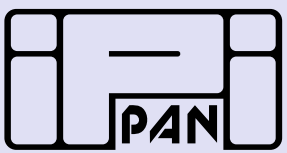
Parsowanie składniowe i jego zastosowania – warsztaty

Alina Wróblewska
alina@ipipan.waw.pl



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 15 kwietnia 2015

Outline

Etapy poprzedzające parsowanie składniowe

Jakie kroki poprzedzają parsowanie składniowe tekstu?

1. segmentacja tekstu na zdania,
2. tokenizacja – podział zdań na tokeny,
3. analiza morfologiczna,
4. lematyzacja – przypisanie tokenom ich form podstawowych,
5. tagowanie – przypisanie tokenom odpowiednich części mowy albo ujednoznacznianie analiz morfoskładniowych.

Outline

Parser Świgra

Parser Świgra online

- Dostępna jest robocza wersja parsera Świgra.
- Obecnie zdania można parsować w przeglądarce Firefox.

Instrukcja parsowania

- Do okna parsera należy wpisać zdanie do parsowania i wcisnąć Analizuj.
- Wyświetla się najbardziej prawdopodobne drzewo.
- Można przejść do innych proponowanych rozbiorów wciskając strzałki na krawędziach.

Przykładowe zdania do analizy

- Tych domów jest dużo.
- W nich można znaleźć różne rzeczy.
- Premier obiecał emerytom dopłaty państwa do leków.

Składnica frazowa

Składnica frazowa (Woliński i in., 2011)

- Zawiera ponad 8 tysięcy zdań automatycznie wybranych z ręcznie zaanotowanego podkorpusu Narodowego Korpusu Języka Polskigo (NKJP, Przepiórkowski i in., 2012).
- Struktury składnikowe przypisane do zdań zostały wygenerowane w sposób półautomatyczny:
  - parser *Świgra* (Woliński, 2004) automatycznie generuje dla każdego zdania zbiór możliwych drzew rozbioru,
  - wygenerowane drzewa są sprawdzane przez lingwistów, którzy albo wybierają najlepsze drzewo, albo odrzucają zdanie, jeśli nie ma poprawnego rozbioru.

Wyszukiwarka drzew ze Składnicy frazowej

`http://treebank.nlp.ipipan.waw.pl`

Wyszukiwarka drzew ze Składnicy frazowej

Język zapytań

- Atrybuty wierzchołków terminalnych:
  - orth – token (wyraz),
  - base – lemat leksemu reprezentowanego przez dany token,
  - pos – część mowy (np. subst, adj, fin, praet).
- Atrybuty wierzchołków nieterminalnych:
  - cat – kategoria składniowa,
  - przypadek, rodzaj, liczba, osoba, aspekt, czas, tryb, itd.

Przykłady zapytań

- [orth=wodą]
- [base=zgnić & pos=praet]
- [cat=fpm & przypadek=bier]
- [cat=fpm & przypadek=bier] $>^*$ [orth=na]
- [cat=fpm] > [cat=fps]

Outline

Parser zależnościowy online

Parser zależnościowy został włączony do Multiserwisu NLP języka polskiego – `http://multiservice.nlp.ipipan.waw.pl`.

Instrukcja parsowania

- Wybierz łańcuch przetwarzania (Select predefined chain of actions:) → 3: **Pantera, DependencyParser**
- Wstaw zdania do sparsowania do okna **Input text** np:
  - Tych domów jest dużo.
  - W nich można znaleźć różne rzeczy.
  - Demokrata Al Gore obiecał emerytom dopłaty państwa do leków.
- Wciśnij **Run**

MaltParser – trenowanie modelu i parsowanie

MaltParser

MaltParser (instalację, dokumentację, itp.) można pobrać ze strony `http://www.maltparser.org`.

Trenowanie MaltParsera

Polecenie trenowania modelu zależnościowego:

```
$ java -jar maltparser-1.8.jar -c nowyModel.mco -m learn
-i sciezka/do/danych/treningowych -a stackeager -l
liblinear -F modelCech
```

Parsowanie

Polecenie parsowania zdań:

```
$ java -jar maltparser-1.8.jar -c nowyModel.mco -m parse
-i sciezka/do/pliku/do/sparsowania -o
sciezka/do/pliku/zwracanego
```

MaltEval (Nilsson i Nivre, 2008)

Narzędzie do ewaluacji i porównywania sparsowanych drzew.

