Parser składniowo-semantyczny dla języka polskiego

część I: preprocessing, konwersja Walentego i uzgodnienia

Wojciech Jaworski

Instytut Informatyki Uniwersytetu Warszawskiego

Instytut Podstaw Informatyki Polskiej Akademii Nauk

26 października 2015

Architektura

- Lematyzacja (morfeusz2)
- Preprocessing
- Parsowanie (walenty)
- Semantyka

Preprocessing

- Tekst reprezentowany jest jako graf.
- Krawędzie etykietowane są tokenami.
- Podczas preprocesingu dodawane są nowe krawędzie i wierzchołki
- Etapy preprocesingu:
  - identyfikacja złożonych znaków interpunkcyjnych
  - dezambiguacja znaczenia znaków interpunkcyjnych i wielkich liter w tym obsługa haplologii, np. 5 interpretacji dla sekwencji

    . . . w XV w. Warszawa . . .
  - rozpoznawanie liczebników zapisanych za pomocą cyfr, dat itp. np, -12.765,45 07.10.2015 13:15:
  - oznaczanie możliwych początków i końców zdań składowych (rozdzielanych przecinkami)
  - rozwijanie skrótów: *np.* $\rightarrow$ *na przykład*. (zasób: słownik skrótów)
  - rozpoznawanie wyrażeń wielosłownych (potrzebny zasób: wyrażenia wielosłowne odmienione przez przypadki i ew. liczby)
  - heurystyka tagująca nieznane tokeny
  - zwijanie interpretacji generowanych przez Morfeusza.

Spis treści

Gramatyka kategorialna (Type Logical Categorial Grammar)

- parsowanie jest dowodzeniem w niekomutatywnej intuicjonistycznej logice liniowej
  - można implementować fragmenty systemu dowodowego uzyskując szybkie parsery i mając gwarancję poprawności
- gramatyka jest w pełni zleksykalizowana (reguły gramatyczne to uniwersalne reguły dowodzenia w logice)
  - ułatwia to integrację z zasobami słownikowymi np. Walentym
- izomorfizm Curriego-Howarda wiąże reguły wnioskowania z termami liniowego rachunku lambda
  - zadaje sposób konstruowania formuł języka reprezentacji znaczenia podczas parsowania
- zamiast występującej w innych formalizmach gramatycznych unifikacji, dopasowywanie termu do wzorca
  - redukcja złożoności obliczeniowej

Spójniki LCG

Wyrażają podstawowe zjawiska występujące w języku.

- $\bullet$ konkatenacja (tworzenie wektorów cech)
- $\&$ niejednoznaczność
- $/$, $\backslash$, $|$ wymaganie argumentu
- $\oplus$ polimorficzne argumenty, sumowanie typów
- 1 opcjonalny argument
- ? wielokrotny argument
- $\&$ parametryzowana niejednoznaczność
- $\star$ tworzenie listy

Umożliwiają bezpośrednie wyrażenie informacji dostarczanej przez Morfeusza i zawartej w Walentym.

zielony adj:sg:nom.voc:m1.m2.m3:pos|adj:sg:acc:m3:pos

$$(adj \bullet sg \bullet (nom \,\&\, voc) \bullet (m1 \,\&\, m2 \,\&\, m3) \bullet pos) \,\&\, (adj \bullet sg \bullet acc \bullet m3 \bullet pos)$$

Walenty

- Słownik walencyjny — podstawowe źródło wiedzy o języku polskim.
- Schematy walencyjne są tłumaczone na wpisy do leksykonu LCG.
- Schematy są przetwarzane na żądanie, dla każdego zdania tylko te, które dotyczą leksemów zawartych w tym zdaniu.
- Aktualnie przetwarzam jedynie składniową warstwę Walentego.

Etapy przetwarzania schematu walencyjnego

- Konwersja schematu do OCamlowego systemu typów.

  subj{np(str),ncp(str,int)}+
  {refl}+{prepnp(o,loc);comprepnp(na temat)}

- Wstawienie do schematów realizacji fraz, podtypów atrybutów i atrybutów równoważnych.

  subj{np(str),ncp(str,int[co,czemu,czy,czyj,...])}+
  {refl}+{prepnp(o,loc);comprepnp(na temat)}

- Dodanie do pozycji w schemacie ról tematycznych i sensów (zaślepka).

  subj,Agnt{np(str),ncp(str,int[co,czemu,czy,czyj,...])}+
  Ptnt{refl}+Arg{prepnp(o,loc);comprepnp(na temat)}

- Uzupełnienie pozycji o informację o opcjonalności.

  subj,Agnt{pro,np(str),ncp(str,int[co,czemu,czy,czyj,...])}+
  Ptnt{refl}+Arg{null,prepnp(o,loc);comprepnp(na temat)}

Etapy przetwarzania schematu walencyjnego c.d.

