Constituency Parsing

Chapter 13 - Speech and Language Processing. Daniel Jurafsky James H. Martin

Fegyó Kinga

2020. március 13.

Outline

1. Konstituens parszolás
2. Többértelműség problematikája
3. CKY parszolás
4. Sekély parszolás

Konstituens parszolás

Konstituens parszolás

Szintaktikai parszolás: felismerünk egy mondatot, és szintaktikai szerkezetet rendelünk hozzá

Környezetfüggetlen nyelvtanok nem mondják meg, ***hogyan***, erre kell találnunk a célnak megfelelő algoritmust.

Hasznosítása: helyesírás-ellenőrzés, szemantikai elemzés révén információkinyerés

What books were written by British women authors before 1800?

Többértelműség problematikája

Többértelműség problematikája

Szerkezei többértelműség:

Grammar	Lexicon
$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid the \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

Figure 13.1 The $\mathscr{L}_1$ miniature English grammar and lexicon.

Többértelműség problematikája

Szerkezeti többértelműség:

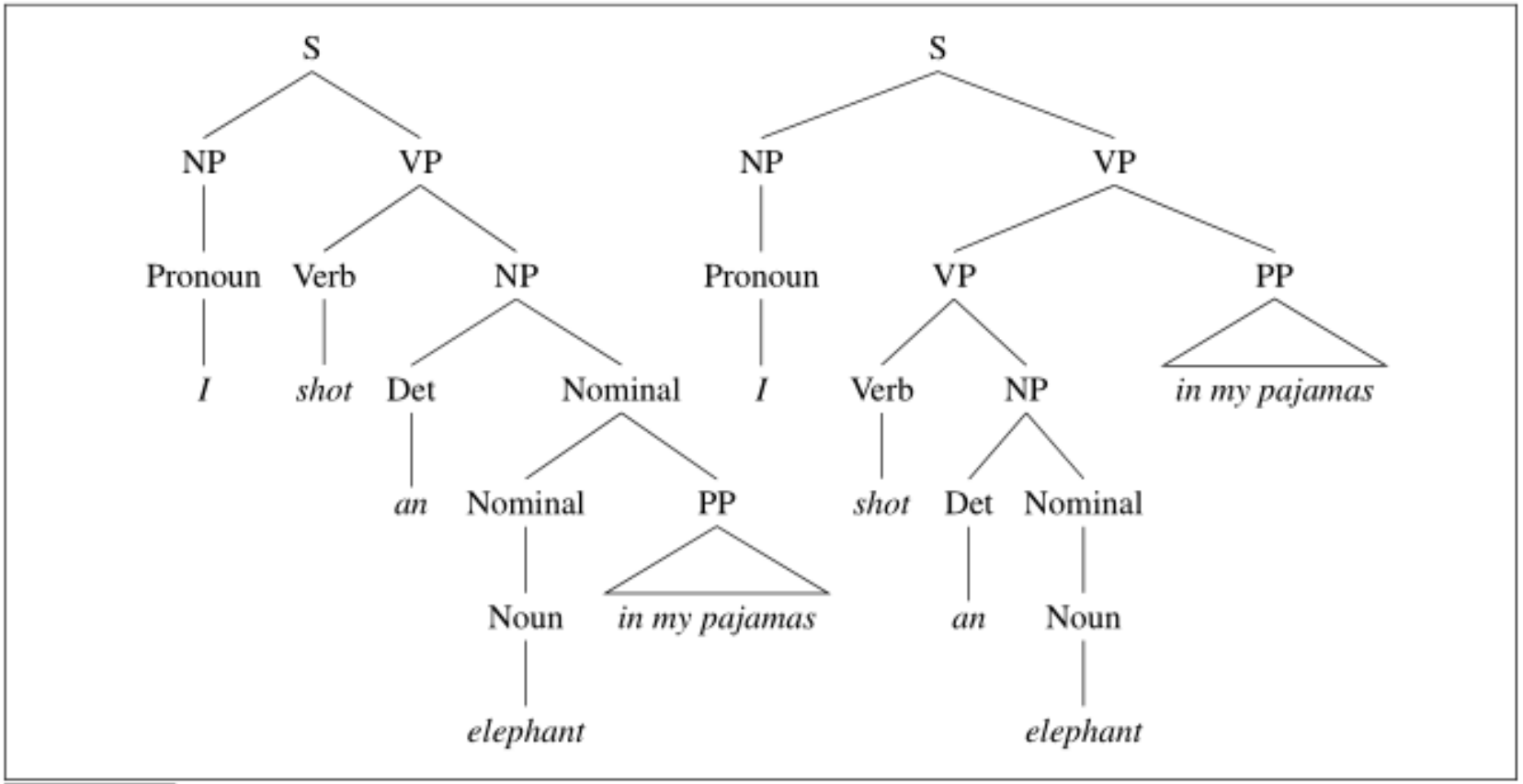

Figure 13.2 Two parse trees for an ambiguous sentence. The parse on the left corresponds to the humorous reading in which the elephant is in the pajamas, the parse on the right corresponds to the reading in which Captain Spaulding did the shooting in his pajamas.

Többértelműség problematikája

- csatolási többértelműség (*attachment ambiguity*)

We saw the Eiffel Tower flying to Paris.

- koordinációs többértelműség (*coordination ambiguity*)

old men and women

[old [men and women]] vs. *[old men] and [women]*

Természetesen előforduló mondatoknak gyakran sok grammatikus, de szemantikailag ésszerűtlen parszolása lehetséges →

szintaktikai egyértelműsítés

CKY parszolás

CKY parszolás

Dinamikus programozás: egy konténerbe szisztematikusan feltöltjük a részfeladataink megoldásait, ennek segítségével tudjuk megoldani a végén a teljes feladatot

Egy konstituens felismerésekor rögzítjük a jelenlétét és elérhetővé tesszük a további műveletek számára → idő- és helytakarékos

→ **Cocke-Kasami-Younger (CKY) algoritmus**

Chomsky-féle Normál Alak (CNF)

CKY algoritmus előfeltétele a CNF nyelvtan:

$$A \rightarrow B\,C \qquad \text{vagy} \qquad A \rightarrow w$$

