Tagery morfosyntaktyczne dla języka polskiego

Łukasz Kobyliński Witold Kieraś

Instytut Podstaw Informatyki Polskiej Akademii Nauk
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

7.12.2015

Wprowadzenie

Cele prezentacji

- podsumowanie obecnego stanu narzędzi do tagowania morfosyntaktycznego w języku polskim,
- porównanie dokładności i wykorzystywanych algorytmów z narzędziami dla innych języków europejskich,
- analiza jakościowa wyników działania poszczególnych tagerów,
- przegląd problemów, które nie zostały rozwiązane przez istniejące tagery,
- stwierdzenie, czy wśród dostępnych narzędzi istnieje tager o pożądanych cechach,
- rekomendacje dotyczące dalszych kroków.

Wprowadzenie

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- nie jest nadgorliwy, można kazać zostawić interpretacje częściowo nieujednoznacznione (np. usunąć tylko bardzo złe interpretacje),
- informuje o poziomie pewności podjętych decyzji,
- działa na niejednoznacznej segmentacji (stworzonej przez Morfeusza lub np. będącej wynikiem zastosowania po Morfeuszu słownika wyrażeń wieloczłonowych),
- daje się (względnie?) łatwo zainstalować i uruchomić na wszystkich platformach, na których jest Morfeusz,
- da się rozszerzyć o uwzględnianie informacji o czasie powstania tekstu.

Plan

1. Tagery języka polskiego – przegląd rozwiązań
2. Tagery morfosyntaktyczne dla innych języków europejskich
3. Tagery języka polskiego – analiza ilościowa
4. Tagery języka polskiego – analiza jakościowa
5. Dyskusja i rekomendacje

Tagery języka polskiego – przegląd rozwiązań

Czym jest tagowanie – przypomnienie

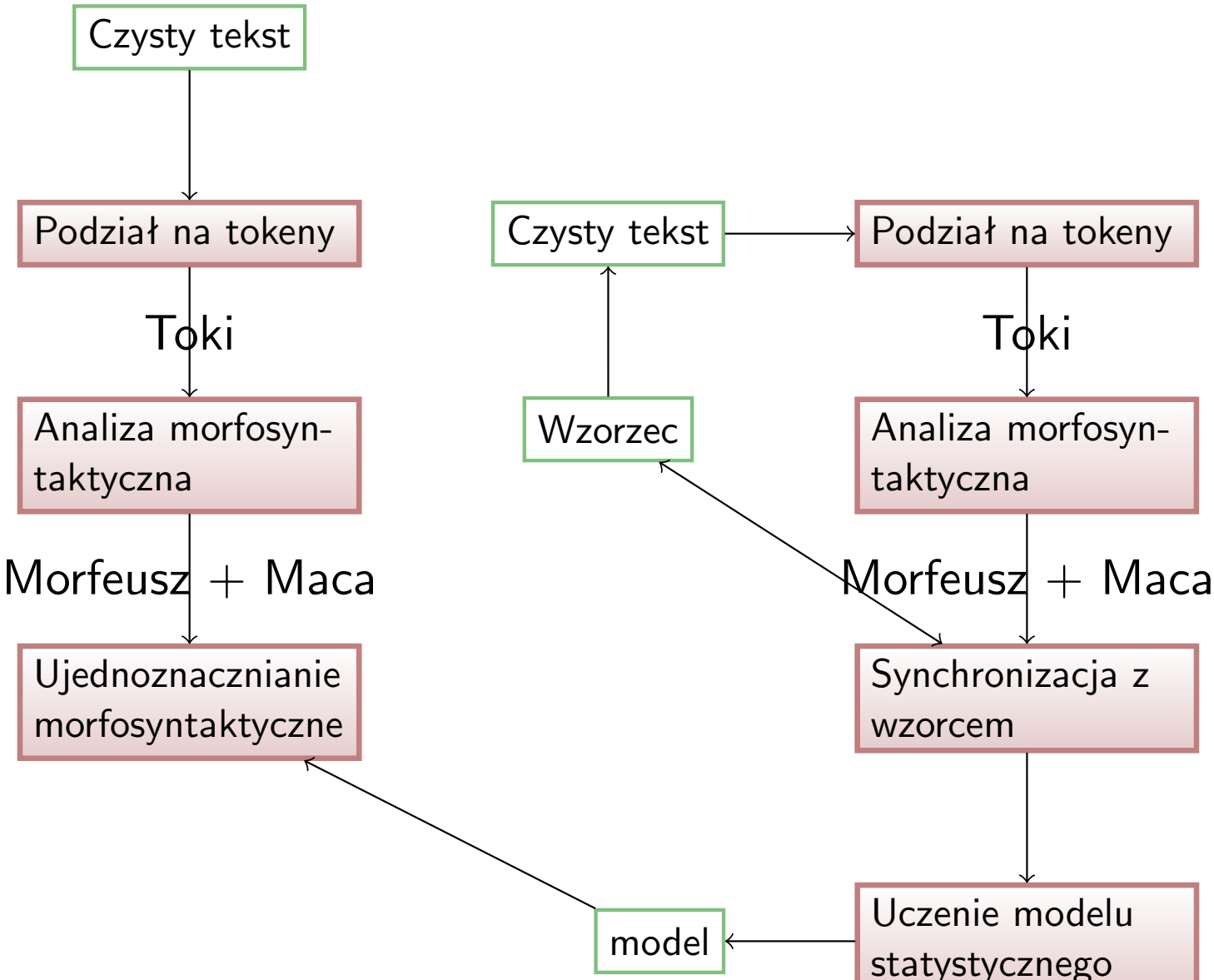
Segment (token) – wyraz lub jego fragment, znak interpunkcyjny, ciąg cyfr lub symboli. Segmenty są ciągłe oraz rozłączne.

Znacznik morfosyntaktyczny (tag) – symbol, który można przypisać segmentowi, określający jego własności morfologiczno-składniowe.

Znakowanie morfosyntaktyczne (tagowanie) – zadanie przypisania ciągowi segmentów ciągu znaczników morfosyntaktycznych.

Segmentacja $\Rightarrow$ **Analiza morfosyntaktyczna** $\Rightarrow$ **Ujednoznacznianie morfosyntaktyczne**

Pełny stos przetwarzania



Tagery morfosyyntaktyczne dla języka polskiego

Tagery „archiwalne”

Dostosowane do tagsetu IPI PAN, modele uczone na korpusie IPI.

- tager Ł. Dębowskiego – statystyczny tager trigramowy,
  weak correctness $= 90,59\%$
- TaKIPI – tager hybrydowy, oparty na drzewach decyzyjnych, częstości unigramów i ręcznie utworzonych regułach.
  weak correctness $= 91.30\%$

Tagery morfosyntaktyczne dla języka polskiego

Tagery uwzględniające tagset NKJP

- Pantera [Acedański 2010] – adaptacja algorytmu Brilla do języków bogatych morfologicznie, takich jak polski,
- WMBT [Radziszewski and Śniatowski 2011] – tager oparty na uczeniu pamięciowym, rozbudowany o wielowarstwowość dla uwzględnienia wielu atrybutów znakowania w języku polskim,
- Concraft [Waszczuk 2012] – tager warstwowy, oparty na Conditional Random Fields (CRF); wyniki dezambiguacji morfosyntaktycznej przekazywane są z jednej warstwy do drugiej,
- WCRFT [Radziszewski 2013] – również oparty na CRF; osobne modele wykorzystywane są do dezambiguacji poszczególnych atrybutów opisu morfosyntaktycznego,

Tagery morfosyntaktyczne dla języka polskiego

Modele dla tagerów zaimplementowanych dla innych języków

- TnT Tagger – statystyczny tager trigramowy, model dla PL przygotowany przez M. Miłkowskiego, dokładność „ok. 88%”,
  `http://zil.ipipan.waw.pl/NKJP%20model%20for%20TnT%20Tagger`
- OpenNLP – tager maksimum entropii, model dla PL przygotowany przez P. Pęzika (nie jest udostępniony publicznie).
  `http://clarin.pelcra.pl/tools/tagger`

Tager Brilla – zasada działania

Uczenie tagera

- wytrenuj tager unigramowy na podstawie zbioru uczącego,
- otaguj zbiór rozwojowy za pomocą tagera unigramowego,
- iteracyjnie znajdź transformacje, które mogą poprawić największą liczbę błędów, wprowadzając jednocześnie jak najmniej pomyłek
- zachowaj zbiór najlepszych transformacji – model.

r	good(r)	bad(r)
Zmień przypadek przyimka z `acc` na `loc`, jeśli kończy się na `na` i jeden z kolejnych tokenów jest w przypadku `loc`.	2496	113
Zmień przypadek przymiotnika z `loc` na `inst`, jeśli jeden z kolejnych tokenów ma przypadek `inst` i kończy się na `em`.	921	29

Tager Brilla – przykładowe reguły

Transformacje Zachodzą według jednego z szablonów:

- $t_i := A$ *if* $t_i = B \wedge \exists_{o \in O_1} t_{i+o} = C$
- $t_i := A$ *if* $t_i = B \wedge \forall_{o \in O_2} t_{i+o} = D$
- $t_i := A$ *if* $t_i = B$ i $i-$te słowo jest z wielkiej litery
- $t_i := A$ *if* $t_i = B$ i $(i-1)-$te słowo jest z wielkiej litery

gdzie:

- $O_1 \in \{\{1\}, \{-1\}, \{2\}, \{-2\}, \{1,2\}, \{-1,-2\}, \{1,2,3\}, \{-1,-2,-3\}\}$,
- $O_2 \in \{\{-2,-1\}, \{-1,1\}, \{1,2\}\}$,
- A, B, C, D – tagi.

