken Seite der Regeln genau ein Nichtterminalsymbol und auf der rechten Seite können sich beliebig viele Terminale und Nichtterminale befinden. Das Nichtterminal auf der linken Seite hat keine Einschränkung welche Symbole sich um es herum befinden müssen. Es hat also keinen Kontext, daher die Bezeichnung kontextfrei. Für die Verarbeitung unseres Beispielsatzes wurden unter anderem folgende Regeln verwendet:

$$\begin{aligned}
S &\rightarrow NP \text{ } ADVP \text{ } VP \text{ } . \\
NP &\rightarrow PRP\$ \text{ } NN \\
PRP\$ &\rightarrow My \\
NN &\rightarrow dog
\end{aligned}$$

Da ein englischer Satz nicht unbedingt *S* als oberstes Nichtterminal hat, sondern auch *FRAG*, *SBARQ* und andere möglich sind, muss in den Grammatiken ein zusätzliches Startsymbol eingeführt werden. Dieses wird zum Beispiel *ROOT* oder *TOP* genannt.

Anhand des vollständigen Regelsatzes der Grammatik einer Sprache, kann man theoretisch jedem grammatikalisch korrekten Satz dieser Sprache seine syntaktische Struktur zuweisen. Es gibt für diverse natürliche Sprachen Sammlungen von annotierten Sätzen, diese werden Treebank oder Korpus genannt. Aus den dort gespeicherten syntaktischen Strukturen kann man unter anderem eine Grammatik für die Sprache ableiten. Hierfür muss man alle verwendeten Verschachtelungen als Regel auffassen.

Ein bekannter Korpus der englischen Sprache ist die Penn Treebank, aus welcher auch die hier verwendete Annotation stammt. Dieser Korpus wird vom Linguistic Data Consortium, mit Sitz in der Universität von Pennsylvania, herausgegeben. Im Rahmen des Penn Treebank Projekts wurden von 1989 bis 1992 über 4,5 Millionen Wörter der Treebank hinzugefügt. Die Texte hierfür stammen zu einem Großteil aus dem Wall Street Journal. Woher die Sätze einer Treebank stammen spielt eine große Rolle, da Parser mit Hilfe von Treebanks trainiert werden. Dazu mehr in Kapitel 2.3.

2.2 Syntaktisches Parsen

Syntaktisches Parsen wird als die “Aufgabe des Erkennens eines Satzes und des Zuweisens einer syntaktischen Struktur” definiert. Ein Parser ist damit ein Programm, das Sätze parst.

Zum Einstieg in das Kapitel wird das Parsen als Suche betrachtet. Das Suchproblem besteht darin, aus allen möglichen Bäumen, welche sich mit der Grammatik generieren lassen, den korrekten Baum zur Eingabe zu finden. Der Suchraum wird also von der Grammatik festgelegt. Ein Baum ist korrekt, wenn das Startsymbol die Wurzel ist und exakt die Eingabe abgedeckt wird. Vereinfachend wird angenommen, dass alle Sätze als oberstes Nichtterminal *S* haben. Deshalb wird dieses als Startsymbol betrachtet. Anhand der zwei Merkmale eines korrekten Baumes kann die Suche gestaltet werden. Es ergeben sich als grundlegende Ansätze die Top-Down und die Bottom-Up Suche.

Beim Top-Down Verfahren wird mit dem Startsymbol *S* begonnen und dieses mit

den Regeln der Grammatik Schritt für Schritt erweitert. Bäume deren Blätter nicht auf die Eingabe passen werden abgelehnt.

Die Bottom-Up Suche beginnt, dem Namen entsprechend, am anderen Ende des Baumes. Im ersten Schritt gibt es nur die Eingabeworte als Blätter. Es wird mit den rechten Seiten der Grammatikregeln der Baum nach oben gebaut. Hier kann also ein Baum ausgeschlossen werden, wenn seine obersten Knoten in keiner Kombination auf keiner rechten Seite einer Produktion vorkommen.

So werden mit der Top-Down Suche keine Bäume gebaut die niemals das Start Symbol als Wurzel haben, dafür entsprechen die Blätter nicht exakt der Eingabe. Beim Bottom-Up Ansatz verhält es sich genau andersherum.

2.2.1 Ambiguität

Das Hauptproblem, welches sich beim Finden des korrekten Baumes ergibt, ist die Mehrdeutigkeit oder auch genannt Ambiguität. Zum einen können Wörter mehrdeutig sein, wie etwa das englische Wort *book*, welches sowohl Verb als auch Nomen sein kann. Zum anderen, und für diesen Kontext relevanter, gibt es die strukturelle Mehrdeutigkeit. Ein Beispielsatz hierfür ist:

I shot an elephant in my pajamas.

In dieser Satzstruktur kann sich *in my pajamas* sowohl auf den Erzähler, als auch auf den Elefanten beziehen und somit gibt es mehr als einen korrekten Baum.

Diese Art der strukturellen Mehrdeutigkeit ist die Anhangs-Mehrdeutigkeit, bzw. im Englischen Attachment Ambiguity. Eine Komponente des Satzes, in diesem Fall die Präpositionalphrase, kann an mehreren Stellen angehängt werden. Kombiniert man die Präpositional- an die Verbalphrase hat der Satz die Bedeutung, dass der Schütze einen Pyjama trägt. Bindet man diese an das Nomen Elefant wird ausgesagt, dass dieser sich im Pyjama befindet. Grammatikalische Korrektheit ist in beiden Fällen gegeben.

Eine weitere Art ist die Koordinations-Mehrdeutigkeit, im Englischen Coordination Ambiguity, welche in Verbindung mit Konjunktionen und Satzverbindungen auftritt. Beispielsweise kann der Satzausschnitt *old men and women* unterschiedlich interpretiert werden. *Old* kann sich auf *men and women* oder nur auf *men* beziehen. Bringt man beide Formen der Mehrdeutigkeit in einen Satz, wie z.B.

I shot wild elephants and lions in my pajamas.

ergeben sich folgende vier Möglichkeiten (die Klammerung stellt die Zusammengehörigkeit dar):

I shot [wild elephants] and [lions] in my pajamas.
I shot [wild elephants and lions] in my pajamas.
I shot [wild elephants] and [lions in my pajamas].
I shot [[wild elephants and lions] in my pajamas].

In den ersten beiden Versionen bezieht sich *in my pajamas* auf *shot*, danach auf *lions*, bzw. *wild elephants and lions*. Auch *wild* wechselt zwischen einem und zwei Bezugswörtern. Die Anzahl kann also exponentiell wachsen. Von diesen Interpretationen beschreibt nur eine den Inhalt des Satzes so, wie er vom Autor gemeint ist. Ein Parser braucht also weitere Kriterien um sich zwischen den verschiedenen Optionen entscheiden zu können. Hierzu mehr in Kapitel 2.3. Ohne zusätzliche Informationen kann der Parser lediglich alle grammatikalisch möglichen Bäume erstellen. Um diese Aufgabe effizient zu bewerkstelligen bietet sich dynamisches Programmieren an.

2.2.2 Dynamische Programmierung

Dynamisches Programmieren ist hier sinnvoll, da beim Erstellen des Baumes im Allgemeinen Mehrfacharbeit anfällt. Baut der Parser zum Beispiel die Bäume von oben nach unten und versucht dabei die Eingabe von links nach rechts abzudecken, dann findet er in der Regel einen Baum der einen Teil des Satzes abdeckt. Da noch nicht die gesamte Eingabe enthalten ist, handelt es sich nicht um einen korrekten Baum. Dennoch wird dieser Baum in der Regel immer wieder gebaut, bzw. ist sogar in der tatsächlichen Lösung enthalten. Dass dieser Teilbaum nicht jedes mal neu gefunden werden muss, bietet es sich an ihn abzuspeichern. Als Algorithmen für zum Parsen haben sich das Chart Parsing, der Earley Algorithmus und der Cocke-Kasami-Younger(kurz CKY) Algorithmus etabliert. Letzterer wird nachfolgend etwas genauer ausgeführt.