Problémás esetek:

1. outputban terminális és nem-terminális
2. outputban egyetlen nem-terminális
3. outputban több mint 2 szimbólum

1. Outputban terminális és nem-terminális

Be kell vezetni egy *dummy* nem-terminálist, ami lefedi az eredeti terminálist, vagyis az $\mathit{INF\text{-}VP} \rightarrow \mathit{to\ VP}$ szabályt a következő szabályokra lehet bontani:

$$\mathit{INF\text{-}VP} \rightarrow \mathit{TO\ VP}$$
$$\mathit{TO} \rightarrow \mathit{to}$$

2. Outputban egyetlen nem-terminális

Szabály, amely outputja egyetlen nem-terminális = **unit production**

Megoldás: újraírni az eredeti szabály outputját az összes non-unit production outputjával, ahova az eredeti szabály vezethet

Ha adott nyelvtanban $A \to B$ egy vagy több unit-production eredménye, és létezik $B \to \gamma$, akkor bevezetjük az $A \to \gamma$ szabályt kihagyva a köztes lépéseket.

3. Outputban több mint 2 szimbólum

Megoldás: új köztes terminálisok bevezetése, amik rövidebb szekvenciákra osztják az eredeti szabályt

Tehát $A \to B\ C\ \gamma$ esetén:

1. $A \to X1\ \gamma$
2. $X1 \to B\ C$

3. Outputban több mint 2 szimbólum

$\mathscr{L}_1$ Grammar	$\mathscr{L}_1$ in CNF
$S \rightarrow NP\ VP$	$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$	$S \rightarrow X1\ VP$
	$X1 \rightarrow Aux\ NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb\ NP$
	$S \rightarrow X2\ PP$
	$S \rightarrow Verb\ PP$
	$S \rightarrow VP\ PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det\ Nominal$	$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal\ Noun$	$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$	$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb\ NP$	$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$	$VP \rightarrow X2\ PP$
	$X2 \rightarrow Verb\ NP$
$VP \rightarrow Verb\ PP$	$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$	$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$	$PP \rightarrow Preposition\ NP$

Figure 13.3 $\mathscr{L}_1$ Grammar and its conversion to CNF. Note that although they aren't shown here, all the original lexical entries from $\mathscr{L}_1$ carry over unchanged as well.