WCRFT i Concraft – zasada działania

WCRFT

- wieloprzebiegowe ujednoznacznianie niezależnych warstw: klasa gramatyczna, liczba, przypadek, itp.
- modele pierwszego rzędu – kontekst analizowany na poziomie obserwacji,
- cechy: forma ortograficzna (*orth*), bigramy *orth*, klasa gramatyczna, bigramy, trigramy *orth*, przypadek, rodzaj, liczba, zgodność gramatyczna

Concraft

- wprowadzenie ograniczeń co do możliwości występowania tagów w danym kontekście w algorytm CRF (ograniczony liniowy model CRF),
- ujednoznacznianie w dwóch warstwach wpływających na siebie (część mowy + przypadek + osoba, pozostałe kategorie gramatyczne),
- cechy: forma ortograficzna, dla OOV: prefiks, sufiks, pocz. zdania,

Przenośność i łatwość wykorzystania

- **Concraft** instalowany i uruchamiany z wykorzystaniem Haskell Platform, która dostępna jest pod wszystkie główne systemy operacyjne,
- **WCRFT, Pantera** – wymagają kompilacji, proces kompilacji dostosowany do środowiska Linuksowego,
- **WMBT** – Python.

Concraft, WCRFT, WMBT – silnie zależą od stosu Corpus2 / Toki / Maca, których kompilacja pod Windows jest możliwa, ale nietrywialna (Visual Studio).

Tagery morfosyntaktyczne dla innych języków europejskich

Tagowanie języka angielskiego

Penn Treebank Wall Street Journal (WSJ) release 3

System	Metoda	Publikacja	Dokładność
BI-LSTM-CRF	Bidirectional LSTM-CRF Model	Huang et al. (2015)	97.55
SCCN	Semi-supervised condensed nearest neighbor	Søgaard (2011)	97.50
Morče/COMPOST	Averaged Perceptron	Spoustová et al. (2009)	97.44
structReg	CRFs with structure regularization	Sun(2014)	97.36
LTAG-spinal	Bidirectional perceptron learning	Shen et al. (2007)	97.33
Stanford Tagger 2.0	Maximum entropy cyclic dependency network	Manning (2011)	97.32
Stanford Tagger 2.0	Maximum entropy cyclic dependency network	Manning (2011)	97.29
Stanford Tagger 1.0	Maximum entropy cyclic dependency network	Toutanova et al. (2003)	97.24

Tagowanie języka angielskiego

System	Metoda	Publikacja	Dokładność
Morče/COMPOST	Averaged Perceptron	Spoustová et al. (2009)	97.23
LAPOS	Perceptron based training with lookahead	Tsuruoka, Miyao, and Kazama (2011)	97.22
SVMTool	SVM-based tagger and tagger generator	Giménez and Márquez (2004)	97.16
Maxent easiest-first	Maximum entropy bidirectional easiest-first inference	Tsuruoka and Tsujii (2005)	97.15
Averaged Perceptron	Averaged Perception discriminative sequence model	Collins (2002)	97.11
GENiA Tagger**	Maximum entropy cyclic dependency network	Tsuruoka, et al (2005)	97.05
MElt	MEMM with external lexical information	Denis and Sagot (2009)	96.96
TnT*	Hidden markov model	Brants (2000)	96.46

Tagery języków europejskich – porównanie

Tager	Język	Rozmiar tagsetu	Korpus treningowy	Dokładność
Concraft	polski	4 000 / 1 000	1M	91,07%
Obeliks	słoweński	1 903	500k	91,34%
Morče	czeski	3 922 / 1 571	2M	95,67%
Featurama	czeski	3 922 / 1 571	2M	95,66%
Morphodita	czeski	3 922 / 1 571	2M	95,75%
BI-LSTM-CRF	angielski	36+12	1M	97,55%

Tagery języka polskiego – analiza ilościowa

Metoda ewaluacji

Miara jakości znakowania

- ze względu na możliwość wystąpienia różnic w segmentacji pomiędzy wynikiem znakowania, a złotym standardem, wykorzystujemy dolne ograniczenie trafności (*accuracy lower bound*, Acc_{lower}) do oceny dokładności tagerów,
- miara ta karze wszelkie zmiany segmentacyjne w stosunku do złotego standardu i traktuje takie tokeny jako sklasyfikowane błędnie,
- token traktowany jest jako oznakowany prawidłowo, jeśli zbiór jego interpretacji ma niepuste przecięcie ze zbiorem interpretacji zwracanych przez tager,
- niezależne sprawdzamy dokładność dla znanych (Acc^{K}_{lower}) i nieznanych słów (Acc^{U}_{lower}), aby ocenić skuteczność ew. modułów odgadywania.

Ewaluacja tagerów

Eksperymenty na milionowym podkorpusie Narodowego Korpusu Języka Polskiego, ver. 1.1, 10-krotna walidacja krzyżowa.

n	Tager	Acc_{lower}	Acc^{K}_{lower}	Acc^{U}_{lower}
1	Pantera	88.95%	91.22%	15.19%
2	WMBT	90.33%	91.26%	60.25%
3	WCRFT	90.76%	91.92%	53.18%
4	Concraft	91.07%	92.06%	58.81%

- Acc_{lower} – łączna dokładność,
- Acc^{K}_{lower} – dokładność dla znanych słów,
- Acc^{U}_{lower} – dokładność dla słów nieznanych (2,8% Morfeusz1, 1,6% Morfeusz2).

Analiza rezultatu działania tagerów

Porównanie wyników

- Wszystkie zwracają prawidłowy tag: **82,78%**
  unikam fin:sg:pri:imperf

  fin:sg:pri:imperf+ fin:sg:pri:imperf+ fin:sg:pri:imperf+ fin:sg:pri:imperf+
- Większość zwraca prawidłowy tag: **7,95%**
  kapitalistów subst:pl:gen:m1

  subst:pl:gen:m1+ subst:pl:gen:m1+ subst:pl:gen:m1+ subst:pl:acc:m1-
- Równowaga w głosowaniu: **2,71%**
  powolny adj:sg:nom:m3:pos

  adj:sg:nom:m3:pos+ adj:sg:nom:m3:pos+ adj:sg:acc:m3:pos- adj:sg:acc:m3:pos-
- Prawidłowy tag w mniejszości: **2,38%**

  twarzy subst:sg:loc:f subst:sg:gen:f- subst:sg:gen:f- subst:sg:gen:f- subst:sg:loc:f+
- Wszystkie się mylą: **4.18%**
  biurka subst:pl:nom:n subst:pl:acc:n- subst:pl:acc:n- subst:sg:gen:n- subst:pl:acc:n-
  (Peggy) McCreary subst:sg:nom:f

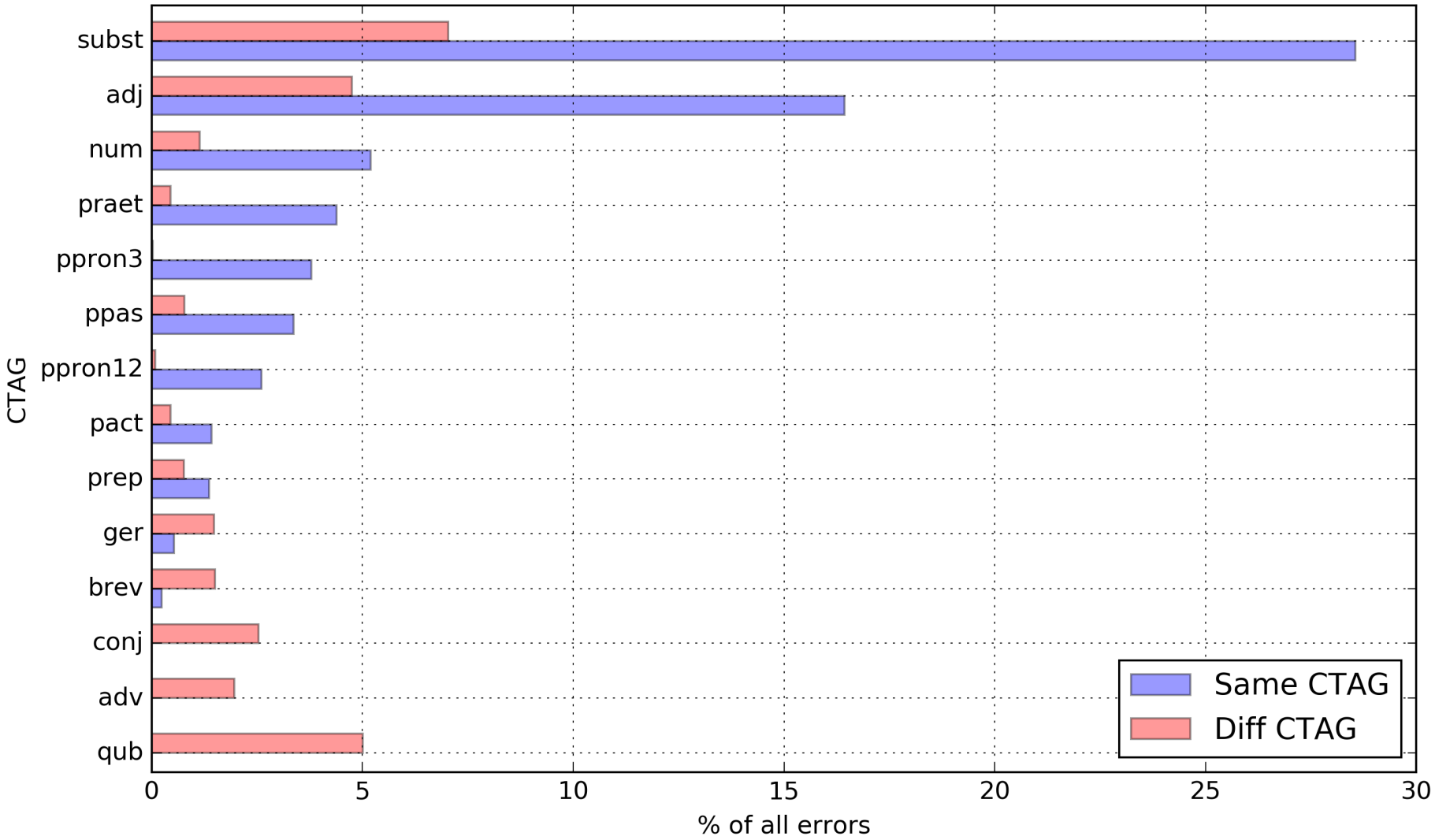
  subst:sg:gen:f- subst:sg:gen:n- subst:sg:nom:n- subst:sg:acc:m1-