- Usunięcie realizacji zawierających leksemy nie występujące w zdaniu.

  subj,Agnt{pro,np(str),ncp(str,int[czemu])}+
  Ptnt{refl}+Arg{null}

- Przetwarzanie leksykalizacji (opiszę w dalszym toku referatu).
- Ukonkretnienie schematu na podstawie klasy fleksemu i jego użycia (wskazanie konkretnych realizacji dla atrybutów str,part,żeby2,. . . oraz zmiana realizacji subj).

  subj,Agnt{pro,np(nomagr),ncp(nomagr,int[czemu])}+
  Ptnt{refl}+Arg{null}

- Uzupełnienie schematu o modyfikatory.

  subj,Agnt{pro,np(nomagr),ncp(nomagr,int[czemu])}+
  Ptnt{refl}+Arg{null}+{null,advp}+{null,prepp}

- Dodanie informacji o cechach semantycznych.

Generowanie wpisu w leksykonie

- Źródła danych: informacja morfosyntaktyczna formy, schemat walencyjny leksemu
- Dla leksemów nie występujących w Walentym generuję schematy typowe dla ich typów leksemów.
- Typy fraz występujące w gramatyce są zbudowane na podstawie typów fraz z Walentego oraz kategorii gramatycznych z Morfeusza

np • *number* • *case* • *gender* • *person*
nump • *number* • *case* • *gender* • *person*
adjp • *number* • *case* • *gender*

- np są frazami nominalnymi z wyłączeniem liczebnikowych.
- Przykładowo token ’zielony’ oznaczony przez Morfeusza jako

zielony adj:sg:nom.voc:m1.m2.m3:pos|adj:sg:acc:m3:pos

zostanie przetłumaczony na następujące wpisy:

adjp • sg • (nom & voc) • (m1 & m2 & m3)

adj • sg • acc • m3

Generowanie wpisu w leksykonie c.d.

prepnp • *prep* • *case*
prepadjp • *prep* • *case*
comprepnp • *prep*

- Frazy przyimkowe zawierają leksem przyimka.
- prepnp i comprepnp obejmują frazy przyimkowo-liczebnikowe.

  cp • *ctype* • *comp*
  ncp • *number* • *case* • *gender* • *person* • *ctype* • *comp*
  prepncp • *prep* • *ctype* • *comp*
  infp • *aspect*
  advp
  fixed • *lex*

- comparp, or, E nie są jeszcze opracowane

Generowanie wpisu w leksykonie c.d.

Typy fraz nie występujące w Walentym

ip • *number* • *gender* • *person*
padvp
prepp • *case*
qub
inclusion
adja
aglt • *number* • *person*
aux-past • *number* • *gender* • *person*
aux-fut • *number* • *gender* • *person*
aux-imp
lex

- pro oraz null są zamieniane w opcjonalny argument 1
- *lex* to typ frazy zawierający tylko jeden leksem, nazywający się tak jak ten leksem, np. nie, się.

Argumenty

- Argumenty wyglądają analogicznie do podstawowych wpisów.
- $\top$ oznacza, że wartość danego pola jest dowolna.
- Opcjonalność argumentów wyraża $\oplus$.
- $/, \backslash$ i $|$ wyrażają położenie argumentu względem nadrzędnika
- Wpis dla przyimka w:loc

$$\text{prepnp} \bullet \text{w} \bullet \text{loc} \,/\, \text{np} \bullet \top \bullet \text{loc} \bullet \top \bullet \top$$

- Wpis dla czasownika w formie osobowej mającego argument {null,prepnp(o,loc);comprepnp(na temat)}:

$$\text{ip} \bullet \textit{number} \bullet \textit{gender} \bullet \textit{person}$$

$$| \; 1 \oplus \text{prepnp} \bullet \text{w} \bullet \text{loc} \oplus \text{comprepnp} \bullet \text{na temat}$$

Uzgodnienie rzeczownika z przymiotnikiem

- Rzeczownik subst:sg:nom.acc:m3,
  którego podrzędnikiem jest przymiotnik (wersja uproszczona)

$$\&_{case\in\{\text{nom,acc}\}} \text{np} \bullet \text{sg} \bullet case \bullet \text{m3} \bullet \text{ter} \mid \text{adjp} \bullet \text{sg} \bullet case \bullet \text{m3}$$