2.2.3 Cocke-Kasami-Younger Algorithmus

Zu Beginn des Algorithmus muss die verwendete Grammatik in die Chomsky-Normalform (kurz CNF) überführt werden. Die CNF zeichnet sich dadurch aus, dass die rechte Seite aus genau zwei Nichtterminalen oder einem Terminal besteht. Das Überführen einer kontextfreien Grammatik in diese Form erfolgt nach einem einfachen Regelsatz und wird nicht hier nicht näher betrachtet. Auch die Rückführung von der CNF in die ursprüngliche Grammatik ist, mit zusätzlich abgespeicherten Informationen, möglich. Diese Rückführung wird gemacht, da die Ergebnisbäume der CNF Struktur und nicht der eigentlichen Grammatik entsprechen. Somit ist auch

eine semantische Analyse komplizierter.

Nach der Überführung in die CNF wird für einen Satz mit n Wörtern eine obere Dreiecksmatrix der Größe $(n+1) \times (n+1)$ aufgestellt. Die Positionen im Satz werden mit 0 beginnend durchnummeriert. Vor dem ersten Wort steht Index 0, nach dem letzten Index n . Die Zelle $[i, j]$ enthält diejenigen Nichtterminale, welche die Eingabe von i bis j abdecken. Dabei gilt $0 \leq i < n-1$, $1 \leq j < n$ und $i < j$. Die Zelle $[0, n]$ gibt somit an, welche Nichtterminale die komplette Eingabe abdecken.

Die Zellen $[i, i+1]$ bilden die Diagonale der Matrix. Hier wird die Eigenschaft der CNF ausgenutzt, dass alle Terminale auf der rechten Seite der Produktionen alleine stehen und somit jedes von mindestens einem Nichtterminal abgedeckt wird.

Die Eigenschaft, dass Nichtterminale auf der rechten Seite immer zu zweit auftreten führt dazu, dass für die eine Zelle $[i, j]$ folgendes gilt. Für jede Regel $A \rightarrow B C$ die i bis j abdeckt, gibt es ein k , mit $i < k < j$, sodass das B den Abschnitt $[i, k]$ und C $[k, j]$ abdeckt. Ersteres befindet sich in der selben Zeile, links von der entsprechenden Zelle. Letzteres ist in der selben Spalte, unterhalb zu finden. Das Füllen der Matrix muss von links nach rechts und von unten nach oben geschehen.

Eine gefüllte Matrix für den Satz *Book the flight through Houston*. ist in Abbildung ?? dargestellt. Hierfür wurde eine sehr kleine Grammatik verwendet die nur knapp mehr als die verwendeten Wörter abdeckt. Diese ist entnommen aus , von wo auch die Abbildung entnommen ist. Neben der CNF Regeln sind auch die ursprünglichen dargestellt.

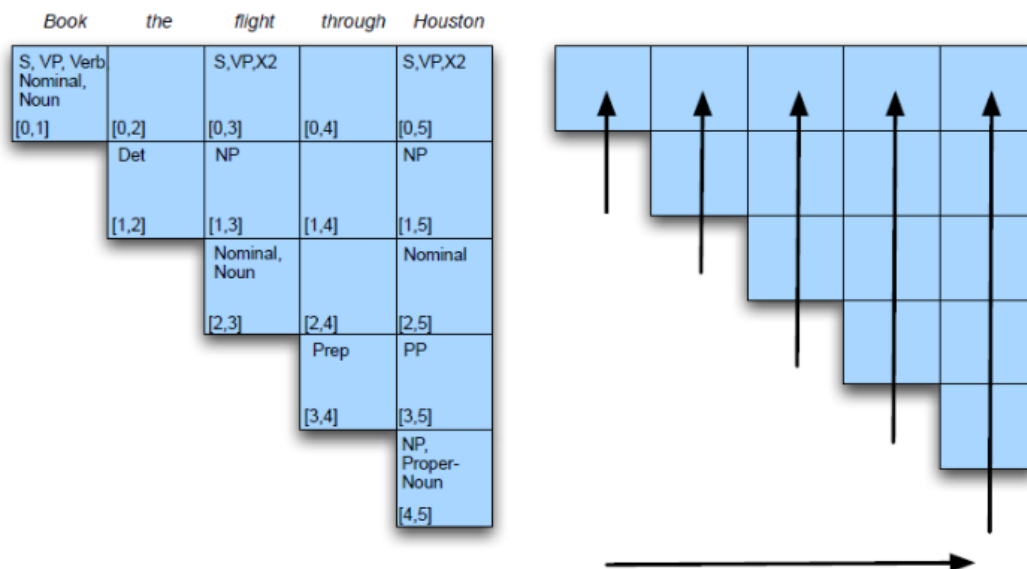


Fig. 2.2: Beispielmatrix für den CKY-Algorithmus, entnommen aus ..

Ohne zusätzliche Informationen gibt die Tabelle nur an, ob ein Satz grammatikalisch korrekt ist. Dies ist er, falls in Zelle $[0, n]$ das Startsymbol steht. Um außerdem die verwendeten Regeln, also den Baum, bzw. die Bäume zur Eingabe zu erhalten,

\mathcal{L}_1 Grammar	\mathcal{L}_1 in CNF
$S \rightarrow NP VP$	$S \rightarrow NP VP$
$S \rightarrow Aux NP VP$	$S \rightarrow XI VP$
	$XI \rightarrow Aux NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb NP$
	$S \rightarrow X2 PP$
	$S \rightarrow Verb PP$
	$S \rightarrow VP PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper-Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det Nominal$	$NP \rightarrow Det Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal Noun$	$Nominal \rightarrow Nominal Noun$
$Nominal \rightarrow Nominal PP$	$Nominal \rightarrow Nominal PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb NP$	$VP \rightarrow Verb NP$
$VP \rightarrow Verb NP PP$	$VP \rightarrow X2 PP$
	$X2 \rightarrow Verb NP$
$VP \rightarrow Verb PP$	$VP \rightarrow Verb PP$
$VP \rightarrow VP PP$	$VP \rightarrow VP PP$
$PP \rightarrow Preposition NP$	$PP \rightarrow Preposition NP$

Fig. 2.3: Grammatik für entnommen aus ..

muss der Algorithmus erweitert werden. Die Erweiterung besteht darin, zu jedem Nichtterminal in jeder Zelle zusätzlich zu speichern, aus welchen zwei anderen Nichtterminalen es entstanden ist. Darüber hinaus darf das gleiche Nichtterminal mehrmals in einer Zelle vorkommen, da es aus unterschiedlichen Nichtterminalen entstanden sein kann. Ein Beispiel zur Tabelle in Abbildung ?? ist in Zelle $[0, 5]$ das Symbol S . Es gibt drei Möglichkeiten dieses zu erzeugen:

$$\begin{aligned}
[0, 1] : Verb \text{ und } [1, 5] : NP \\
[0, 3] : VP \text{ und } [3, 5] : PP \\
[0, 3] : X2 \text{ und } [3, 5] : PP
\end{aligned} \tag{2.2}$$

Damit ergeben sich für die Eingabe drei verschiedene Syntaxbäume. Diese werden erstellt, indem man die Tabelle, beginnend mit dem jeweiligen Startsymbol, rückwärts durchläuft.

2.3 Statistisches Parsen

In diesem Abschnitt werden Mittel vorgestellt, welche dem Parser helfen sich zwischen mehreren, grammatikalisch korrekten Bäumen eines Eingabesatzes zu entscheiden. Hierfür benötigt er für jeden errechneten Baum zusätzlich die Wahrscheinlichkeit mit welcher dieser semantisch korrekt ist. Das heißt, jede Lösung ist mit einer Wahrscheinlichkeit versehen und es kann diejenige, dessen Wert am höchsten ist als Lösung gewählt werden.

2.3.1 Probabilistische Kontextfreie Grammatiken

Die Anforderung, einem Baum eine Wahrscheinlichkeit zuzuweisen, lässt sich mit dem Konzept der probabilistischen kontextfreien Grammatiken (abgekürzt: PCFG) umsetzen. Abgesehen von zwei Erweiterungen haben die Produktionen einer PCFG die gleiche Form wie die der ursprünglichen kontextfreien Grammatik. Erstens wird jeder Regel eine Wahrscheinlichkeit p , mit $0 \leq p \leq 1$, hinzugefügt.

$$A \rightarrow \beta[p] \quad (2.3)$$

Diese gibt an, mit welcher Wahrscheinlichkeit die rechte Seite und der Vorbedingen der linken Seite auftritt.

$$p := P(A \rightarrow \beta | A) \quad (2.4)$$

Es handelt sich also um eine bedingte Wahrscheinlichkeit.