Podział na klasy gramatyczne

		Acc_{lower} (%)			
klasa	liczność	PANTERA	WMBT	WCRFT	Concraft
subst	331570	85,21	86,25	87,36	88,29
interp	223542	99,63	99,97	99,97	99,97
adj	128703	76,53	81,10	81,56	82,52
prep	115818	97,04	97,28	97,54	98,05
qub	68079	92,98	93,82	92,91	92,92
fin	59458	98,64	98,70	98,81	98,94
praet	53326	90,90	88,96	89,80	89,69
conj	44840	95,17	95,41	94,61	93,96
adv	42750	95,31	95,59	95,29	94,77
inf	19213	98,91	99,20	99,09	99,14
comp	17842	97,26	97,29	96,84	96,88
num	16160	33,40	56,40	60,32	55,99

Podział na klasy gramatyczne

Concraft: błąd wyboru klasy vs błąd w ramach klasy (test: 100k)



Najczęstsze błędy

Concraft: najczęstsze błędy wyboru klasy gramatycznej

Concraft	NKJP	liczność	% błędów
adj	subst	199	1.9422
conj	qub	178	1.7373
subst	adj	162	1.5811
adv	qub	159	1.5518
subst	ger	152	1.4835
qub	conj	151	1.4737
subst	brev	140	1.3664
ger	subst	128	1.2493
num	adj	108	1.0541
ppas	adj	108	1.0541
qub	adv	91	0.8882
adj	ppas	71	0.6930
adj	num	71	0.6930

Najczęstsze błędy

Concraft: najczęstsze błędy doboru tagu w ramach klasy `subst`

Concraft	NKJP	liczność	% błędów
sg:nom:m3	sg:acc:m3	191	1.8641
sg:acc:m3	sg:nom:m3	153	1.4933
sg:acc:n	sg:nom:n	134	1.3078
sg:nom:n	sg:acc:n	117	1.1419
pl:nom:m3	pl:acc:m3	89	0.8686
pl:acc:f	pl:nom:f	78	0.7613
pl:nom:f	pl:acc:f	71	0.6930
pl:acc:m3	pl:nom:m3	68	0.6637
sg:gen:m1	sg:acc:m1	67	0.6539
sg:acc:m1	sg:gen:m1	53	0.5173
pl:nom:n	pl:acc:n	48	0.4685
sg:gen:f	pl:gen:f	47	0.4587

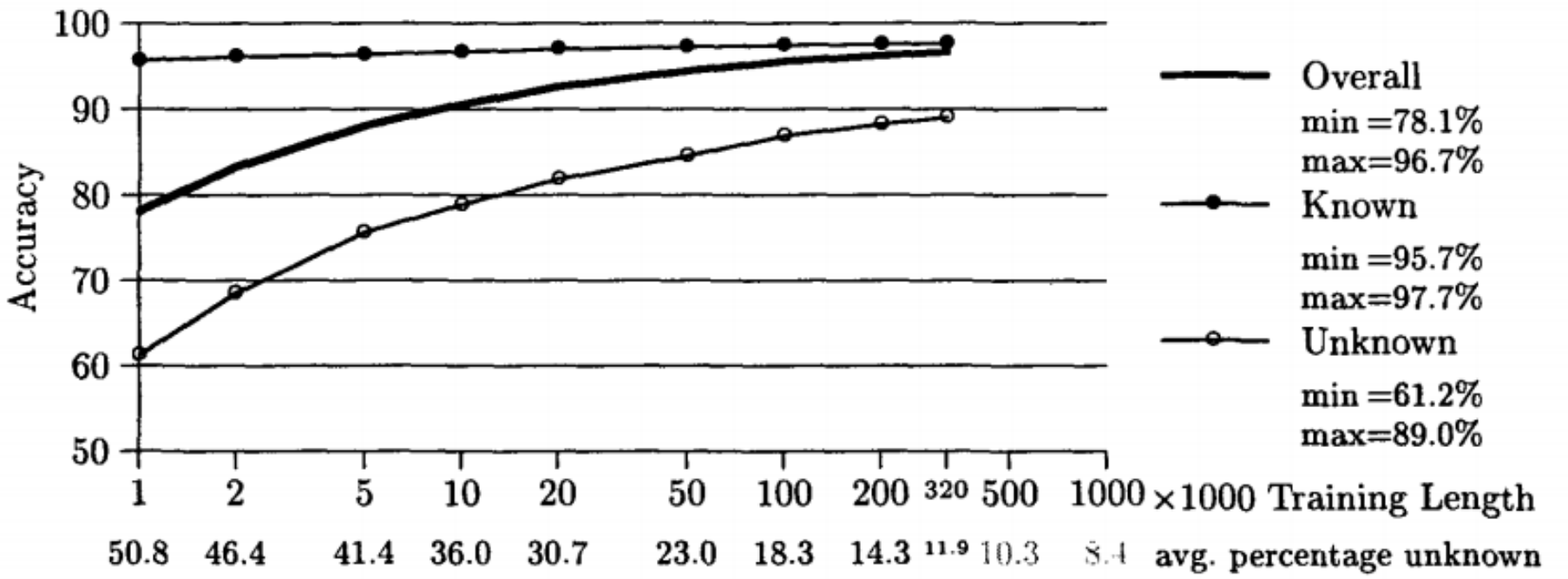
Najczęstsze błędy

Concraft: najczęstsze błędy doboru tagu w ramach klasy `adj`

Concraft	NKJP	liczność	% błędów
sg:nom:m3:pos	sg:acc:m3:pos	90	0.8784
sg:acc:m3:pos	sg:nom:m3:pos	72	0.7027
sg:nom:m1:pos	sg:nom:m3:pos	57	0.5563
sg:nom:n:pos	sg:acc:n:pos	51	0.4978
pl:nom:m3:pos	pl:acc:m3:pos	49	0.4782
pl:nom:f:pos	pl:acc:f:pos	46	0.4490
pl:acc:f:pos	pl:nom:f:pos	44	0.4294
sg:nom:m3:pos	sg:nom:m1:pos	42	0.4099
sg:acc:n:pos	sg:nom:n:pos	32	0.3123
pl:acc:m3:pos	pl:nom:m3:pos	28	0.2733
pl:nom:n:pos	pl:acc:n:pos	27	0.2635
pl:nom:m3:pos	pl:nom:f:pos	25	0.2440
pl:acc:n:pos	pl:nom:n:pos	21	0.2050