CKY felismerés

CNF nyelvtan garantálja, hogy minden nem-terminális nódusnak pontosan 2 gyereke van, így az egész fa szerkezetét dekódolhatjuk egy kétdimenziós mátrixban.

$$_0\ \textit{Book}\ _1\ \textit{that}\ _2\ \textit{flight}\ _3$$

n hosszú mondat reprezentációjához $(n{+}1) \times (n{+}1)$ mátrix felső háromszöge kell

$[i,j]$ cella: olyan nem-terminálisok halmaza, amik az input i és j pozíciói közti konstituenseit tartalmazzák

CKY felismerés

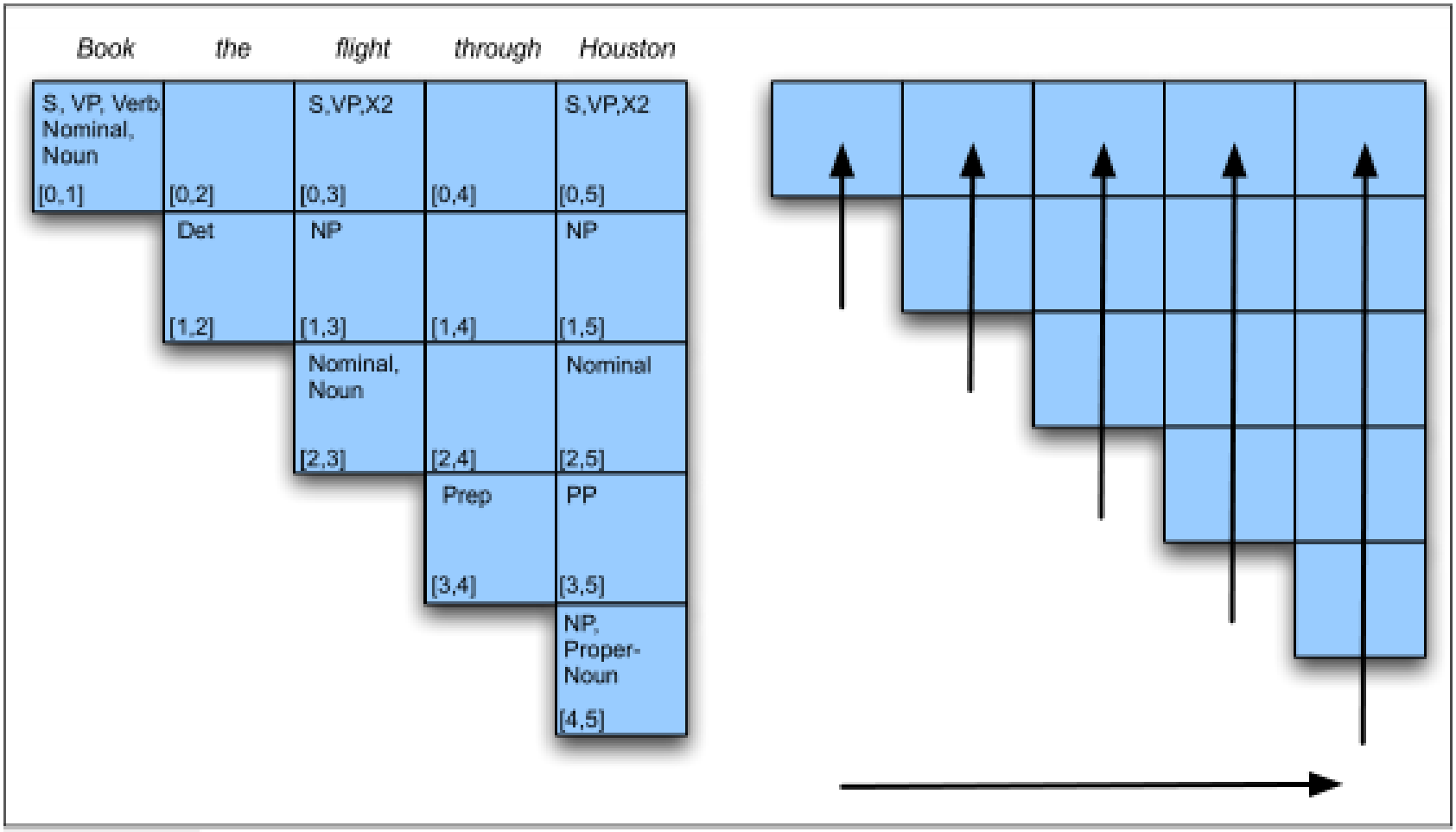

Figure 13.4 Completed parse table for *Book the flight through Houston.*

CKY felismerés

```
function CKY-PARSE(words, grammar) returns table

  for j ← from 1 to LENGTH(words) do
      for all {A | A → words[j] ∈ grammar}
              table[j − 1, j] ← table[j − 1, j] ∪ A
      for i ← from j − 2 downto 0 do
          for k ← i + 1 to j − 1 do
              for all {A | A → BC ∈ grammar and B ∈ table[i,k] and C ∈ table[k, j]}
                      table[i,j] ← table[i,j] ∪ A
```

Figure 13.5 The CKY algorithm.

CKY parszolás

Felismerő algoritmust két ponton kell megváltoztatni:

1. minden nem-terminálishoz mutatókat rendelni, amik a mátrix azon elemeire mutatnak, amikből eredeztethető