Zweitens muss für jedes Nichtterminal die Summe der Wahrscheinlichkeiten aller seiner Produktionen 1 ergeben.

$$\sum_{\beta} P(A \rightarrow \beta) = 1 \quad (2.5)$$

Die Wahrscheinlichkeit für einen Baum errechnet sich dann durch das Produkt der Wahrscheinlichkeiten aller verwendeten Regeln. Eine Grammatik heißt konsistent, falls die Summe der Wahrscheinlichkeiten aller möglichen Sätze 1 ergibt. Inkonsistenz tritt auf, falls es eine Regel der Form $A \rightarrow A$ gibt.

Mit dieser neuen Art von Grammatik ergeben sich auch neue Algorithmen zum Berechnen der Bäume. Sowohl der CKY als auch der Earley Algorithmus sind um den Faktor der Wahrscheinlichkeit erweiterbar, wobei der CKY Algorithmus mehr Verwendung findet.

Der in Kapitel 2.2.3 vorgestellte CKY Algorithmus wird dahingehend erweitert, dass in jeder Zelle für jedes Nichtterminal die Wahrscheinlichkeit gespeichert wird. Damit erhält man eine dreidimensionale $(n + 1) \times (n + 1) \times V$ Matrix, wobei V die Anzahl an Nichtterminalen in der Grammatik ist.

Einen Nutzen kann man aus der zugefügten Wahrscheinlichkeit nur dann ziehen, wenn ihr numerischer Wert Sinn ergibt. Um diesen Wert für jede Regel zu berechnen gibt es zwei Möglichkeiten. Zum einen kann dieser aus einer vollständig annotierten Treebank errechnet werden. Hierfür wird jedes Auftreten einer Regel und des entsprechenden Nichtterminals gezählt und dividiert:

$$P(A \rightarrow \beta) = \frac{\text{Anzahl}(A \rightarrow \beta)}{\sum_{\gamma} \text{Anzahl}(A \rightarrow \gamma)} = \frac{\text{Anzahl}(A \rightarrow \beta)}{\text{Anzahl}(A)} \quad (2.6)$$

Falls man keine solche Treebank zur Verfügung hat, gibt es noch eine zweite Möglichkeit die Werte der PCFG festzulegen. Hierzu arbeitet der Parser einen unannotierter Textkorpus durch und versieht die Sätze mit Tags. Zu Beginn haben alle Produktionen eines Nichtterminals die selbe Wahrscheinlichkeit. Der Korpus wird iterativ durchlaufen und nach jedem Durchgang werden die Wahrscheinlichkeiten der Regeln angepasst. Das Anpassen passiert wie in der ersten vorgestellten Möglichkeit, da zu diesem Zeitpunkt eine annotierte Treebank vorhanden ist. Das Verfahren endet, wenn die Wahrscheinlichkeiten konvergieren.

Die resultierenden Werte in der Grammatik hängen also davon ab, mit welchem Korpus sie berechnet wurden. Hierbei spielen Größe und Textart eine große Rolle. Um beim Parsen eines Textes möglichst gute Ergebnisse zu erhalten, sollte der Parser eine PCFG verwenden, welche mit einem Text des selben Genres erstellt wurde. So kann man beispielsweise mit einer Treebank aus technischen Handbüchern, unabhängig von ihrer Größe, beim Parsen eines privaten Briefes keine guten Ergebnisse erwarten, da sich die Sprache zu sehr unterscheidet.

Außerdem weist dieses neue Konzept auch zwei Nachteile auf. Das erste Problem der PCFG ergibt sich aus der Kontextfreiheit. Die Wahrscheinlichkeit einer Produktion ist immer gleich, egal an welcher Stelle im Satz sie auftritt. In der tatsächlichen natürlichen Sprache ist das aber im Allgemeinen nicht der Fall. Beispielsweise kann die gewählte Produktion einer Nominalphrase abhängig davon sein ob die Phrase Objekt oder Subjekt des Satzes ist. Da diese Information ohne weiteres nicht in der Grammatik nicht berücksichtigt werden kann, muss der Mittelwert aus beiden Fällen gebildet werden. Dies führt dazu, dass entweder im Falle des Objekts oder des Subjekts häufig die falsche Regel angewendet wird.

Als zweite Schwachstelle ergibt sich, dass die einzelnen Wörter eine zu kleine Rolle spielen. Als Beispiel hierfür dient die bereits erklärte Anhangs-Mehrdeutigkeit aus Kapitel 2.2. Wiederum wird eine Präpositionalphrase betrachtet, welche entweder an eine Verbal- oder eine Nominalphrase angebunden wird. An welche von beiden hängt allein von der Treebank ab. Dort kommt entweder die Produktion

$$VP \rightarrow \alpha \ NP \ PP$$

oder die Kombination aus

$$VP \rightarrow \alpha \ NP$$

und

$$NP \rightarrow NP \ PP$$

öfter vor. α steht für eines der hier möglichen Verb-Nichtterminale, welches aber in beide Fällen immer das selbe ist und deswegen keine Rolle spielt. Wesentlich bessere Resultate sind hier erzielbar, wenn man das Verb aus *VP*, das Nomen aus *NP* und die Präposition aus *PP* in die Entscheidungsfindung mit einbezieht. Es kann aus der Treebank die Information gewonnen werden, ob die gegebene Präposition sich öfter auf das Nomen oder das Verb bezieht. Angenommen es gibt eine Präposition die ausschließlich mit Verben in Verbindung steht. In der Treebank sind es nun aber die Nominalphrasen, an welche öfter Präpositionalphrasen gebunden werden. Dann wird die eben angenommene Präposition mit einer PCFG, welche mit der Treebank errechnet wurde, immer falsch zugeordnet.

Diese Schwäche der PCFG macht sich ebenso bei Koordinations-Mehrdeutigkeit bemerkbar. Beim Verbinden von Phrasen durch Konjunktionen kann wieder die konkrete Konjunktion und die Beziehung zu den entsprechenden Wörtern der zu verbindenden Phrasen betrachtet werden. Betrachtet man nochmal das Beispiel aus Abschnitt 2.2: *old men and women*. Hier könnte eine Treebank die Information liefern, dass die Wörter *men* und *women* per *and* öfter direkt miteinander verbunden werden, als dass nur eines von beiden das Adjektiv *old* zugeordnet bekommt.

Eine mögliche Verbesserung der PCFG ergibt sich durch das Erweitern der Nichtterminale, um die Information wessen Kind es ist. Es wird also ein *NP* als NP^S geschrieben, wenn *S* der Elternknoten ist oder als NP^{VP} , falls es *VP* ist. Hierdurch ergeben sich zwei neue Regeln in der Grammatik und damit auch zwei neue Wahrscheinlichkeiten. Mit diesem Konzept kann dem Nichtterminal, obwohl die Kontextfreiheit im grammatikalischen Sinne nicht verletzt wird, ein Kontext gegeben werden. Allerdings wird, wenn jedes Nichtterminal für jeden möglichen Elternknoten eine neue Produktion erhält, die Grammatik enorm aufgeblasen. Nebeneffekt dieser neuen Größe ist, dass bei gleich bleibender Treebank für jede Regel weniger Trainingsdaten zur Verfügung stehen und damit die erhaltenen Wahrscheinlichkeitswerte ungenauer sind. Für einen gegebenen Korpus muss also ein Split-and-Merge Algorithmus ausgeführt werden, der errechnet wie weit eine Aufteilung der Nichtterminale sinnvoll ist.