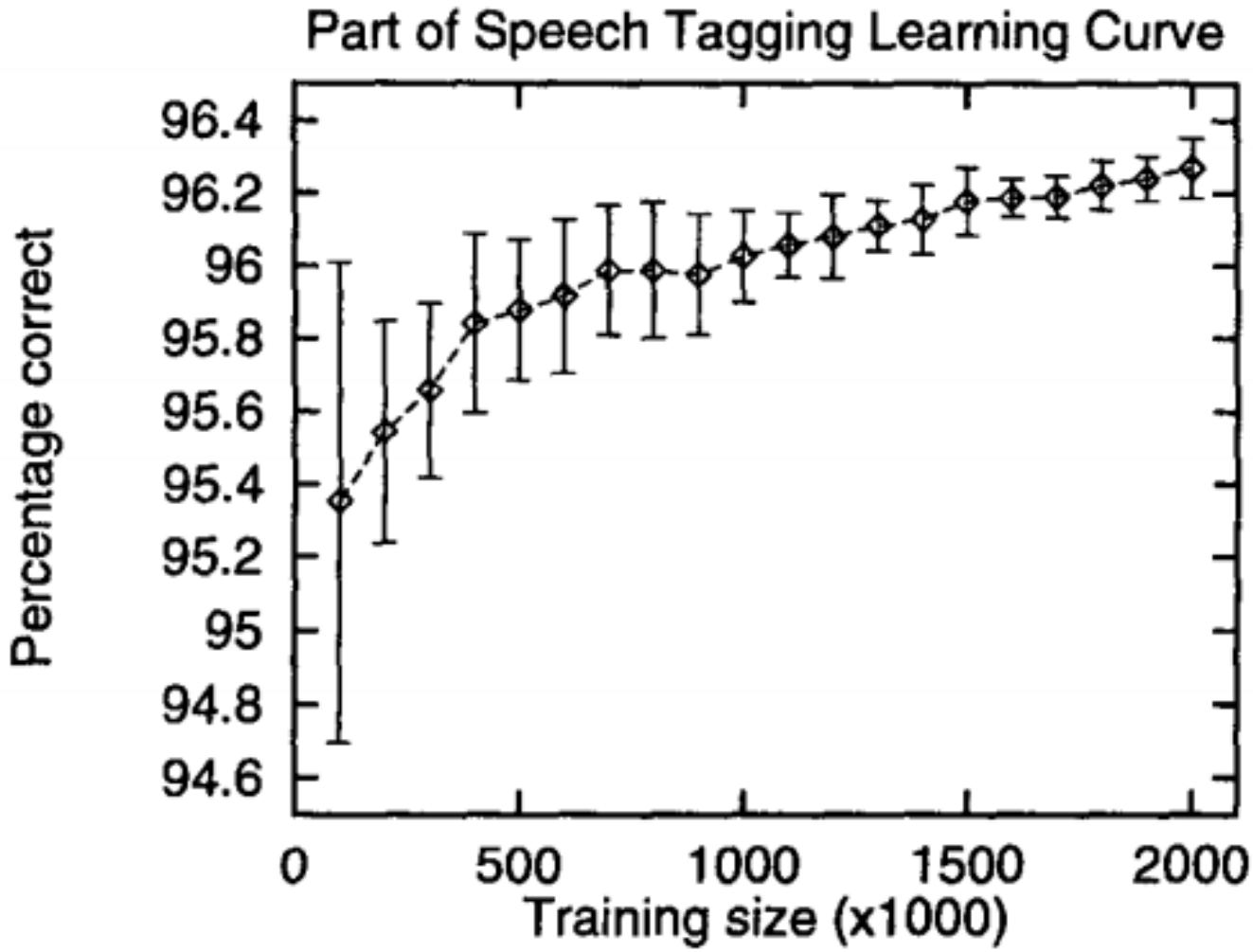
Rozmiar danych treningowych

TnT Tagger – NEGRA corpus, 30 000 tokenów testowych.



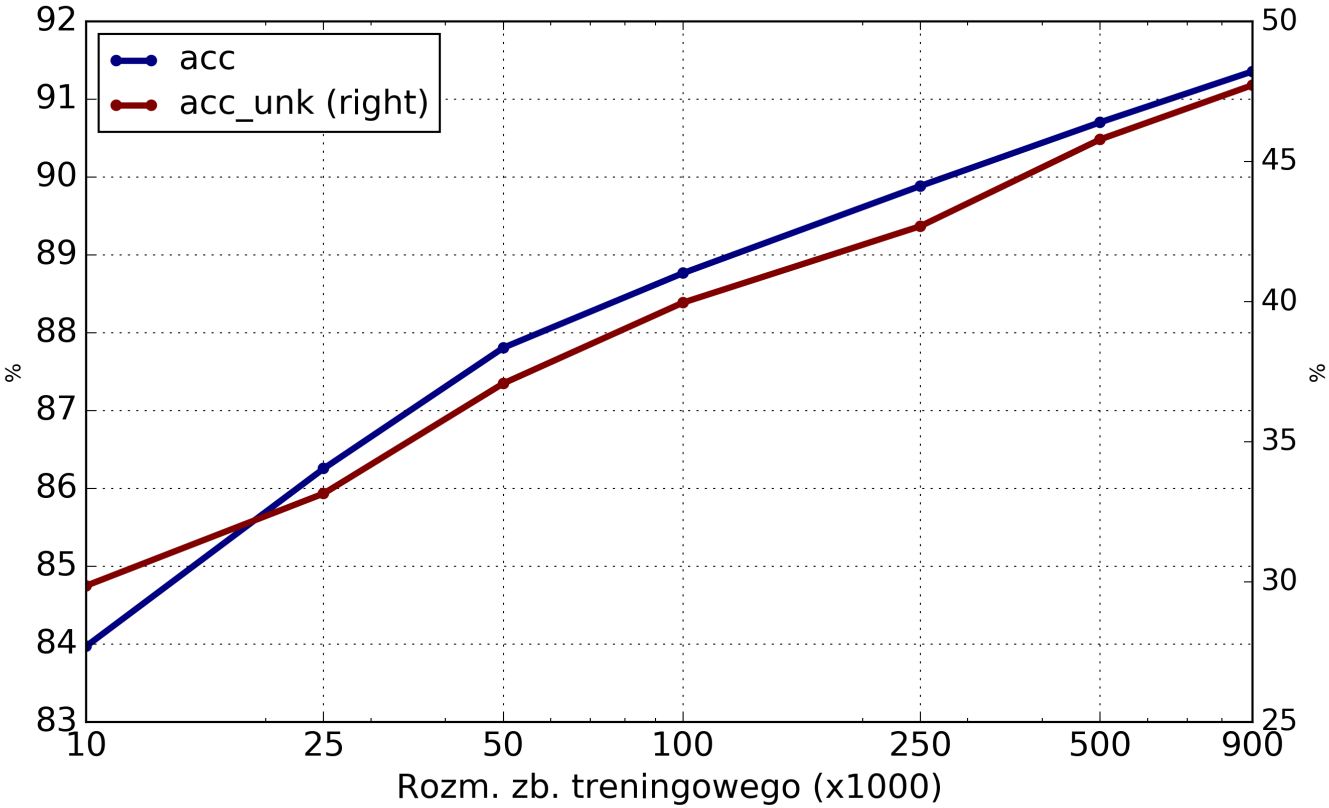
Rozmiar danych treningowych

MBT Tagger – WSJ corpus, kroswalidacja krzyżowa



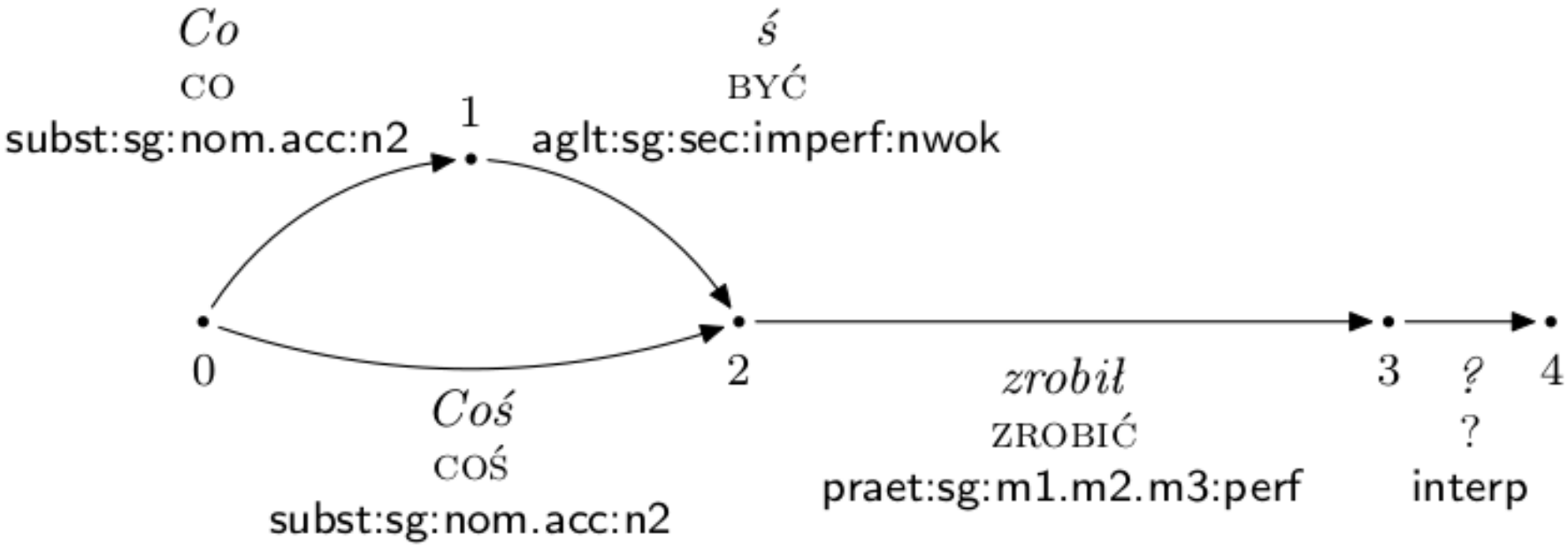
Rozmiar danych treningowych

Concraft – NKJP 1M, 100 000 tokenów testowych.



Problem niejednoznaczności segmentacji

Na czym polega problem?



Problem niejednoznaczności segmentacji

Jak często występują niejednoznaczności?

