Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel

Recski Gábor

MTA SZTAKI
Nyelvtechnológiai Kutatócsoport
recski@sztaki.hu

TLP20
2010. november 25.

Tartalom

A feladat

- NP-chunking: diszjunkt NP-szósorozatok azonosítása
- jellemzően a nem-rekurzív NP-ket keresi
- gépi tanuló algoritmusok kedvelt terepe
- angol nyelvre 94 – 96% közötti pontosság

A hunchunk rendszer

- Maximum Entrópiás tanulást és Rejtett Markov Modellt (HMM) használ
- minimális és maximális NP-ket is keres
- nem-rekurzív magyar NP-ken 94.75%-os F-pontszám

A magyar NP-nyelvtan

- környezetfüggetlen nyelvtan (CFG)
- 38 újraírószabály
- Hivatkozik
  - morfológiai jegyekre
  - projekciós szintekre

A nyelvtan és a KR-kód

NLTK parser

- nyílt forráskódú, Python nyelven írt
- számos parser algoritmust implementál
- támogatja a jegyekre hivatkozó újraírószabályokat

KR-kód

- KR: morfológiai annotációs formalizmus
- a KR-kódok a nyelvtan terminális szimbólumai
- a KR-beli jegyek és a nyelvtan által használt jegyek között egyértelmű megfeletetés van

Példa

óráján

NLTK	`NOUN[POSS=1, CAS=[SUE=1]]`
KR	`NOUN<POSS><CAS<SUE>>`

Névmások elkülönítése

- a névmásokat a KR-kód nem különbözteti meg a főnevektől
- nekünk szükségünk lesz rá:

ez	NOUN<PRON<DEM>>
mindenki	NOUN<PRON<GEN>>
valami	NOUN<PRON<INDEF>>
aki	NOUN<PRON<REL>>
ki	NOUN<PRON<WH>>

Egyéb kiegészítések

- Alapértelmezett jegyek megjelenítése

```
NOUN → NOUN[PERS=0, PRON=0, CAS=0, PLUR=0, POSS=0]
```

- Képzési adatok kódolása

harmadikkal
NUM[ORD]/NUM<CAS<INS>> →
→ NUM[CAS=[INS=1], SRC=[STEM=NUM, DERIV=ORD]]

A nyelvtan bővítése

A nyelvtan használatához szükség lesz

- AdjP-nyelvtanra

```
ADJ -> ADJ ADJ
ADJ -> ADV ADJ
```

- NumP-nyelvtanra

```
NUM -> NUM NUM
NUM -> ADV NUM
NUM -> ADJ NUM
```

A nyelvtan javítása (1)

Amiért szükség volt a névmásokra:

(1) *minden pofon*
(2) *néhány villanykörte*

```
(1)  NOUN -> NOUN[PRON=GEN] NOUN
(2)  NOUN -> NOUN[PRON=INDEF] NOUN
```

A nyelvtan javítása (2)

A mutató névmás egyezése:

ez a pincér
ezek a hajók
attól a pasastól

A szükséges szabály:

```
NOUN[POSS=?a, PLUR=?b, CAS=?c, D=?d] ->
-> NOUN[PRON=DEM, BAR=0, POSS=?a, PLUR=?b, CAS=?c]
ART NOUN[PRON=0, BAR=2, POSS=?a, PLUR=?b, CAS=?c, D=0]
```

A nyelvtan javítása (3)

Az AdjP-nyelvtan egy hiányossága:

(1) *a korsónak támasztott könyvet* olvasta
(2) *az ókori mór hódítóktól származó esküvést* hallották

A szükséges szabály:

```
(1)  ADJ ->
     NOUN ADJ[SRC=[STEM=VERB[], DERIV='PERF_PART']]
(2)  ADJ ->
     NOUN ADJ[SRC=[STEM=VERB[], DERIV='IMPERF_PART']]
```

A pontosság javulása

A parsert lépésenként kiértékeltük a Szeged Treebank alapján készült NP-korpuszon

Fejlesztés	F-pontszám
Kornai 1985	81.76%
AdjP, NumP	84.18%
Névmások	85.45%
“Ez a” szerkezet	86.68%
Deverbális melléknevek	87.87%
Írásjelek és kötőszavak	89.36%

Ez az eredmény még elmarad a `hunchunk` 94.75%-os teljesítményétől

Hibrid megoldás

- A `hunchunk` rendszer tanításakor felhasználjuk a szabályalapú elemző kimenetét.
- A parser kimenetét megfeleltetjük a `hunchunk` által használt chunk-címkéknek (`B-NP`, `I-NP`, `E-NP`, `1-NP`, `O`)

	Pontosság	Fedés	F-pontszám
`hunchunk`	94.61%	94.88%	**94.75%**
`hunchunk`+parser jegyek	95.29%	95.68%	**95.48%**

Köszönöm a figyelmet!


Kornai A.
The internal structure of Noun Phrases.
Approaches to Hungarian, 1:79–92, 1985.


S. Bird, E. Klein, and E. Loper.
Natural language processing with Python.
O'Reilly Media, 2009.


Recski G., Varga D., Zséder A., and Kornai A.
Fonévi csoportok azonosítása magyar-angol párhuzamos korpuszban.
VI. Magyar Számitógépes Nyelvészeti Konferencia, 2009.


Rebrus P., Kornai A., and Varga D.
Egy általános célú morfológiai annotáció.
Általános Nyelvészeti Tanulmányok, 2010.