2.3.2 Probabilistische Lexikalisierte Kontextfreie Grammatiken

Ein weiterer Ansatz, um bessere Parserergebnisse zu erhalten, ist das Lexikalisieren der Grammatik. Hierfür muss der Begriff des Kopfes, im Englischen Head, eingeführt werden. Die Idee ist, dass jede syntaktische Einheit einen Kopf, in Form eines Wortes aus dieser Einheit, besitzt. Es wird das Wort genommen, welches “im Satz am

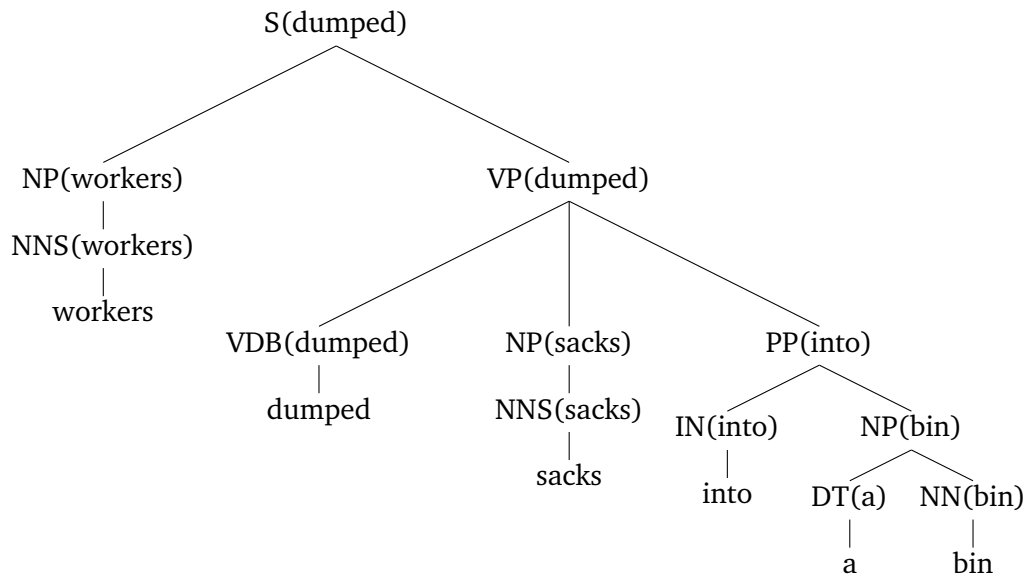


Fig. 2.4: Lexikalisierte Baum, entnommen aus

grammatikalisch wichtigsten ist". Da jedes Nichtterminal einer syntaktischen Einheit oder einem POS-Tag entspricht, kann jedem Nichtterminal und jeder Produktion der Grammatik ein Kopf zugewiesen werden. Für das Finden des richtigen Wortes gibt es verschiedene Schritte. Begonnen wird indem jedes POS-Nichtterminal sein entsprechendes Kind als Head wählt. Dieses Wort wird dann im Baum nach oben weitergeben. Für jede syntaktische Einheit gibt es Regeln, anhand derer einer der Köpfe der Kinder ausgewählt und als eigener eingesetzt wird. Ein Beispiel ist der lexikalisierte Baum für den Satz /textitworkers dumped sacks into a bin, dargestellt in Abbildung ??.

Zusätzlich zum Kopfwort kann auch noch dessen POS-Tag abgespeichert werden. Somit wird beispielsweise $S(\textit{dumped})$ zu $S(\textit{dumped}, VBD)$ und $NNS(\textit{workers})$ zu $NNS(\textit{workers}, NNS)$. Die Wahrscheinlichkeit der Regel

$$VP(\textit{dumped}, VBD) \rightarrow VBD(\textit{dumped}, VBD) \ NP(\textit{sacks}, NNS) \ PP(\textit{into}, IN) \quad (2.7)$$

ergibt sich wieder aus dem Inhalt der Treebank, nämlich durch die Formel

$$\frac{\text{Anzahl}(VP(\textit{dumped}, VBD) \rightarrow VBD(\textit{dumped}, VBD) \ NP(\textit{sacks}, NNS) \ PP(\textit{into}, IN))}{\text{Anzahl}(VP(\textit{dumped}, VBD))} \quad (2.8)$$

Dieser Wert ist 0, falls der Korpus keinen Satz mit *dumped sacks into* enthält. Existiert kein Satz in welchem sich $VP(\textit{dumped}, VBD)$ finden lässt, so ist dieser Wert nicht definiert.

Aufgrund dieses Problems bedarf es anderer Berechnungsvorschriften, wie zum Beispiel Collins Model 1. Es wird jede Produktion folgendermaßen betrachtet: H ist der Kopf, L_i sind die Nichtterminale links und R_i die rechts davon. Alle

Nichtterminale bleiben weiterhin lexikalisiert. Das Kopfwort wird mit h bezeichnet. Damit ergibt sich die Form

$$A \rightarrow L_n \dots L_1 H R_1 \dots R_m$$

Zusätzlich wird an der Stelle L_{n+1} und R_{m+1} das Nichtterminal *STOP* eingefügt um anzuzeigen, dass hiernach die Regel zu Ende ist. In drei Schritten wird die Wahrscheinlichkeit der Produktion errechnet:

1. Es wird der Kopf mit der Wahrscheinlichkeit $P_H(H|A, h)$ generiert.
2. Alle Elemente rechts vom Kopf werden mit $\prod_{i=1}^{m+1} P_R(R_i|A, h, H)$, also einschließlich dem *STOP*, generiert.
3. Alle Elemente links von H werden mit $\prod_{i=1}^{n+1} P_L(L_i|A, h, H)$ generiert.

Für die Regel 2.7 errechnet sich die Wahrscheinlichkeit über

$$\begin{aligned}
 P &= P_H(VBD|VP, dumped) \\
 &\times P_R(NP(sacks, NNS)|VP, dumped, VBD) \\
 &\times P_R(PP(into, IN)|VP, dumped, VBD) \\
 &\times P_R(STOP|VP, dumped, VBD) \\
 &\times P_L(STOP|VP, dumped, VBD)
 \end{aligned} \tag{2.9}$$

Im Gegensatz zu vorher muss also nicht die Kombination aus $NP(sacks, NNS)$, $PP(into, IN)$ und $VBD(dumped, VBD)$ vorhanden sein. Es genügt, wenn jedes einzeln, als Kind von $VP(dumped, VBD)$ auf der entsprechenden Seite des Heads in der Treebank zu finden ist. Darüber hinaus wird das Modell noch um die Information der Distanz erweitert. Es wird also zusätzlich berücksichtigt wie weit ein Nichtterminal vom Kopf der Regel entfernt ist. Der Vollständigkeit halber ist außerdem zu erwähnen, dass es neben Model 1 noch zwei weitere Modelle gibt. Diese Modelle bilden die Grundlage für den Collins Parser.

Konzept

In diesem Kapitel wird das Konzept nach dem implementiert wurde vorgestellt. In Abbildung ?? wird dargestellt, wie die Leistung eines Parsers evaluiert wird. Der Input-Text ist der rohe Text den der Parser mit syntaktischer Annotation versehen soll. Das Parser-Modell ist die Grundlage auf der ein Parser arbeitet. Der Goldstandard ist der Input-Text mit, als korrekt angesehener, Annotation. Dieser wird mit der Ausgabe des Parser verglichen. Daraus lassen sich Kennzahlen zur Leistung des Parsers errechnen, die am Ende ausgegeben werden. Um mehrere Parser miteinander zu vergleichen werden die Resultate der einzelnen in eine Tabelle zusammengefasst und dargestellt. Alle Elemente werden nachfolgend ausführlich vorgestellt.

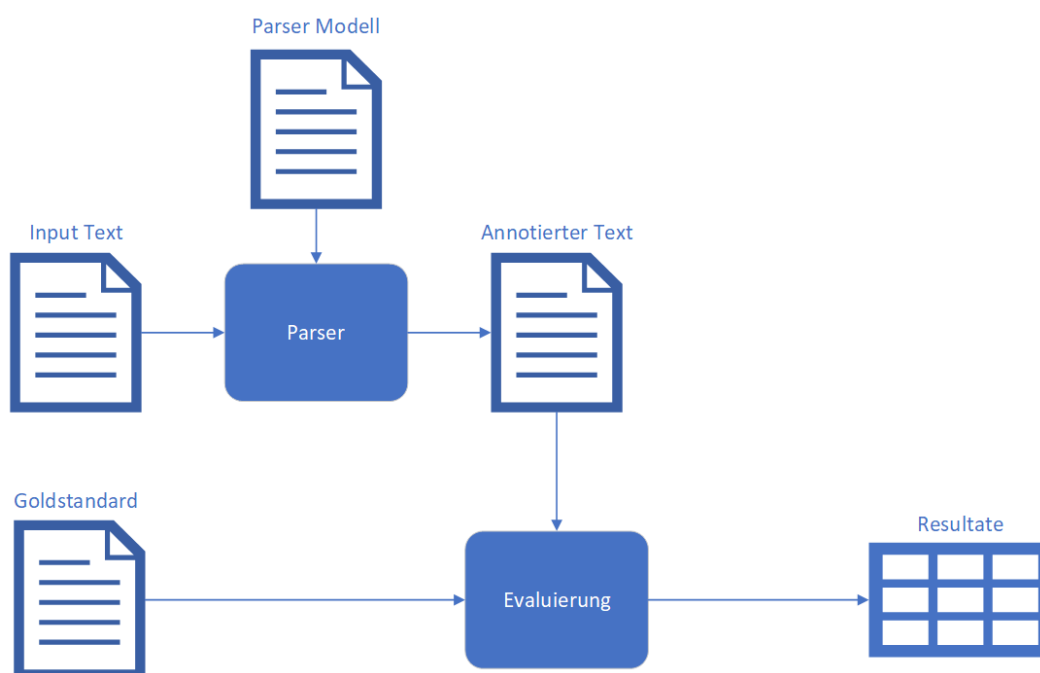


Fig. 3.1: Konzeptueller Aufbau

3.1 Input Text

Die Eingabe des Parsers sind Sätze, die nach Tokens getrennt sind. Das heißt, alles was ein POS-Tag bekommt wird mit Leerzeichen getrennt. Ein solcher Satz wird