Korpus NKJP 1M, Morfeusz 1 SGJP:

645 wystąpień na 1 095 118 segmentów (0.0589%)

kiedyś	234 / 1
gdzieś	172 / 1
miałem	99 / 98
udziałem	40 / 0
musiałem	28 / 28
sms-a	6 / 0
działam	6 / 0
doń	5 / 4
tyłem	4 / 0
pis-em	4 / 0
podziałem	3 / 0
piekłem	3 / 0
czekałem	3 / 3
jadłem	3 / 3

pis-owi	3
winnym	2
prl-em	2
wyłom	2
rozdziałem	2
hiv-em	2
pit-ów	2
działem	1
rop-em	1
tir-a	1
kor-owcy	1
urm-em	1
kor-em	1
dj-a	1

ipn-em	1
sms-ów	1
pgr-ach	1
zus-em	1
vat-em	1
siadłem	1
msz-ów	1
zoz-ów	1
mosir-em	1
vip-om	1
msz-ecie	1
zoz-owi	1
czemuś	1

Problem niejednoznaczności segmentacji

Jak często występują niejednoznaczności?

Próbka 100M NKJP, Morfeusz 1 SGJP:

40 354 wystąpień na 101 052 527 segmentów (0.0399%)

kiedyś	12751	czemuś	171	pit-ów	65
miałem	8350	działam	153	Łks-em	60
gdzieś	6171	działem	151	pis-em	44
udziałem	4988	piekłem	130	siadłem	43
musiałem	2173	zus-em	95	skok-i	39
czekałem	537	sms-em	95	gks-em	39
tyłem	523	tir-ów	93	padłem	36
doń	414	zoz-ów	86	pgr-ów	30
podziałem	411	tir-a	85	vip-a	30
sms-a	357	zus-owi	82	pis-owi	28
vip-ów	305	azs-em	81	skok-ów	24
winnym	256	vat-em	76	pks-em	24
sms-ów	207	rozdziałem	76	dj-ów	20
jadłem	199	wyłom	65	dj-e	20

Problem niejednoznaczności segmentacji

Jak często występują niejednoznaczności?

Korpus NKJP 1M, Morfeusz 2:

2583 wystąpień na 1 095 118 segmentów (0.2359%)

coś	777
ktoś	382
czym	334
kiedyś	234
gdzieś	172
miałem	99
czegoś	97
kogoś	82
czymś	63
kimś	42
udziałem	40

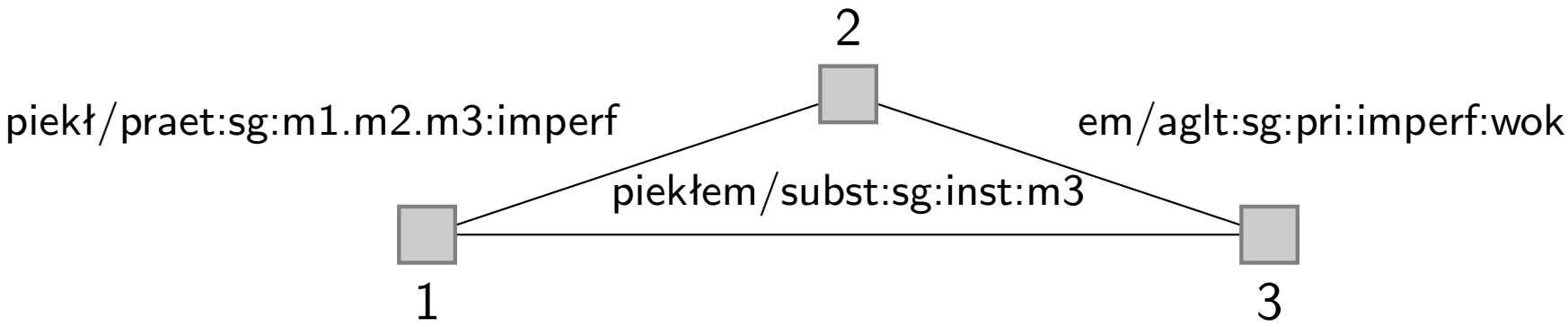
komuś	31
tom	19
działam	6
jam	5
doń	5
oścież	5
tyłem	4
musiałem	3
podziałem	3
piekłem	3
czekałem	3

jadłem	3
rozdziałem	2
wyłom	2
bom	2
coście	1
czyżbyś	1
czyżem	1
siadłem	1
czemuś	1
działem	1

Problem niejednoznaczności segmentacji

Możliwe rozwiązania: tagset pośredni (A. Radziszewski).

Wprowadźmy tagset pośredni, który pozwoli uniknąć części niejednoznaczności



piekłem

piec fin:sg:m1:pri:imperf:prt
piec fin:sg:m2:pri:imperf:prt
piec fin:sg:m3:pri:imperf:prt
piekło subst:sg:inst:n

Problem niejednoznaczności segmentacji

Możliwe rozwiązania: tagset pośredni

Rozwiązanie analogiczne do trybu `-p composite` w Morfeuszu 2:

```
$ echo piekłem — morfeusz_analyzer
[0,1,piekł,piec:v,praet:sg:m1.m2.m3:imperf,_,_]
[0,2,piekłem,piekło,subst:sg:inst:n2,pospolita,_]
[1,2,em,być,aglt:sg:pri:imperf:wok,_,_]

$ echo piekłem — morfeusz_analyzer -p composite
[0,1,piekłem,piec:v,praet:sg:m1.m2.m3:pri:imperf,_,_
0,1,piekłem,piekło,subst:sg:inst:n2,pospolita,_]
```

Problem niejednoznaczności segmentacji

Możliwe rozwiązania: dostosowanie tagera do przetwarzania DAGów

Na przykład, dla tagera Concraft: konieczna jest modyfikacja całego stosu przetwarzania:

- modyfikacja istniejących formatów zapisu korpusów (XML, Plain),
- modyfikacja narzędzi: Corpus2, Maca,
- reimplementacja algorytmu w tagerze, aby był w stanie przetwarzać dane w reprezentacji grafowej.

Poziom pewności ujednoznaczniania morfosyntaktycznego

Oczekujemy, że: tager informuje o poziomie pewności podjętych decyzji

Tagery oparte na metodach uczenia maszynowego mogą zwracać prawdopodobieństwa brzegowe wyboru poszczególnych interpretacji.

Implementacja w Concraft:

```
$ ~/.cabal/bin/concraft-pl tag -m model.gz < test.plain
```

rzucała space
    rzucać praet:sg:f:imperf 1.000
światło space
    światło adv:pos 0.000
    światło subst:sg:acc:n 0.929
    światło subst:sg:nom:n 0.071
    światło subst:sg:voc:n 0.000

tylko space
    Tylka subst:sg:voc:f 0.000
    tylko conj 0.209
    tylko qub 0.791
na space
    na interj 0.000
    na prep:acc 1.000
    na prep:loc 0.000
podłogę space
    podłoga subst:sg:acc:f 1.000

Tagery języka polskiego – analiza jakościowa

Tagowanie w NKJP

W NKJP300 zapytanie:

- `[pos=prep & base!=temu][case=voc]`: 5879 wyników.
- `[pos=prep][pos="praet|fin|inf|impt"]`: 88 271 wyników.
- `[pos=prep & base!=temu & case=$1][pos=adj & base!=który & case!=$1][pos=subst & case=$1]`: 798 794 wyników.
- `[pos=prep & base!=temu & case=$1][pos=subst & case!=$1]`: 5 770 963 wyników.

Tagowanie w NKJP

W NKJP300 zapytanie:

- `[pos=prep & base!=temu][case=voc]`: 5879 wyników.
- `[pos=prep][pos="praet|fin|inf|impt"]`: 88 271 wyników.
- `[pos=prep & base!=temu & case=$1][pos=adj & base!=który & case!=$1][pos=subst & case=$1]`: 798 794 wyników.
- `[pos=prep & base!=temu & case=$1][pos=subst & case!=$1]`: 5 770 963 wyników.

Tagowanie w NKJP

W NKJP300 zapytanie:

- `[pos=prep & base!=temu][case=voc]`: 5879 wyników.
- `[pos=prep][pos="praet|fin|inf|impt"]`: 88 271 wyników.
- `[pos=prep & base!=temu & case=$1][pos=adj & base!=który & case!=$1][pos=subst & case=$1]`: 798 794 wyników.
- `[pos=prep & base!=temu & case=$1][pos=subst & case!=$1]`: 5 770 963 wyników.

Tagowanie w NKJP

W NKJP300 zapytanie:

- `[pos=prep & base!=temu][case=voc]`: 5879 wyników.
- `[pos=prep][pos="praet|fin|inf|impt"]`: 88 271 wyników.
- `[pos=prep & base!=temu & case=$1][pos=adj & base!=który & case!=$1][pos=subst & case=$1]`: 798 794 wyników.
- `[pos=prep & base!=temu & case=$1][pos=subst & case!=$1]`: 5 770 963 wyników.

Tagowanie w NKJP

W NKJP300 zapytanie:

- `[pos=prep & base!=temu][case=voc]`: 5879 wyników.
- `[pos=prep][pos="praet|fin|inf|impt"]`: 88 271 wyników.
- `[pos=prep & base!=temu & case=$1][pos=adj & base!=który & case!=$1][pos=subst & case=$1]`: 798 794 wyników.
- `[pos=prep & base!=temu & case=$1][pos=subst & case!=$1]`: 5 770 963 wyników.