- Formuła

$$\&_{x\in\{a_1,a_2,\ldots,a_n\}} \varphi(x)$$

  jest równoważna

$$\varphi(a_1) \,\&\, \varphi(a_2) \,\&\, \cdots \,\&\, \varphi(a_n)$$

- Pełny wpis

$$\&_{case\in\{\text{nom,acc}\}} \text{np} \bullet \text{sg} \bullet case \bullet \text{m3} \bullet \text{ter}\{$$

$$/\,1 \oplus \text{adjp} \bullet \text{sg} \bullet case \bullet \text{m3},$$

$$\backslash ?\text{adjp} \bullet \text{sg} \bullet case \bullet \text{m3}\}$$

Uzgodnienie czasownika

- Czasownik w czasie przeszłym z podmiotem np i aglutynatem

$$\&_{gender\in\{m1,m2,m3\}} \&_{person\in\{pri,sec\}} \text{ip} \bullet \text{sg} \bullet gender \bullet person\{$$

$$| \ 1 \oplus \text{np} \bullet \text{sg} \bullet \text{nom} \bullet gender \bullet person,$$

$$| \ \text{aglt} \bullet \text{sg} \bullet gender \bullet person\}$$

- Czasowniki posiłkowe i aglutynaty są podrzędnikami leksemów do których przynależy schemat.
- Czas przeszły czasownika typu winien

$$\&_{gender\in\{m1,m2,m3\}} \&_{person\in\{pri,sec\}} \text{ip} \bullet \text{sg} \bullet gender \bullet person\{$$

$$| \ \text{aux-past} \bullet \text{sg} \bullet gender \bullet person,$$

$$| \ \text{aglt} \bullet \text{sg} \bullet gender \bullet person\}$$

Uzgodnienia liczebnika

- Uzgodnienie dla formy „dwóch” z interpretacją num:pl:gen.loc:m1.m2.m3.f.n2:congr

$$\&_{case\in\{gen,loc\}} \&_{gender\in\{m1,m2,m3,f,n2\}} \text{nump} \bullet \text{pl} \bullet \textit{case} \bullet \textit{gender} \bullet \text{ter}$$

$$/\ \text{np} \bullet \text{pl} \bullet \textit{case} \bullet \textit{gender} \bullet \text{ter}$$

- oraz z interpretacją num:pl:nom.acc.voc:n1.p1.p2:rec

$$\&_{case\in\{nom,acc,voc\}} \&_{gender\in\{n1,p1,p2\}} \text{nump} \bullet \text{sg} \bullet \textit{case} \bullet \text{n2} \bullet \text{ter}$$

$$/\ \text{np} \bullet \text{pl} \bullet \textit{case} \bullet \textit{gender} \bullet \text{ter}$$

- Fraza liczebnikowa jako podmiot czasownika w formie osobowej

$$\text{ip} \bullet \textit{number} \bullet \textit{gender} \bullet \textit{person}$$

$$|\ \text{nump} \bullet \textit{number} \bullet \text{nom} \bullet \textit{gender} \bullet \textit{person}$$

Poprzyimkowe formy zaimków osobowych

- Kluczowa jest forma „nie”: jeśli nie uwzględnimy poprzyimkowości, uzyskamy dużą liczbę błędnych rozbiorów dla zdań z negacją.
- Wymaganie poprzyimkowości realizujemy za pomocą podniesienia typu.
- Na przykład formie „nią” ppron3:sg:acc.inst:f:ter:_:praep zamiast typu

$$\mathrm{np} \bullet \mathrm{sg} \bullet (\mathrm{acc} \,\&\, \mathrm{inst}) \bullet \mathrm{f} \bullet \mathrm{ter}$$

nadajemy typ

$$\&_{\mathit{prep}} \; \&_{\mathit{case} \in \{\mathrm{acc},\mathrm{inst}\}} \; \mathrm{prepnp} \bullet \mathit{prep} \bullet \mathit{case}$$

$$\backslash(\mathrm{prepnp} \bullet \mathit{prep} \bullet \mathit{case} \,/\, \mathrm{np} \bullet \mathrm{sg} \bullet \mathit{case} \bullet \mathrm{f} \bullet \mathrm{ter})$$

Zaimki względne i pytajne

- Spełniają dwie funkcje:
  - realizują argumenty czasownika,
  - zastępują ip przez cp.
- „Kto pyta?” kto subst:sg:nom:m1

$$\mathrm{cp} \bullet \mathrm{int} \bullet \mathrm{kto} \,/ (\mathrm{ip} \bullet \top \bullet \top \bullet \top \mid \mathrm{np} \bullet \mathrm{sg} \bullet \mathrm{nom} \bullet \mathrm{m1} \bullet \mathrm{ter})$$