MaltEval – ewaluacja sparsowanych drzew zależnościowych

Jakość sparsowanych drzew można sprawdzić porównując je z ręcznie zaanotowanymi drzewami przy pomocy miar LAS i UAS:

```
$ java -jar MaltEval.jar -g sciezka/do/pliku/gold -s
sciezka/do/pliku/sparsowanego --Metric LAS
```

MaltEval – wizualizacja drzew zależnościowych

Porównanie drzew:

```
$ java -jar MaltEval.jar -g sciezka/do/pliku/gold -s
sciezka/do/pliku/sparsowanego -v 1
```

Format CoNLL drzew zależnościowych

Format CoNLL

Na potrzeby konkurencji w ramach CoNLL w 2006 i 2007 roku został zdefiniowany ujednolicony format kodowania drzew zależnościowych.

1. `ID` – identyfikator tokenu,
2. `FORM` – token,
3. `LEMMA` – forma podstawowa
4. `CPOSTAG` – ogólna część mowy,
5. `POSTAG` – część mowy,
6. `FEATS` – cechy morfologiczne
7. `HEAD` – identyfikator nadrzędnika,
8. `DEPREL` – typ relacji zależnościowej.

Format CoNLL – Przykład

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Na	na	prep	prep	acc	4	adj
2	wszelki	wszelki	adj	adj	sg|acc|m3|pos	3	adj
3	wypadek	wypadek	subst	subst	sg|acc|m3	1	comp
4	weźmiemy	wziąć	fin	fin	pl|pri|perf	0	pred
5	cię	ty	ppron12	ppron12	sg|acc|m1|sec|nakc	4	obj
6	za	za	prep	prep	acc	4	comp
7	wielbłąda	wielbłąd	subst	subst	sg|acc|m2	6	comp
8	.	.	interp	interp	_	4	punct

Outline

Dostępność parsera XLE

XLE jest dostępny dla celó niekomercyjnych po podpisaniu odpowiedniej licencji (zob. Obtaining XLE na stronie `http://www2.parc.com/isl/groups/nltt/xle/`).

Dokumentacja XLE

Dokumentacja XLE jest dostępna na stronie `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`

XLE

- System do parsowania i generowania zdań zgodnie z formalizmem LFG.
- Wersje XLE dostępne na systemy Mac OS X, Linux i Solaris Unix.

Parser XLE

Parsowanie XLE

- Polecenie uruchomienia systemu XLE:
  `$ xle`
- Polecenie stworzenia parsera z podaną gramatyką:
  `% create-parser sciezka/do/POLFIE.lfg`
- Polecenie parsowania:
  `% parse {Zdanie do sparsowania}`

Parser XLE-Web

- Dostępny na stronie
  `http://clarino.uib.no/iness/xle-web`.
- Parser aktualnie nie obsługuje języka polskiego.

Bank struktur LFG dla języka polskiego

Bank polskich struktur LFG

`http://clarino.uib.no/iness/page`

Wizualizacja analiz w INESS

- wybrać z menu po lewej stronie ‘Treebank selection’
- wybrać język ‘Polish’
- wybrać zbiór ‘pol-pargram’
- wybrać zdanie

System INESS umożliwia

- dezambiguację analiz (dla zalogowanych użytkowników),
- zaawansowane wyszukiwanie,
- pobieranie analiz w wybranym formacie,
- parsowanie (dla określonych języków),
- ...

Ankieta

Ankieta dla uczestników warsztatów

- Prosimy o wypełnienie ankiety po zakańczeniu uczestnictwa w warsztatach
- Ankieta jest dostępna na stronie `http://www.interankiety.pl/interankieta/3581bd3042422b2013fbeebf4524a2b3`.
- Ankieta będzie aktywna do 19 kwietnia.

Dziękuję za uwagę!