Podstawa analizy

Tagery Pantera, Concraft, WCRF i WMBT (w dwu wersjach) trenowane na 90% NKJP1M.

Analiza została przeprowadzona na pozostałych 10% NKJP1M (ok. 120 tys. segmentów) oznakowanych czterema tagerami w zestawieniu z ręcznym znakowaniem wzorcowym.

Różne tagsety dla języka polskiego?

Obecnie funkcjonują równolegle dwa tagsety języka polskiego

- tagset NKJP, używany do anotacji korpusu, a także w większości innych zasobów językowych,
- tagset Morfeusza.

Skutkuje to sytuacją, w której w sposób niejawny dokonywana jest ciągła konwersja pomiędzy tagsetami:

```
tagset_from=sgjp
tagset_to=nkjp
override=n1:n
override=n2:n
override=n3:n
override=p1:m1
override=p2:n
override=p3:n
```

```
tagset_from=morfeusz2; tagset_to=nkjp
override=dig:num; override=nie:conj
override=romandig:num
override=prefa:ign
override=prefppas:ign
override=prefs:ign; override=prefv:ign
override=naj:ign; override=cond:ign
override=substa:ign
```

Dwa tagsety, dwa opisy

Różnice słownikowe

- Brak interpretacji `adv` dla: JESZCZE, ZNÓW, ZNOWU, JUŻ, WRESZCIE.
- Brak interpretacji `conj` dla: DOŚĆ, JESZCZE, RÓWNIEŻ, TAKŻE, TEŻ.
- Brak interpretacji `qub` dla: ABSOLUTNIE, GENERALNIE, GŁÓWNIE, JAK, JAKOŚ, PRAKTYCZNIE, PRAWDOPODOBNIE, PRZYPUSZCZALNIE, RZECZYWIŚCIE, RZEKOMO, SZCZEGÓLNIE, WŁAŚCIWIE, WYŁĄCZNIE.
- Brak interpretacji `prep` dla: BLISKO, BLIŻEJ, NAJBLIŻEJ, WYJĄWSZY, APROPOS, CELEM, X. . .

Inne różnice

- Formy *bliska* i *swojemu* w wyrażeniach *z bliska* i *po swojemu* w NKJP są `adjp`, w Morfeuszu zaś — `adj`.
- PÓŁTORA w NKJP to wyłącznie `subst`, w Morfeuszu — wyłącznie `num`.
- Liczebniki PÓŁ i ĆWIERĆ w NKJP wymają liczby mnogiej, w Morfeuszu 2.0 — pojedynczej.
- `Burk` w NKJP nie odpowiadają tym w Morfeuszu, np. *łupnia*, *zamian*, *przemian*, *ciemku*, *jaw* są tylko formami rzeczowników. Z drugiej strony *Burkina* i *Faso* w NKJP to rzeczowniki, w Morfeuszu — burkinostki, *propos* w NKJP to kublik lub przyimek, w Morfeuszu — burkinostka, itd.

Najczęściej mylone znaczniki

- subst:sg:acc:m3 vs. subst:sg:nom:m3
- conj vs. qub
- subst:sg:acc:n vs. subst:sg:nom:n
- adj:sg:acc:m3:pos vs. adj:sg:nom:m3:pos
- subst:pl:acc:m3 vs. subst:pl:nom:m3
- subst:pl:acc:f vs. subst:pl:nom:f
- adv vs. qub
- praet:sg:m1:perf vs. praet:sg:m3:perf
- praet:sg:m1:imperf vs. praet:sg:m3:imperf
- subst:sg:acc:m1 vs. subst:sg:gen:m1
- subst:pl:nom:f vs. subst:sg:gen:f
- adj:sg:nom:m1:pos vs. adj:sg:nom:m3:pos
- prep:acc vs. prep:loc

Najczęściej mylone znaczniki

- subst:sg:acc:m3 vs. subst:sg:nom:m3
- conj vs. qub
- subst:sg:acc:n vs. subst:sg:nom:n
- adj:sg:acc:m3:pos vs. adj:sg:nom:m3:pos
- subst:pl:acc:m3 vs. subst:pl:nom:m3
- subst:pl:acc:f vs. subst:pl:nom:f
- adv vs. qub
- praet:sg:m1:perf vs. praet:sg:m3:perf
- praet:sg:m1:imperf vs. praet:sg:m3:imperf
- subst:sg:acc:m1 vs. subst:sg:gen:m1
- subst:pl:nom:f vs. subst:sg:gen:f
- adj:sg:nom:m1:pos vs. adj:sg:nom:m3:pos
- prep:acc vs. prep:loc

Najczęściej mylone znaczniki

- subst:sg:acc:m3 vs. subst:sg:nom:m3
- conj vs. qub
- subst:sg:acc:n vs. subst:sg:nom:n
- adj:sg:acc:m3:pos vs. adj:sg:nom:m3:pos
- subst:pl:acc:m3 vs. subst:pl:nom:m3
- subst:pl:acc:f vs. subst:pl:nom:f
- adv vs. qub
- praet:sg:m1:perf vs. praet:sg:m3:perf
- praet:sg:m1:imperf vs. praet:sg:m3:imperf
- subst:sg:acc:m1 vs. subst:sg:gen:m1
- subst:pl:nom:f vs. subst:sg:gen:f
- adj:sg:nom:m1:pos vs. adj:sg:nom:m3:pos
- prep:acc vs. prep:loc

Najczęściej mylone znaczniki

- subst:sg:acc:m3 vs. subst:sg:nom:m3
- conj vs. qub
- subst:sg:acc:n vs. subst:sg:nom:n
- adj:sg:acc:m3:pos vs. adj:sg:nom:m3:pos
- subst:pl:acc:m3 vs. subst:pl:nom:m3
- subst:pl:acc:f vs. subst:pl:nom:f
- adv vs. qub
- praet:sg:m1:perf vs. praet:sg:m3:perf
- praet:sg:m1:imperf vs. praet:sg:m3:imperf
- subst:sg:acc:m1 vs. subst:sg:gen:m1
- subst:pl:nom:f vs. subst:sg:gen:f
- adj:sg:nom:m1:pos vs. adj:sg:nom:m3:pos
- prep:acc vs. prep:loc

Homonimia

przekładała	**paczki** [paczek:subst:pl:acc:m3] z ręki do ręki
W Krakowie, w	**krypcie** [krypeć:subst:pl:acc:m3] pod kościołem księży pijarów
pokrywa śnieżna sięga pół	**metra** [metro:subst:sg:gen:n] .
start na 200	**metrów** [metr:subst:pl:acc:m1] stylem klasycznym
mundurach z czerwonymi	**kitami**[kit:subst:pl:inst:m3] na czakach
częstować się wigilijnymi	**potrawami** [potraw:subst:pl:inst:m3] .
o	**systemie** [systema:subst:sg:loc:f] oświaty
informowaliśmy o	**związkach** [związka:subst:pl:loc:f] zawodowych
rozwieszała w	**ogrodzie** [ogroda:subst:sg:loc:f] bieliznę
dwudziestu pięciu	**wierszach** [wiersza:subst:pl:loc:f] poematu Lukrecjusza
wydatków na budowę i	**remonty** [remonta:subst:sg:gen:f] kaplic
Zdradzisz nam	**sekrety** [sekreta:subst:pl:acc:f] urody
zatrzaskuje lodówkę pełną	**puszek**[puszek:subst:sg:acc:m3] z filmem
Gonią mnie	**potwory** [potwora:subst:sg:gen:f]
zapadnij w	**moczary** [moczara:subst:sg:gen:f] na lewym brzegu
Na	**karcie** [kart:subst:sg:loc:m3] do głosowania drukuje się odcisk
stoi sobie w	**paletach** [paleto:subst:pl:loc:n] pod folią
Skryły nas	**liście** [liście:subst:sg:acc:n] i iluzja niewidzialności
w ekipie	**gości** [gościa:subst:pl:gen:f] zapanowała zrozumiała euforia

Homonimia

przekładała	**paczki** [paczek:subst:pl:acc:m3] z ręki do ręki
W Krakowie, w	**krypcie** [krypeć:subst:pl:acc:m3] pod kościołem księży pijarów
pokrywa śnieżna sięga pół	**metra** [metro:subst:sg:gen:n] .
start na 200	**metrów** [metr:subst:pl:acc:m1] stylem klasycznym
mundurach z czerwonymi	**kitami** [kit:subst:pl:inst:m3] na czakach
częstować się wigilijnymi	**potrawami** [potraw:subst:pl:inst:m3] .
o	**systemie** [systema:subst:sg:loc:f] oświaty
informowaliśmy o	**związkach** [związka:subst:pl:loc:f] zawodowych
rozwieszała w	**ogrodzie** [ogroda:subst:sg:loc:f] bieliznę
dwudziestu pięciu	**wierszach** [wiersza:subst:pl:loc:f] poematu Lukrecjusza
wydatków na budowę i	**remonty** [remonta:subst:sg:gen:f] kaplic
Zdradzisz nam	**sekrety** [sekreta:subst:pl:acc:f] urody
zatrzaskuje lodówkę pełną	**puszek** [puszek:subst:sg:acc:m3] z filmem
Gonią mnie	**potwory** [potwora:subst:sg:gen:f]
zapadnij w	**moczary** [moczara:subst:sg:gen:f] na lewym brzegu
Na	**karcie** [kart:subst:sg:loc:m3] do głosowania drukuje się odcisk
stoi sobie w	**paletach** [paleto:subst:pl:loc:n] pod folią
Skryły nas	**liście** [liście:subst:sg:acc:n] i iluzja niewidzialności
w ekipie	**gości** [gościa:subst:pl:gen:f] zapanowała zrozumiała euforia

