Morfologik

Marcin Miłkowski
IFiS PAN

koordynator pl.OpenOffice.org

Adres projektu: *morfologik.blogspot.com*

Morfologik

- Trzy składniki projektu:
  - Automatycznie generowany słownik form wyrazowych
  - Reguły korektora LanguageTool
  - Biblioteka Stempelator (Java)

Słownik form

- Słownikowy anotator (POS tagger)
  - W formacie FSA (biblioteka Stempelator)
  - Oparty na słowniku alternatywnym (format ispell/myspell)
  - Generowany automatycznie
  - Dodatkowa funkcja: lematyzacja

Słownik form

- Słownik ispella:
  - Słownik definiujący flagi afiksów: prefiksów i sufiksów
  - Słownik wyrazów
- Poszczególne flagi odpowiadają wielu przypadkom, rodzajom...

- *Świstak mówi: niemożliwe!*

Słownik form

- Jednak Zbigniew Płotnicki napisał pracę dyplomową u dra Jacka Jelonka, którą był słownik morfologiczny oparty na ispellu (program waspell).
  - Waspell miał być udostępniony na licencji GNU (tak jak słownik alternatywny)...
  - Do dziś niedostępny, ale:
    - Wiedziałem, że to możliwe
    - Miałem wersję binarną

Słownik form

- Generowanie słownika, algorytm prosty:
  - 1. Wygenerowanie wszystkich form ze słownika ispella, wraz z podaniem bazowych flag i końcówek (7 pól płaskiej bazy danych)
  - 2. Wypisanie, przy użyciu tablicy asocjacyjnej, wszystkich form

Słownik form

- Generowanie słownika, algorytm prosty:
  - 1. Wygenerowanie wszystkich form ze słownika ispella, wraz z podaniem bazowych flag i końcówek (7 pól płaskiej bazy danych)

```
Abbie Abba   M   M   a   ie  [^cdghijklłnrtvwzż]a
Abbo  Abba   M   M   a   o   [^l]a
Abbocie  Abbot  OoSsT  O   t   cie    [^s]t
Abbot Abbot  OoSsT  OoSsT  0   0
```

Słownik form

- Generowanie słownika, algorytm prosty:
  - 2. Wypisanie, przy użyciu tablicy asocjacyjnej:

```
Mmn  n  a     a  subst:pl:gen:f
OSTos   O  ów owa   ów subst:sg:gen:m
OSTos   O  ów owa   ów subst:sg:gen:m1
OSTos   O  ów owem  ów subst:sg:inst:m
OSTos   O  ów owem  ów subst:sg:inst:m1
OSTos   O  ów owie  ów subst:pl:nom.voc:m
OSTos   O  ów owie  ów subst:pl:nom.voc:m1
OSTos   O  ów owowi ów subst:sg:dat:m
```

  - Pole 1 to zestaw flag formy podstawowej wyrazu: pozwala odróżnić np. wyrazy typu „niezdara” rodzaju męskiego od „niezdara” rodzaju żeńskiego

Słownik form

- Prosty algorytm nie uwzględnia wyrazów nieodmiennych oraz nieregularnych (=bez flagi ispella)
  - Rozwiązanie hybrydowe:
    - Słownik ręcznie dopisanych form
    - Zgadywanie, na podstawie słownika odmian słownika alternatywnego, przynajmniej części flagi:

```
wystrużże   wystrugać verb:irreg
wlepiwszy   wlepić pant:perf
```

Formy anotowane na podstawie końcówek (-wszy & verb:irreg = pant:perf, adj -kszy = adj:comp)

Słownik form

- Zestaw znaczników oparty na tagsecie IPI
- Część znaczników stosowana inaczej (błędnie), zwłaszcza aglt (z powodu tokenizatora):

```
abdykowali      abdykować verb:praet:pl:ter:m1:?perf
abdykowaliby    abdykować verb:aglt:praet:pl:ter:m1:?perf
```

Słownik form

- Dodatkowe znaczniki:
  - **pneg**: forma może być zanegowana (np. „biały”)
  - **neg**: forma faktycznie zanegowana (np. „niebiały)
- Bałagan w oznaczeniach deklinacji
  - (m1...)