sequenziert genannt. Die meisten Wörter brauchen keine weitere Verarbeitung, da sie bereits ein Leerzeichen zum nächsten Wort haben. Ein typisches Beispiel für die Notwendigkeit der Aufteilung ist *he's*. Hier erkennt der Parser nur dass es sich um die Wörter *he* und *is* handelt, wenn ein Leerzeichen vor dem Apostroph eingeschoben wird. Bekommt man aus einem Korpus keine Version des Satzes, bei dem dieser Arbeitsschritt schon erledigt ist, muss man einen Tokenizer vorschalten. Hierfür gibt es unterschiedliche Anbieter, wie z.B. Stanford oder OpenNLP. Allerdings wird in der Implementierung kein Tokenizer verwendet. Die Sätze der Eingabe müssen über ein Textdokument übergeben und zeilenweise getrennt sein. Der Parser erstellt für jede Zeile einen Baum.

3.2 Parser Output

Als Ausgabe gibt der Parser die annotierten Sätze in der Reihenfolge, in der sie eingegeben wurden, zurück. Es wird hierfür wieder ein Textdokument erstellt. Beispielsweise lautet die annotierte Version des Satzes

It goes 150 miles an hour .

```
( (S (NP (PRP It)) (VP (VBZ goes) (NP (NP (CD 150) (NNS miles))
(NP (DT an) (NN hour)))) (. .) )
```

Das äußerste Klammerpaar hat kein führendes Nichtterminal und könnte deshalb weggelassen werden. Es enthält typischerweise das Startsymbol der Parser, also *ROOT*, *TOP* oder ähnliches. Dieses Nichtterminal wird aber für die Evaluierung aus dem Ergebnisbaum herausgenommen, weil es keine syntaktische Information liefert. Da der Satz des Goldstandards ebenfalls diese unannotierten Klammern besitzt, werden sie aus beiden Bäumen nicht gelöscht sondern im Evaluationsschritt ignoriert.

3.3 Parser Modell

Ein Parser erzielt, je nachdem welches Modell ihm zu Grunde liegt, sehr unterschiedliche Ergebnisse. Daher ist dieses Modell nicht fest im Parser verankert, sondern wird ihm als Datei übergeben. Somit kann man selbst Modelle anhand einer Treebank erstellen lassen um dann beispielsweise auf einem Parser die unterschiedlichen Modelle vergleichen. Alle verwendeten Parser bringen ein trainiertes Modell mit.

3.4 Goldstandard

Als Goldstandard werden die korrekten Bäume bezeichnet. Diese wurden entweder von Menschenhand erstellt oder durch einen Parser produziert und dann von Menschen nachkorrigiert. Es muss für ein eingegebenes Textdokument, welches vom Parser bearbeitet werden soll auch ein Textdokument geben, das für jeden dieser Sätze die annotierte Version enthält. Die gratis verfügbaren Treebanks weisen oft ein unterschiedliches Sortiment an Dateien und Formaten auf.

Der, in dieser Arbeit verwendete, Korpus heißt “The NAIST-NTT TED Talk Treebank” und stammt aus dem Jahr 2014. Er umfasst etwa 23.000 Wörter in 1.200 Sätzen, welche aus 10 Reden gewonnen wurden. Zur Annotierung wurde der Berkeley Parser verwendet und dessen Resultat per Hand verbessert. Als Tag-Set wurde wieder die Vorlage der Penn Treebank verwendet. In dieser Treebank sind für jeden Satz eine Rohversion, eine sequenzierte und eine annotierte Version enthalten. Da es sich bei den Texten um Gesprochenes handelt, sind noch weitere Informationen, wie zum Beispiel Zeitmessung eines Satzes und ähnliches vorhanden. Das wird hier nicht weiter verwendet.

Bei Verwendung anderer Goldstandards können diverse Probleme auftreten. Zwei davon sind hier kurz durchdacht aber nicht implementiert worden.

Falls das Problem auftritt, dass nur die korrekten Lösungsbäume zur Eingabe verfügbar sind, so kann man sich z.B. mit NLTK die Blätter jedes Lösungsbaums ausgeben lassen. Hierdurch erhält man, den in seine Token aufgeteilten Satz.

Ist der Lösungsbaum nicht einzeilig, sondern erstreckt sich ähnlich zu 2.1 über mehrere Zeilen, so muss hier auch eine Vorverarbeitung geschehen. Entweder kann anhand der Klammerung erkannt werden wo die Grenzen des Baumes liegen oder es müssen alle Zeilen, die mit einer Art Leerzeichen (Tabulator, u.ä.) beginnen, an die vorherige gehangen werden.

3.5 Parser

Alle Parser erhalten die Eingabe in selber Form, allerdings kann jeder Parser noch eine individuelle Vorverarbeitung benötigen. Dazu mehr in Abschnitt 4.3. Im Rahmen dieser Arbeit sind drei Parser verglichen worden. Das sind der Stanford-Parser, der Berkeley-Parser und der Parser aus dem OpenNLP Paket. Jeder verwendet zum Annotieren die Penn Treebank Tags und liefert die Ausgabe standardmäßig ohne die zusätzlichen relationalen Tags.

3.5.1 Stanford-Parser

Der Stanford-Parser bringt ein Modell namens *englishPCFG* mit. Es handelt sich dabei um eine unlexikalisierte PCFG für die englische Sprache.

3.5.2 Berkeley-Parser

3.5.3 OpenNLP-Parser

3.6 Evaluierung

Für die Evaluierung wird das Ergebnis der Parser zeilenweise mit dem Goldstandard verglichen. Zu diesem Zweck werden die Konstituenten der Bäume betrachtet. Eine Konstituente beschreibt einen Knoten im Baum und enthält die Information über das Nichtterminal, den Start- und den Endpunkt. Die Punkte geben den Index des ersten und letzten Wortes dieses Teilbaums an. Mit jedem Wort wird der Zähler inkrementiert. Für den Beispielsatz *They learn much faster.* (Baum in Abbildung 3.2) sind die Konstituenten in Tabelle 3.1 dargestellt.