Homonimia

palet

rzeczownik *daw.* [SJPDor.]
m3 0014u

	l. p.	l. m.
M.	palet	palety
D.	paletu	paletów
C.	paletowi	paletom
B.	palet	palety
N.	paletem	paletami
Ms.	palecie	paletach
W.	palecie	palety

paleta

rzeczownik [SJPDor.]
ż 0099

	l. p.	l. m.
M.	paleta	palety
D.	palety	palet
C.	palecie	paletom
B.	paletę	palety
N.	paletą	paletami
Ms.	palecie	paletach
W.	paleto	palety

paleto *'palto'*

rzeczownik *daw.*
n2 0184

	l. p.	l. m.
M.	paleto	paleta
D.	paleta	palet
C.	paletu	paletom
B.	paleto	paleta
N.	paletem	paletami
Ms.	palecie	paletach
W.	paleto	paleta

Homonimy męskie

oceny prawidłowości odstrzału	**kozła** [subst:sg:gen:m3] jest wiek
pasjonuje się nie	**wężami** [subst:pl:inst:m3], lecz psami
od półtora do pięciu tysięcy	**bojowników** [subst:pl:gen:m2] czeczeńskich
ocenia się długość życia	**gwarków** [subst:pl:gen:m3] na piętnaście
do pełnienia funkcji	**kapelanów** [subst:pl:gen:m2] wojskowych
z dwójką takich wiesz	**maluchów** [subst:pl:gen:m2] .
pomagała przy doręczaniu	**klientom** [subst:pl:dat:m2] towarów
koncert folkowo-rockowej Kapeli —	**górali** [subst:pl:gen:m2] z Żywca
Ojciec mojej przyjaciółki był	**admirałem** [subst:sg:inst:m2], mąż kontradmirałem
przechodził ten	**bokser** [subst:sg:acc:m3] z drugiego piętra
ten wątek jest jak	**smok** [subst:sg:nom:m3] z głowami , które odrastają

Wahliwość rodzajowa

te	**grzyby** [subst:pl:nom:m3] też nie są osolone
kieliszek do ust i poi go	**szampanem** [subst:sg:inst:m3] . To dlatego nigdy
wyrwano Panu	**zęba** [subst:sg:gen:m3], że o operacji nie
nie wyjmując	**papierosa** [subst:sg:gen:m3] z ust
rosyjskich	**śmigłowców** [subst:pl:gen:m2] doczekało się reakcji
crepes,	**naleśniki** [subst:pl:nom:m2] , z nadzwyczajną mieszaniną
o niebieskich	**migdałach** [subst:pl:loc:m2]
dekorowania jej plasterkami	**ogórka** [subst:sg:acc:m2] i jajka
założył mu błyskawicznie	"**nelsona**" [subst:sg:gen:m3] i zanim

Segmenty nierozpoznane I

Rzeczowniki w NKJP1M:

Segmenty nierozpoznane II

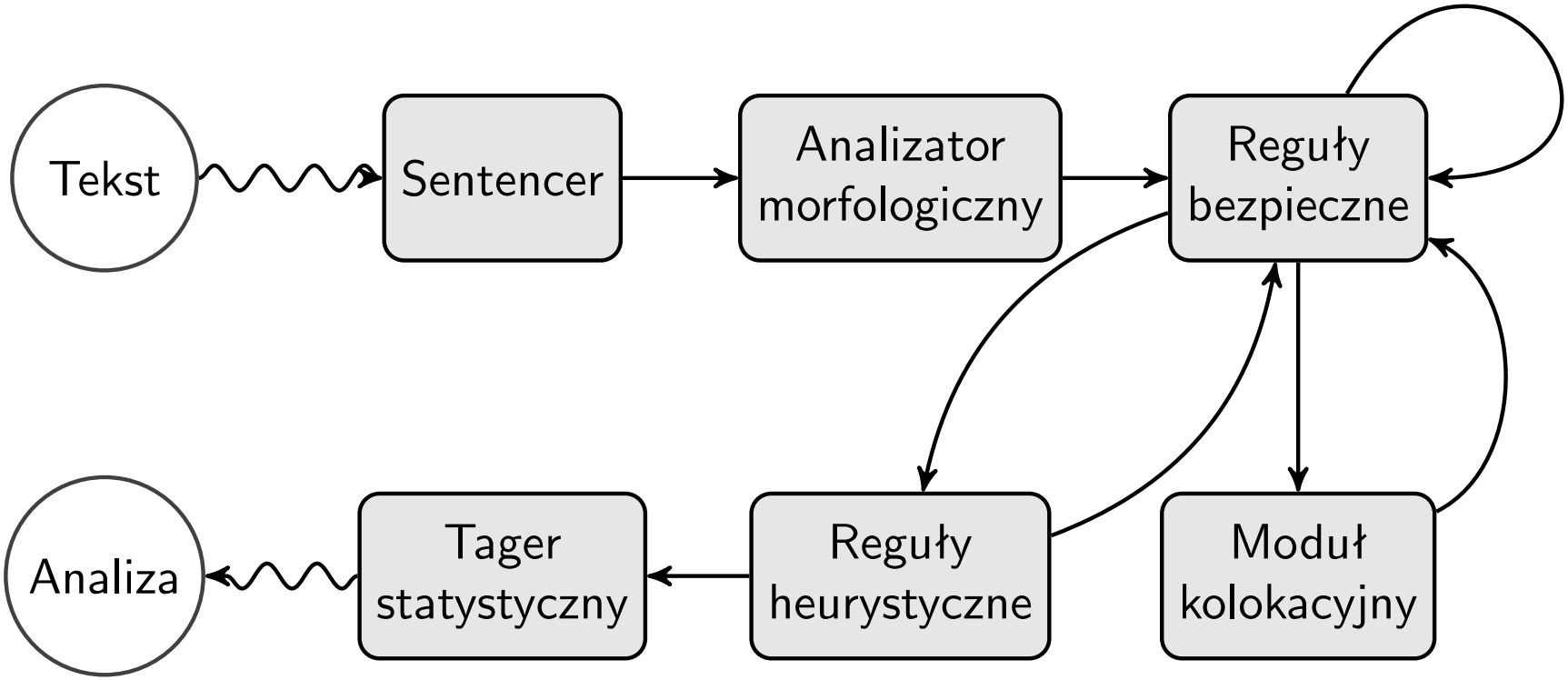
Ok. 25% błędów tagerów bez zgadywaczy stanowią słowa nierozpoznane. Wśród nich zarysowują się cztery wyraźne grupy:

- rzeczowniki rodzaju M1 — nazwiska i imiona męskie,
- rzeczowniki rodzajów M3, F i N — nazwy własne geogr., nazwy firm, organizacji itp., skrótowce,
- przymiotniki i liczebniki — zapisy liczbowe,
- skróty.

Ogółem Morfeusz 2.0 w NKJP1M nie rozpoznaje:

- 10 148 rzeczowników,
- 460 przymiotników (nie licząc zapisów liczbowych),
- 90 czasowników,
- 20 liczebników (nie licząc zapisów liczbowych).

Jak to się robi w Czechach? I



Jak to się robi w Czechach? II

Komponenty:

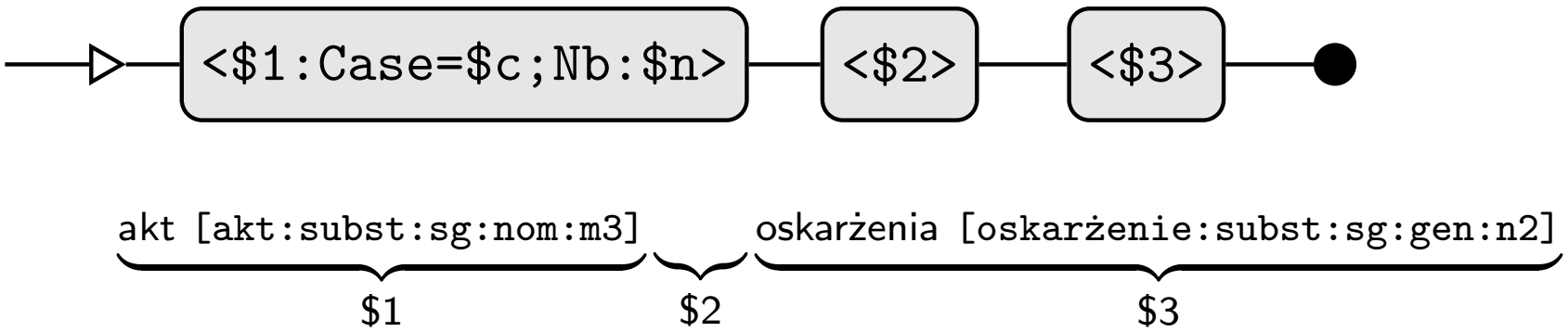
- obszerny słownik (ok. 800 000 haseł, w tym ok. 200 000 nazw własnych),
- duży zestaw reguł (ok. 2600) tworzonych ręcznie w formalizmie *LanGr*,
- moduł kolokacyjny *Phras*,
- analizatory statystyczne *MorČe* i *MorphoDiTa*.