- „W związku z czym pyta?” czym subst:sg:inst.loc:n2

$$\&_{prep} \; \&_{case \in \{\mathrm{inst,loc}\}} \; [\mathrm{cp} \bullet \mathrm{int} \bullet \mathrm{co} \,/ (\mathrm{ip} \bullet \top \bullet \top \bullet \top \mid \mathrm{comprepnp} \bullet prep)]$$

$$\backslash (\mathrm{comprepnp} \bullet prep \,/\, \mathrm{np} \bullet \mathrm{sg} \bullet case \bullet \mathrm{n2} \bullet \mathrm{ter})$$

- Możliwe kryterium na przyimek złożony: występowanie w zdaniu przez zaimkiem pytajnym lub względnym.

Zaimki względne i pytajne

- „O którą książkę pyta?” którą adj:sg:acc.inst:f:pos

$$\&_{prep} \; \&_{case \in \{acc,inst\}} \; [[cp \bullet int \bullet który \,/\, (ip \bullet \top \bullet \top \bullet \top \mid prepnp \bullet prep \bullet case)]$$

$$\backslash (prepnp \bullet prep \bullet case \,/\, np \bullet sg \bullet case \bullet f \bullet ter)]$$

$$/(np \bullet sg \bullet case \bullet f \bullet ter \backslash adjp \bullet sg \bullet case \bullet f)$$

- Wykorzystanie podniesienia typu jest możliwe dzięki temu, że zaimki te są na początku zdania.
- Nie można byłoby z niego skorzystać, gdyby stały pomiędzy argumentami czasownika
- Jest to szczęśliwy zbieg okoliczności, albo przesłanka za tym, że opis w LCG odzwierciedla składnię języka polskiego

Leksykalizacje

balansować:
{lex(prepnp(na,loc),sg,XOR('krawędź','skraj'),atr1({np(gen)}))}

- Konwersja nazw fraz na nazwy fleksemów.
  balansować:
  {lex([prep(loc),'na';subst(sg,loc),XOR('krawędź','skraj')],
  atr1({np(gen)}))}
- Zamiana list fleksemów na argumenty.
  balansować:
  {lex(prep(loc),'na',ratr({lex(subst(sg,loc),XOR('krawędź','skraj'),
  atr1({np(gen)}))}))}
- Utworzenie schematów z leksykalizacji.
  balansować: {lex(1,prep(loc),'na')}
  lex(1,na): ratr({lex(2,subst(sg,loc),XOR('krawędź','skraj'))})
  lex(2,krawędź): atr1({np(gen)})
  lex(2,skraj): atr1({np(gen)})

Leksykalizacje

lex(1,na): ratr({lex(2,subst(sg,loc),XOR('krawędź','skraj'))})
lex(2,krawędź): atr1({np(gen)})
lex(2,skraj): atr1({np(gen)})

- Rozwinięcie modyfikacji.
  lex(1,na): {lex(2,subst(sg,loc),XOR('krawędź','skraj'))}
  lex(2,krawędź): {null;np(gen)}
  lex(2,skraj): {null;np(gen)}
- Rozwinięcie list lematów.
  lex(1,na): {lex(2,subst(sg,loc),'krawędź')}
  lex(1,na): {lex(2,subst(sg,loc),'skraj')}
- Generowanie wpisu w leksykonie

$$\text{lex} \bullet 1 \bullet \text{na} \bullet \text{prep} \bullet \text{loc} \,/\, \text{lex} \bullet 2 \bullet \text{krawędź} \bullet \text{subst} \bullet \top \bullet \text{loc} \bullet \top \bullet \top$$

$$\text{lex} \bullet 2 \bullet \text{krawędź} \bullet \text{subst} \bullet \textit{number} \bullet \textit{case} \bullet \text{f} \bullet \text{ter}$$
$$\mid \text{np} \bullet \top \bullet \text{gen} \bullet \top \bullet \top$$

Spis treści

Testy

- Test wykonany na pierwszych 3746 zdaniach ze Składnicy
- Bez wstępnej anotacji.

1760	0,470	parsed
1663	0,444	not parsed
200	0,053	unknown token
72	0,019	timeout (100s)
51	0,014	error

- Rodzaje błędów:
  - 39 generowanie leksykonu
  - 7 Stack overflow
  - 5 Naruszenie ochrony pamięci
- Średnie czasy (przybliżone zwn. wielokrotne uruchamianie parsera):

0,41-0,43s	preprocessing
0,01s	generowanie wpisów do leksykonu
2,80-3,91s	parsowanie

- Na czas preprocesingu składa się głównie czas uruchamiania Morfeusza.