Falls es, für eine Konstituente des Parsers, im Goldstandard eine mit gleichem

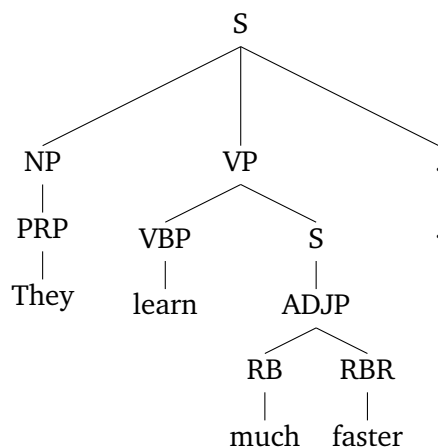


Fig. 3.2: Syntaxbaum zu *They learn much faster.*

Nichtterminal und gleicher Spanne gibt, wird diese als richtig bewertet. Hier muss bei der Ausführung entschieden werden, welche Nichtterminale gewertet werden. Zum einen kann man alle in das Resultat einbeziehen. Zum anderen können die POS-Tags herausgelassen und nur die syntaktischen Tags aus Tabelle 2.2 bewertet werden. Der Grund ist, dass man so Parser einbringen kann, die als Eingabe einen mit POS-Tags versehenen Text bekommen. Diese Tags verfälschen dann die Korrektheitsrate der Parser und müssen ignoriert werden. Da in dieser Arbeit ausschließlich Parser genutzt werden, die als Eingabe sequenzierten Text bekommen, betrachten

Nichtterminal	Text	Start	Ende
S	They learn much faster .	0	5
NP	They	0	1
PRP	They	0	1
VP	learn much faster	1	4
VBD	learn	1	2
S	much faster	2	4
ADJP	much faster	2	4
RB	much	2	3
RBR	faster	3	4
.	.	4	5

Tab. 3.1: Konstituenten zu Abbildung 3.2

wir hier den Fall in dem alle Tags relevant sind. Der Evaluierungsmechanismus ist angelehnt an ... Als Bewertungskriterien werden folgende vier Werte errechnet: *Precision*, *Recall*, F_1 und *Relative Cross Brackets* (kurz: RCB). Siehe Formeln (3.1) bis (3.4).

$$Precision = \frac{\# \text{ korrekte Konstituenten}}{\# \text{ Konstituenten im Parseroutput}} \quad (3.1)$$

$$Recall = \frac{\# \text{ korrekte Konstituenten}}{\# \text{ Konstituenten im Goldstandard}} \quad (3.2)$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.3)$$

$$RCB = \frac{\# \text{ Parser Konst. die Goldstandard Konst. kreuzen}}{\# \text{ Konst. im Parseroutput}} \quad (3.4)$$

Die *Precision*, im Deutschen *Genauigkeit*, gibt an, wieviele der Konstituenten des Parsers korrekt sind. *Recall*, oder *Trefferquote*, beschreibt zu welchem Teil die Konstituenten des Goldstandards mit den Parser-Konstituenten abgedeckt wurden. F_α ist das gewichtete harmonische Mittel aus beiden. Im Rahmen dieser Arbeit wurde nur F_1 betrachtet, d.h. beide werden gleich gewichtet. Beim RCB-Wert handelt es sich um einen Indikator der ausschließlich für Syntaxbäume eingesetzt wird. Er gibt an, wieviele der Konstituenten des Parsers sich mit denen des Standards kreuzen. Kreuzen ist in diesem Kontext folgendermaßen definiert. Seien $k1$ und $k2$ zwei Konstituenten mit unterschiedlichen Start- und Endpunkten. Außerdem gilt o.B.d.A., dass $k1$ den niedrigeren Startpunkt hat. Dann kreuzen sie, falls der Endpunkt von $k1$ größer als der Startpunkt und kleiner als der Endpunkt von $k2$ ist. Anders ausgedrückt, sie sind nicht ineinander enthalten aber teilen sich mindestens

ein Wort. Aus der Definition ergibt sich, für $k1$ aus dem Parser und $k2$ aus dem Goldstandard, dass $k1$ nicht korrekt sein kann. Angenommen es wäre korrekt, so müsste es ein $k1'$ im Goldstandard mit identischen Grenzen geben. $k1'$ beginnt vor $k2$ und hört auch vor diesem auf und kann somit kein Eltern- oder Kindknoten von diesem sein. Da sie sich aber mindestens ein Wort teilen würde für dieses Wort gelten, dass es Blatt von zwei verschiedenen Teilbäumen wäre, also zwei Eltern hätte. Hier ergibt sich also ein Widerspruch. Aus den Formeln folgt damit außerdem: $RCB \leq 1 - Precision$. Für alle vier Kennzahlen liegt der Wertebereich zwischen 0 und 1. *Precision*, *Recall* und F_1 sollten möglichst hoch sein und *Relative Cross Brackets* möglichst niedrig.

Implementierung

Dieses Kapitel beschreibt, wie das Konzept aus Kapitel 3 umgesetzt wurde. Es wird sowohl auf die selbst implementierten Operatoren, als auch auf die von mir verwendeten RapidMiner-Operatoren eingegangen.

4.1 RapidMiner Studio

Die Plattform für das Parser-Framework wurde auf RapidMiner Studio festgelegt. Hierbei handelt es sich um eine Data Science Plattform, in der Operatoren im Steckkasten-Prinzip miteinander verbunden werden. Zwischen diesen Operatoren werden Daten in unterschiedlicher Form übergeben. Die Operatoren selbst führen verschiedene Aufgaben aus. Das eigentliche RapidMiner Studio bietet hauptsächlich Operatoren für die Verarbeitung von Tabellen an. Allerdings sind über den RapidMiner Marketplace auch eine Vielzahl von Erweiterungen erhältlich. Im Rahmen dieser Arbeit wird die Text Processing Extension verwendet. RapidMiner bietet zusätzlich die Möglichkeit eigene Operatoren zu programmieren, was im Rahmen dieser Arbeit auch genutzt wurde. Das Studio, die Erweiterungen und die eigenen Operatoren wurden in Java geschrieben.

4.2 Verwendete Operatoren

Hier werden die bereits bestehenden Operatoren, die ich im Prozess verwendet habe, beschrieben.

Zum Einlesen des Parser-Modells wurde der Operator **Read File** aus dem Ordner *Utility\Files* verwendet. Dieser hat als Parameter den Pfad, für die zu öffnende Datei. Er bietet einen Output-Port, über den die geöffnete Datei später als `com.rapidminer.operator.nio.file.SimpleFileObject` verarbeitet werden kann. Das Dateiformat der Modelle ist *.bin* für den OpenNLP-Parser, *.gr* für den Berkeley-Parser und *.ser.gz* für den Stanford-Parser. Wichtig ist hier aber nicht die geöffnete Datei, da alle Parser diese selbst nochmal öffnen. Die Parser bekommen

als Übergabe den Pfad und dieser kann aus dem `SimpleFileObject` entnommen werden.

Aus der Text-Processing-Extension wird der Operator **Read Document** verwendet. Er wird an zwei Stellen verwendet. Zum einen wird über ihn der Input-Text für den Parser eingelesen. Zum anderen wird der Goldstandard für die Evaluierung damit in den Prozess gebracht. In beiden Fällen wird nur der Parameter *file* auf den entsprechenden Dateipfad gesetzt und alle anderen Parameter auf ihrem Default-Wert belassen. Auch sein Input-Port wird nicht verwendet. Über den Output-Port wird die geöffnete Datei als `com.rapidminer.operator.text.Document` an die nachfolgenden Operatoren übergeben.

Der letzte verwendete Operator ist **Execute Script** aus dem Ordner *Utility\Scripting*. Dessen Aufgabe ist das Ausführen von Java Code. Verwendung findet er lediglich in Kombination mit dem Stanford-Parser. Die verwendeten Packages von Stanford werden hier einmal initialisiert, da es sonst einen Fehler seitens RapidMiner gibt. Hierfür wird ein Objekt der Klasse `edu.stanford.nlp.trees.TreebankLanguagePack` erstellt und wieder auf `null` gesetzt. Weder die Input- noch die Output Ports des Operators werden genutzt, wodurch er vor den anderen Operatoren ausgeführt wird, und die Stanford-Packages korrekt initialisiert werden. Um diese Strategie der Fehlerumgehung nutzen zu können sind zwei zusätzliche Schritte notwendig. Zum einen muss die Datei *stanford-parser.jar*, welche die entsprechenden Packages beinhaltet, in den lib Ordner von RapidMiner Studio gelegt werden. Diesen findet man typischerweise unter *Programme\RapidMiner\RapidMiner Studio\lib*. Das zweite und schwerwiegendere Problem ist der Bedarf einer Large License von RapidMiner. Für die Dauer dieser Arbeit wurde mir diese von RapidMiner kostenlos zur Verfügung gestellt. Der reguläre Preis beträgt allerdings .

4.3 Eigene Operatoren

Hier werden die Operatoren und alle dazu gehörigen Klassen, die für dieses Parser-Framework entwickelt wurden, vorgestellt.

Zur Entwicklung einer eigenen Erweiterung bietet RapidMiner einen Guide an. Als Entwicklungsumgebung wurde Eclipse (Version: Oxygen.3a Release (4.7.3a)) in Kombination mit dem Gradle Plugin gewählt.

Alle entwickelten Operatoren erben von der Klasse `com.rapidminer.operator.Operator` und die Methode `doWork()` wird aufgerufen, wenn der Operator im Workflow durchlaufen wird.