2. engedélyezni, hogy egy nem-terminálisnak több változatát is bevezethessük a mátrixba

Önkényes parszolás visszaadása: kijelölünk egy S szimbólumot a $[0,n]$ cellából, majd rekurzívan visszafejtjük a konstituenseket

CKY a gyakorlatban

CNF miatt a parszerünk által visszaadott fa nem egyezik a szintakták által rajzoltakkal, ami megnehezíti a szintaxisalapú szemantikai elemzéseket

Utófeldolgozás: kellő információt megtartunk, hogy a fánkat vissza tudjuk alakítani az eredeti nyelvtan szerint

→ olyan megoldás is szóba jöhet, ami elfogadja az önkényes, nem CNF környezetfüggetlen nyelvtanokat

Sekély parszolás

Chunking

Olyan diszjunkt szegmentumok felismerése és osztályozása, amelyek a nagyobb, tartalmas szavakat tömörítő szófajokkal címkézett nem-rekurzív frázisokat alkotják: *NP*, *VP*, *AdjP*, *PP*.

$[_{NP}$ The morning flight] $[_{PP}$ from] $[_{NP}$ Denver] $[_{VP}$ has arrived.]

Gépi tanulásos chunking

State-of-the-art chunking módszer: felügyelt gépi tanulás, ahol annotált adatok a tanító anyag.

IOB tagging: csoport kezdete (**B**), közepe (**I**), csoporton kívüli elem (**O**), címkék száma *2n + 1*, ahol *n* a csoportok típusainak száma

The	*morning*	*flight*	*from*	*Denver*	*has*	*arrived*
B_NP	I_NP	I_NP	B_PP	B_NP	B_VP	I_VP