Możliwe zastępniki

Polskie zasoby, które (potencjalnie) można wykorzystać:

- SGJP (ew. Polimorf), być może rozszerzony o więcej nazw własnych, np. wszystkie nazwiska z bazy PESEL,
- Spejd,
- *Słownik elektroniczny jednostek frazeologicznych* (SEJF),
- *Słownik paradygmatów polskich frazeologizmów czasownikowych* (VERBEL),
- być może również ramki frazeologiczne Walentego (?),
- omówione wcześniej tagery statystyczne.

SEJF



Przykładowe zastosowanie:

- (...) skierowano do sądu **akt** `[akta:subst:pl:gen:n]` oskarżenia.
- (...) odpisy aktu **oskarżenia** `[ger:sg:gen:n:perf:aff]` (...)

SEJF

W NKJP300 jest 4116 ciągów pasujących do zapytania `[base~"akt"] [base~"oskarżenie"]`. Ale tylko 2182 spełnia warunki nałożone na to wyrażenie w SEJF-ie. Reszta to błędy, zwykle polegające na wyborze rzeczownika AKTA.

Podobnie dla `[base~"armia"] [base~"czerwony"]` — 1983 wyników, ale tylko 120 spełnia ograniczenia SEJF-u. Reszta to w większości błędy polegające na wyborze rzeczownika CZERWONA.

Dla ciągu `[base~"brud"] [base="," ] [base~"smród"] [base="i"] [base~"ubóstwo"]` Pantera uparcie wybiera rzeczownik SMRÓD M1. Jedyne poprawne dopasowanie jest w dopełniaczu.

SEJF

W NKJP300 jest 4116 ciągów pasujących do zapytania `[base~"akt"] [base~"oskarżenie"]`. Ale tylko 2182 spełnia warunki nałożone na to wyrażenie w SEJF-ie. Reszta to błędy, zwykle polegające na wyborze rzeczownika AKTA.

Podobnie dla `[base~"armia"] [base~"czerwony"]` — 1983 wyników, ale tylko 120 spełnia ograniczenia SEJF-u. Reszta to w większości błędy polegające na wyborze rzeczownika CZERWONA.

Dla ciągu `[base~"brud"] [base=","] [base~"smród"] [base="i"] [base~"ubóstwo"]` Pantera uparcie wybiera rzeczownik SMRÓD M1. Jedyne poprawne dopasowanie jest w dopełniaczu.

SEJF

W NKJP300 jest 4116 ciągów pasujących do zapytania `[base~"akt"] [base~"oskarżenie"]`. Ale tylko 2182 spełnia warunki nałożone na to wyrażenie w SEJF-ie. Reszta to błędy, zwykle polegające na wyborze rzeczownika AKTA.

Podobnie dla `[base~"armia"] [base~"czerwony"]` — 1983 wyników, ale tylko 120 spełnia ograniczenia SEJF-u. Reszta to w większości błędy polegające na wyborze rzeczownika CZERWONA.

Dla ciągu `[base~"brud"] [base=","] [base~"smród"] [base="i"] [base~"ubóstwo"]` Pantera uparcie wybiera rzeczownik SMRÓD M1. Jedyne poprawne dopasowanie jest w dopełniaczu.

Dyskusja i rekomendacje

Możliwe ulepszenia

- Uspójnienie korpusu treningowego ze słownikiem.
- Zwiększenie ilości danych treningowych.
- Rozszerzanie słownika o nazwy własne.
- Wykorzystanie dodatkowej informacji słownikowej (kwalifikatory, pospolitość).
- Wykorzystanie innych zasobów lingwistycznych obejmujących np. frazeologię.
- Zastosowanie ręcznie pisanych reguł — przynajmniej do kontroli jakości znakowania.

Struktura danych treningowych w NKJP1M

Kategoria	Udział w korpusie
Dzienniki	25,5%
Pozostałe periodyki	23,5%
Książki publicystyczne	1,0%
Literatura piękna	16,0%
Literatura faktu	5,5%
Typ informacyjno-poradnikowy	5,5%
Typ naukowo-dydaktyczny	2,0%
Internetowe interaktywne (blogi, fora, Usenet)	3,5%
Internetowe nieinteraktywne (statyczne strony, Wikipedia)	3,5%
Quasi-mówione (protokoły sesji parlamentu)	2,5%
Mówione medialne	2,5%
Mówione konwersacyjne	5,0%
Inne teksty pisane	3,0%
Książka niebeletrystyczna nieklasyfikowana	1,0%

Podsumowanie

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- nie jest nadgorliwy, można kazać zostawić interpretacje częściowo nieujednoznacznione (np. usunąć tylko bardzo złe interpretacje),
  obecnie: Pantera, Concraft; po zmianie parametru: WCRFT, WMBT,
- informuje o poziomie pewności podjętych decyzji,
  obecnie: Concraft, możliwe: WCRFT, WMBT; trudniejsze dla Pantery
- działa na niejednoznacznej segmentacji (stworzonej przez Morfeusza lub np. będącej wynikiem zastosowania po Morfeuszu słownika wyrażeń wieloczłonowych),
  obecnie: tylko tagset pośredni, możliwe: w każdym po modyfikacjach całego stosu,

Podsumowanie

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- nie jest nadgorliwy, można kazać zostawić interpretacje częściowo nieujednoznacznione (np. usunąć tylko bardzo złe interpretacje),
  obecnie: Pantera, Concraft; po zmianie parametru: WCRFT, WMBT,
- informuje o poziomie pewności podjętych decyzji,
  obecnie: Concraft, możliwe: WCRFT, WMBT; trudniejsze dla Pantery
- działa na niejednoznacznej segmentacji (stworzonej przez Morfeusza lub np. będącej wynikiem zastosowania po Morfeuszu słownika wyrażeń wieloczłonowych),
  obecnie: tylko tagset pośredni, możliwe: w każdym po modyfikacjach całego stosu,

Podsumowanie

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- nie jest nadgorliwy, można kazać zostawić interpretacje częściowo nieujednoznacznione (np. usunąć tylko bardzo złe interpretacje),
  obecnie: Pantera, Concraft; po zmianie parametru: WCRFT, WMBT,
- informuje o poziomie pewności podjętych decyzji,
  obecnie: Concraft, możliwe: WCRFT, WMBT; trudniejsze dla Pantery
- działa na niejednoznacznej segmentacji (stworzonej przez Morfeusza lub np. będącej wynikiem zastosowania po Morfeuszu słownika wyrażeń wieloczłonowych),
  obecnie: tylko tagset pośredni, możliwe: w każdym po modyfikacjach całego stosu,

Podsumowanie (2)

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- daje się (względnie?) łatwo zainstalować i uruchomić na wszystkich platformach, na których jest Morfeusz,
  obecnie: żaden; trzeba przygotować dystrybucję binarną Macy pod Windows; Haskell Platform + Concraft lub kompilacja innych tagerów,
- da się rozszerzyć o uwzględnianie informacji o czasie powstania tekstu,
  obecnie: żaden, możliwe: w opartych na ML jako dodatkowa cecha.

Podsumowanie (2)

Cechy pożądanego tagera (M. Woliński)

Chciałbym tager, który:

- daje się (względnie?) łatwo zainstalować i uruchomić na wszystkich platformach, na których jest Morfeusz,
  obecnie: żaden; trzeba przygotować dystrybucję binarną Macy pod Windows; Haskell Platform + Concraft lub kompilacja innych tagerów,
- da się rozszerzyć o uwzględnianie informacji o czasie powstania tekstu,
  obecnie: żaden, możliwe: w opartych na ML jako dodatkowa cecha.

Podziękowania

Podziękowania za sugestie i uwagi dla:

- Adama Radziszewskiego
- Jakuba Waszczuka
- Szymona Acedańskiego

Dziękujemy za uwagę!

Bibliografia I


Acedański, Szymon, 2010.
A morphosyntactic Brill tagger for inflectional languages.
In *Advances in Natural Language Processing*.


Brill, Eric and Jun Wu, 1998.
Classifier combination for improved lexical disambiguation.
In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98. Stroudsburg, PA, USA: Association for Computational Linguistics.


Śniatowski, Tomasz and Maciej Piasecki, 2012.
Combining Polish morphosyntactic taggers.
In Pascal Bouvry, Mieczysław A. Kłopotek, Franck Leprévost, Małgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybiński (eds.), *Security and Intelligent Information Systems*, volume 7053 of *LNCS*. Springer-Verlag.

Bibliografia II


Radziszewski, Adam, 2013.
A tiered CRF tagger for Polish.
In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka (eds.), *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.


Radziszewski, Adam and Szymon Acedański, 2012.
Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers.
In *Proceedings of TSD 2012*, LNCS. Springer-Verlag.


Radziszewski, Adam and Tomasz Śniatowski, 2011a.
A Memory-Based Tagger for Polish.
In *Proceedings of the LTC 2011*.

Bibliografia III

van Halteren, Hans, Walter Daelemans, and Jakub Zavrel, 2001.
Improving accuracy in word class tagging through the combination of machine learning systems.
Comput. Linguist., 27(2):199–229.

Waszczuk, Jakub, 2012.
Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language.
In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India.