Die Übergabe von Daten zwischen Operatoren erfolgt über Klassen, die das Interface `com.rapidminer.operator.IObject` implementieren, wie z.B. `ExampleSet` oder `Document`.

Im Rahmen dieser Arbeit wurden folgende Operatoren erstellt: **Berkeley Parser**, **OpenNLP Parser**, **Stanford Parser**, **Compare Results** und **Show Results**

Alle Parser-Operatoren haben die gleichen Ports zur Ein- und Ausgabe.

ioobjectInputGrammar Über diesen Port wird das Modell des Parser eingelesen. Dieses wird in `doWork()` auf ein `SimpleFileObject` gecastet, um den Dateipfad auslesen zu können. Er wird mit dem Read-File-Operator verbunden.

ioobjectInputText Hier liest der Operator den Input-Text für den Parser ein. Das `IObject` wird auf ein `Document` gecastet, um den Inhalt als String zu bekommen. Dieser String wird dann nach Zeilen getrennt, da jeder Parser immer eine Zeile eingegeben bekommt. Er wird mit dem Read-Document-Operator verbunden.

nameOutputExt Es wurde ein `PortExtender` verwendet. Das heißt, falls man diesen Port verbindet wird der selbe Port nochmals erstellt. Somit kann man die Parser Ausgabe öfters verwerten. Dies kann man z.B. nutzen, um in der Evaluation unterschiedliche Parameter zu setzen. Extender müssen im Konstruktor mit der Methode `start()` aktiviert werden. Über die, von diesem Extender erstellten, Ports wird der Name des Parsers in einem `Document` nach außen gegeben. Hierfür gibt es einen Konstruktor der Klasse `Document` dessen Übergabeparameter ein String ist.

resultOutputExt Dieser Extender gibt den geparsen Text in einem `Document` aus. Hier sind alle einzeln verarbeiteten Zeilen wieder zu einem String zusammengefügt worden.

4.3.1 Berkeley-Parser

Der Code für diesen Parser ist aus der Klasse `edu.berkeley.nlp.PCFG.LA.BerkeleyParser` entnommen. Diese Klasse wird beim Kommandozeilenaufwurf des Parsers angesprochen. Die `main`-Methode sowie `outputTrees(...)` wurden leicht abgewandelt übernommen.

Die Funktionalität der `main` wurde in die `doWork()` geschrieben. Über den eigentlichen Kommandozeilenaufwurf können Optionen übergeben werden, die in ein Objekt

der Klasse `edu.berkeley.nlp.PCFG.LA.BerkeleyParser.Options` geparkt werden. Hier wird stattdessen ein Objekt dieser Klasse erstellt und die Optionen manuell gesetzt. Der Wert des Attributs `grFileName` wird auf den Dateipfad des eingelesenen Modells gesetzt. Bei allen anderen werden die Default-Werte beibehalten.

Die Ein- und Ausgabe in der Originalklasse war ein `BufferedReader` und ein `PrintWriter`. Das wurde abgewandelt, da die Eingabe jetzt zeilenweise als `String` zur Verfügung steht und die Ausgabe auch in einen `String` geschrieben werden soll. Der Parser bekommt seine Eingabe jetzt über eine `for`-Schleife, in der über das Array der Eingabesätze iteriert wird. Zuvor wurden per `while`-Schleife und die `BufferedReader` Methode `readLine()` der Parser-Input erhalten.

Die Methode `outputTrees(...)` wurde als Methode des Operators aufgenommen. Hierzu wurde der Rückgabe Typ von `void` auf `String` geändert und der Übergabeparameter zur Datenausgabe von `PrintWriter` auf `String` abgeändert. Das führt dazu, dass statt `outputData.write("...")` nun `outputText += "..."` verwendet wird. Außerdem wird die Möglichkeit, die Ausgabe als `.png` zu erhalten, entfernt. Zuletzt wird der Methode ein `return outputText` angehängt. In der `doWork()` wird dieser `String` als `Document` verpackt und über den `PortExtender` an die Output-Ports geliefert.

4.3.2 OpenNLP-Parser

OpenNLP bietet ein Manual zum Integrieren des Parsers in eine Anwendung mit Hilfe seiner API. Der Code aus diesem Manual wurde in die `doWork()` des Operator geschrieben. Dieser Code wurde von einem zusätzlichen `try-catch`-Block umgeben, um eine `FileNotFoundException` zu fangen. Diese Exception wird beim Öffnen der Datei des Parser-Modells geworfen. Per `for`-Schleife werden die Sätze der Eingabe in den Parser geschickt.

Über die `String` Methode `outputText.replaceAll("\\(TOP ", "(");` wird das Startsymbol des Parser entfernt. Um zu verhindern, dass das Auftreten des Wortes "TOP" im Eingabesatz gelöscht wird, wurde die Klammer mit in den regulären Ausdruck aufgenommen. Worte des Satzes haben in der Formatierung immer eine schließende Klammer nach sich, aber nie eine öffnende vorher.

4.3.3 Stanford-Parser

Beim Herunterladen des Parser wird eine Demo-Klasse mitgeliefert, in welcher der Code zum Starten des Parsers aufgeführt wird. Hieraus wurden die notwendigen Befehle in die `doWork()` des Operators übernommen. Zu beachten ist hier noch,

dass der Parser den Satz nicht als `String` entgegen nimmt, sondern die Token des Satzes in einem `String-Array`.

Das Startsymbol dieses Parser lautet "ROOT" und wird analog zum OpenNLP-Parser aus 4.3.2 entfernt.

4.3.4 Compare Results

In diesem Operator findet die Evaluierung statt. Es wird die Ausgabe von genau einem Parser mit dem Goldstandard verglichen und die Kennzahlen errechnet.

Hierzu gibt es drei Input-Ports, deren `IObject` jeweils zu einem `Document` gecastet wird. Der erste Port erhält den Namen des Parsers, der zweite die Ausgabe der Parser und der dritte den Goldstandard. Das Ergebnis wird in einem einzeiligen `com.rapidminer.example.ExampleSet` ausgegeben. Dadurch kann man später mehrere Evaluierungen in einer Tabelle untereinander anzeigen lassen.

Der Operator hat zwei `Boolean-Parameter`. `PARAMETER_REMOVE_SUFFIX` gibt an, ob den syntaktischen Tags die Suffixe entfernt werden sollen. Das heißt, dass NP-SBJ als NP gewertet wird. `PARAMETER_COUNT_ONLY_SYNTACTIC_TAGS` entscheidet ob alle Tags oder nur die syntaktischen in die Bewertung zählen. Die Parameter wurden entsprechend dem Extension Manual erstellt. Default-Wert bezüglich der Suffixe ist `true` und bezüglich der syntaktischen Tags `false`.

Für die Evaluierung benötigt dieser Operator zwei zusätzliche Hilfsstrukturen. Zum einen das Enum **PennTag**. Hier sind alle Tags der Penn Treebank aufgelistet. Jedes Literal hat einen `String` als Attribut, der die Schreibweise angibt. Somit kann über die statische Methode `stringToPennTag(String s, boolean removeSuffix)` ein `String` auf ein Enum-Literal abgebildet werden. Hierzu werden alle Tags durchlaufen und deren Attribut mit dem Übergabestring verglichen. Das entsprechende Literal wird zurückgegeben. Für den Fall, dass dem `String` kein Tag entspricht, wird das zusätzlich eingeführte Literal `Empty`, mit leerem `String` als Attribut, zurückgegeben. Das passiert beispielsweise dann, wenn im Parser oder Goldstandard zusätzliche Tags verwendet werden. `removeSuffix` sorgt dafür, dass die Tags ab dem ersten Bindestrich abgeschnitten werden und somit nur das ursprüngliche betrachtet wird. Zum anderen wird die Klasse **ParseTreeNode** verwendet. Ein Objekt dieser Klasse entspricht einer Konstituente aus Kapitel 3.6. Realisiert wurde dieses Konzept über eine `PennTag` und zwei `int` Membervariablen für den Typ und die Grenzen. Über die Methode `equals(Object obj)` wird getestet, ob sowohl Tag, als auch Grenzen des übergebenen `ParseTreeNodes` mit dem aufrufendem übereinstimmen.