The	*morning*	*flight*	*from*	*Denver*	*has*	*arrived*
B_NP	I_NP	I_NP	O	B_NP	O	O

Gépi tanulásos chunking

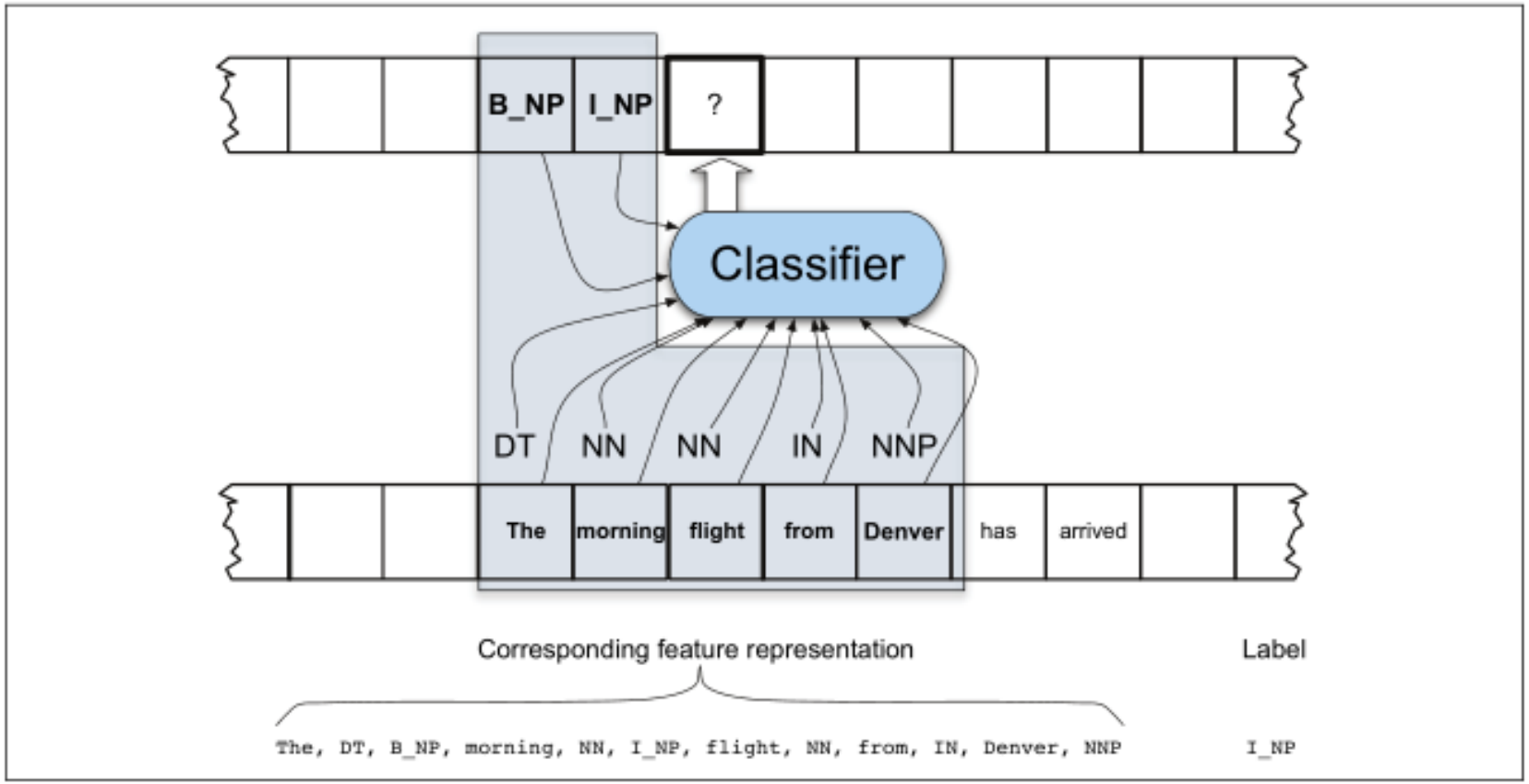

Figure 13.8 A sequence model for chunking. The chunker slides a context window over the sentence, classifying words as it proceeds. At this point, the classifier is attempting to label *flight*, using features like words, embeddings, part-of-speech tags and previously assigned chunk tags.

Chunking rendszerek kiértékelése

A csoportosítási folyamat outputjait hasonlítjuk hús-vér annotátorok gold-standard válaszaihoz, de nem szavankénti egyezést nézünk.

Befolyásoló tényezők:

- ***Precision***: rendszer által adott helyes csoportok száma / rendszer által adott összes csoport száma
- ***Recall***: rendszer által adott helyes csoportok száma / szövegben található összes csoport száma

F-measure:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad F_1 = \frac{2PR}{P+R}$$