Anotacje – skąd pneg?

- Dodatkowy znacznik został wprowadzony w celu ulepszenia reguły błędnej pisowni rozłącznej:
  - „Prosimy o nie palenie”
- W słowniku alternatywnym flaga prefiksu *b* jest przypisana do podzbioru wyrazów, które w podręcznikowych regułach ortografii pisze się łącznie z „nie” → mniej fałszywych alarmów

Testowanie anotacji

- Nie istnieje jeszcze niezawodny i dokładny walidator znaczników (np. sprawdzający kompletność anotacji)
- Ale: błędy w oznaczeniu są wyławiane podczas testowania reguł gramatycznych i stylistycznych
  - Stąd np. „pneg”

Słownikowy POS tagger

- Wady:
  - Brak jednoznacznej anotacji
    - W korekcie gramatycznej to jest zaleta, statystyczne taggery nie są trenowane na niegramatycznych zdaniach
  - Brak form spoza słownika
    - Słownik alternatywny został wzbogacony o prawie 500 tysięcy form nazw własnych (na moją prośbę)
- Zaleta:
  - Szybkość, działa bez parsera

Słownikowy POS tagger

- Słownik alternatywny jest aktualizowany na bieżąco
  - Trzeba aktualizować tablicę asocjacyjną
  - Zasób słownika się powiększa

LanguageTool

- Regułowy korektor gramatyczny
  - Jednostka przetwarzania: zdanie (definiowane syntaktycznie, przez interpunkcję)
  - Błędy wykrywane przez zgodność z regułami wzorców; inne korektory stosują
    - Gramatyki sformalizowane (pełne, np. Link Grammar)
    - Statystyczne wzory zdań poprawnych lub błędnych
    - Rozwiązania hybrydowe (Wordnet + Google)

LanguageTool

- Modułowy:
  - Dla każdego języka osobny POS tagger
  - Probablistyczny tagger z OpenNLP został zastąpiony słownikowym (w najnowszej wersji)
  - Obsługuje angielski, niemiecki, polski; są już taggery dla włoskiego, hiszpańskiego i francuskiego, w przygotowaniu dla Malayalam
  - Ma działać w OpenOffice.org, ale też np. w *controlled technical English* (OCE)

LanguageTool

- Plik reguł w XML-u; siła wyrazu z grubsza podobna do zapytania do Poliqarpa (możliwe zapewne przekształcenie XSLT do zapytania)
- Wprowadzane elementy relacyjne (pomijanie określonej liczby wyrazów, wyjątki i zakresy wyjątków)
- Brakuje jeszcze kilku możliwości specyfikowania relacji między elementami

LanguageTool

- Fałszywe alarmy – problem korektorów gramatycznych i stylistycznych (*vide* korektor w MS Word)
- Testowanie reguł na prawdziwych tekstach pokazuje często błędność anotacji lub konieczność wprowadzenia nowego znacznika: sprzężenie zwrotne

Plany

- Statystyczny lub regułowy segmentator (chunker; np. do frazy nominalnej) *zamiast* parsera
- Wprowadzenie dodatkowych poziomów analizy: akapit i cały tekst
- Ulepszenie sugestii i wprowadzenie koniugatora (słownika generującego zadane formy)
- Nightly builds

Plany

- Sprawdzanie formy polskich wyrazów online (w czasie sprawdzania pisowni)
- Aplikacja do testowania i pisania reguł online (konieczne do współpracy z Włochami, Francuzami itd.)
- Testowanie na korpusie błędów (z Internetu, uczniów, gazet...)
- LT jako wtyczka do Firefoksa
- Udoskonalenie LT jako modułu OOo