Der Operator selbst bedient sich wiederum zweier Hilfsmethoden. Mit

```
private static String formatSentence(String s)
```

wird jede Zeile so formatiert, dass vor und hinter jeder Klammer ein Leerzeichen ist. Außerdem gibts es keine aufeinanderfolgenden Leerzeichen. Dadurch kann der String nach Leerzeichen gesplittet werden. Durch

```
private static List<ParseTreeNode> parseSentence(String s, boolean removeSuffix,  
boolean countOnlySyntacticTags)
```

wird eine formatierte Zeile *s* in eine Liste von Konstituenten umgewandelt. Hierfür wird ein `Stack<ParseTreeNode>` genutzt. Der Satz wird nach Leerzeichen in Klammern, Nichtterminale und Terminale gesplittet. Diese werden mit einer `for`-Schleife durchlaufen. Damit ergeben sich vier Fälle.

- Fall 1: Öffnende Klammer. Es wird ein neuer `ParseTreeNode` erstellt und dessen Startpunkt auf die aktuelle Wortposition im Satz gesetzt. Diese Konstituente wird auf den Stack gelegt.
- Fall 2: Nichtterminal. Die oberste Konstituente des Stacks wird geholt, ihr Typ auf das gelesene Nichtterminal gesetzt und sie selbst wieder auf den Stack zurückgelegt.
- Fall 3: Terminal. Der Wortzähler wird inkrementiert. Das eigentliche Terminal wird nirgends abgespeichert, da diese in der Parser-Ausgabe und im Goldstandard identisch sind.
- Fall 4: Schließende Klammer. Der oberste `ParseTreeNode` wird vom Stack geholt und sein Endpunkt auf den aktuellen Stand des Wortzählers gesetzt. Gilt `countOnlySyntacticTags == true`, so werden nur diejenigen `ParseTreeNodes` an die Ausgabe gehängt, deren `PennTag` zu den syntaktischen Tags gehört. Falls nicht, so wird jede Konstituente der Ausgabeliste hinzugefügt.

Diese Methode gibt also alle zu bewertenden Konstituenten des Satzes zurück.

In der `doWork()` des Operators werden sowohl Parser-Ausgabe als auch Goldstandard in nach Zeilen aufgeteilt. Für jede Zeile der beiden Versionen werden die beschriebenen Hilfsmethoden aufgerufen. Die Parser-Konstituenten werden durchlaufen und im Goldstandard nach passendem Gegenstück gesucht. Hierdurch bekommt man die Anzahl an korrekten Konstituenten. Für den *RCB*-Wert werden nochmals die `ParseTreeNodes` des Parsers durchlaufen und alle gezählt, die mindestens eine

Konstituente des Goldstandards kreuzen. Diese beiden Zahlen, sowie die Anzahl an Parser- und Goldstandard-Konstituenten werden für jede Zeile aufsummiert. Dadurch können, nach dem Vergleich aller Zeilen, entsprechend der Formeln (3.1) bis (3.4), die Kennzahlen für den Parser errechnet werden.

Zuletzt wird ein `ExampleSet` mit den Attributen `Name`, `Precision`, `Recall`, `F1`, `Crossing Brackets`, sowie der totalen Anzahlen der Konstituenten, aus denen die Kennzahlen berechnet wurden, erstellt. Diese eine Zeile, mit den Werten des Parsers, wird in dieses Set eingefügt, welches dann an den `OutputPort` übergeben wird.

4.3.5 Show Results

Dieser Operator dient lediglich dem Zusammenführen von mehreren Parser-Evaluationen.

Es wird für die Input-Ports wieder ein `PortExtender` verwendet, um beliebig viele `Compare-Results-Operatoren` anschließen zu können. Der Output-Port liefert eine Tabelle, in der pro Zeile die Ergebnisse eines Parsers stehen.

In der `doWork()` werden alle Input-Ports durchlaufen und die Werte in die Ergebnistabelle übertragen. Diese wird zuletzt an den Output-Port geliefert.

Evaluierung

In diesem Kapitel wird die Leistung der Parser anhand der Ted-Talks-Treebank getestet.

Dieser Korpus enthält zehn Texte unterschiedlicher Länge, die alle von jedem Parser bearbeitet werden. In Tabelle 5.1 sind die Ergebnisse aufgeführt. Für die Messung wurden alle Tags gewertet, nicht nur die syntaktischen. In einer Zeile stehen die Werte für die entsprechende Datei. Für den totalen Wert wurde die Anzahl der unterschiedlichen Konstituenten über alle zehn Dateien aufsummiert und anhand dieser Zahlen nach entsprechender Formel berechnet. Somit wird die unterschiedliche Größe der Texte berücksichtigt.

Der Berkeley-Parser nimmt mit Precision- und Recall-Resultaten von über 93% den ersten Platz ein. Der OpenNLP-Parser liegt, mit weniger als 1% Vorsprung in beiden Werten, vor dem Stanford-Parser. Beide weisen einen gleich hohen Anteil an kreuzenden Konstituenten auf.

Für die Interpretation dieser Ergebnisse muss beachtet werden, dass die Modelle nicht mit den selben Texten trainiert wurden. Es handelt sich um die vortrainierten Modelle der Parser-Anbieter. Um aussagekräftigere Ergebnisse zu erhalten, müssen eigene Modelle erstellt werden.

Dateiname	Berkeley Parser				Stanford Parser				OpenNLP Parser			
	Precision	Recall	F_1	RCB	Precision	Recall	F_1	RCB	Precision	Recall	F_1	RCB
SheatHembrey_2011	93,3%	93,2%	93,2%	2,7%	89,4%	89,2%	89,3%	4,4%	89,9%	89,3%	89,6%	3,6%
RobertLang_2008	93,8%	93,4%	93,6%	2,4%	89,4%	89,1%	89,3%	3,5%	89,2%	89,1%	89,2%	3,3%
JessaGamble_2010G	94,3%	94,4%	94,4%	2,2%	88,3%	89,6%	88,9%	3,8%	90,3%	90,3%	90,3%	2,9%
MihalyCsikszentmihaly_2004	93,8%	93,0%	93,4%	3,8%	87,8%	85,2%	86,4%	6,0%	89,6%	89,1%	89,3%	5,5%
YvesBahar_2009	91,9%	92,0%	91,9%	1,4%	88,2%	88,8%	88,5%	3,7%	89,4%	89,6%	89,5%	1,4%
KatherineFulton_2007	92,8%	92,3%	92,5%	3,5%	85,7%	83,4%	84,5%	4,8%	87,2%	87,3%	87,2%	6,0%
HannaRosin_2010W	93,1%	92,8%	92,9%	4,1%	90,0%	89,3%	89,7%	4,2%	89,3%	88,5%	88,9%	4,8%
HansRosling_2010S1	92,4%	92,6%	92,5%	2,9%	88,9%	89,6%	89,2%	3,6%	87,4%	87,2%	87,3%	4,4%
StefanaBroadbent_2009G	92,8%	92,1%	92,5%	3,8%	88,5%	88,0%	88,2%	5,6%	87,9%	87,8%	87,8%	5,5%
AnthonyAtala_2009P	94,5%	94,3%	94,4%	2,0%	89,0%	87,9%	88,4%	3,9%	90,3%	90,4%	90,3%	3,4%
Total	93,4%	93,1%	93,3%	3,0%	88,7%	87,9%	88,3%	4,4%	89,1%	88,8%	89,0%	4,4%

Tab. 5.1: Evaluation der Parser für die Ted-Talks-Treebank

Schluss

Zuletzt wird ein Fazit gezogen und ein Ausblick gegeben.

List of Figures

2.1	Syntaxbaum zum Satz <i>My dog also likes eating sausage.</i>	6
2.2	Beispielmatrix für den CKY-Algorithmus, entnommen aus	10
2.3	Grammatik für entnommen aus	11
2.4	Lexikalisierte Baum, entnommen aus	15
3.1	Konzeptueller Aufbau	17
3.2	Syntaxbaum zu <i>They learn much faster.</i>	20

List of Tables

2.1	Penn Treebank POS Tags	4
2.2	Penn Treebank Syntaktische Tags	5
3.1	Konstituenten zu Abbildung 3.2	21
5.1	Evaluation der Parser für die Ted-Talks-Treebank	32

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und dabei keine anderen als die angegebenen Hilfsmittel benutzt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde bisher weder gesamt noch in Teilen einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Bayreuth, Juni 14, 2019

Klaus Freiberger

