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 zaadresowane przez istniejące tagery,
- rekomendacje dotyczące dalszych kroków.

Plan

1. Tagery języka polskiego – przegląd rozwiązań
2. Tagery języka polskiego – analiza ilościowa
3. Tagery morfostynaktyczne innych języków
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 morfosyntaktycznzy (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

TAGOWANIE Podział na tokeny - Toki Analiza morfosyntaktyczna - Maca + Morfeusz Ujednoznacznianie morfosyntaktyczne

UCZENIE Ponowna analiza morfosyntaktyczna otagowanego tekstu - Toki + Maca + Morfeusz + synchronizacja Uczenie modelu statystycznego na podstawie wzorca

Tagery morfostynaktyczne dla języka polskiego

Tagery historyczne

- tager Łukasza Dębowskiego,
- TaKIPI.

Tagery morfostynaktyczne 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 morfostynaktyczne dla języka polskiego

Modele dla tagerów zaimplementowanych dla innych języków

- TnT Tagger
- OpenNLP

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 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 pojedynczych 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.

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

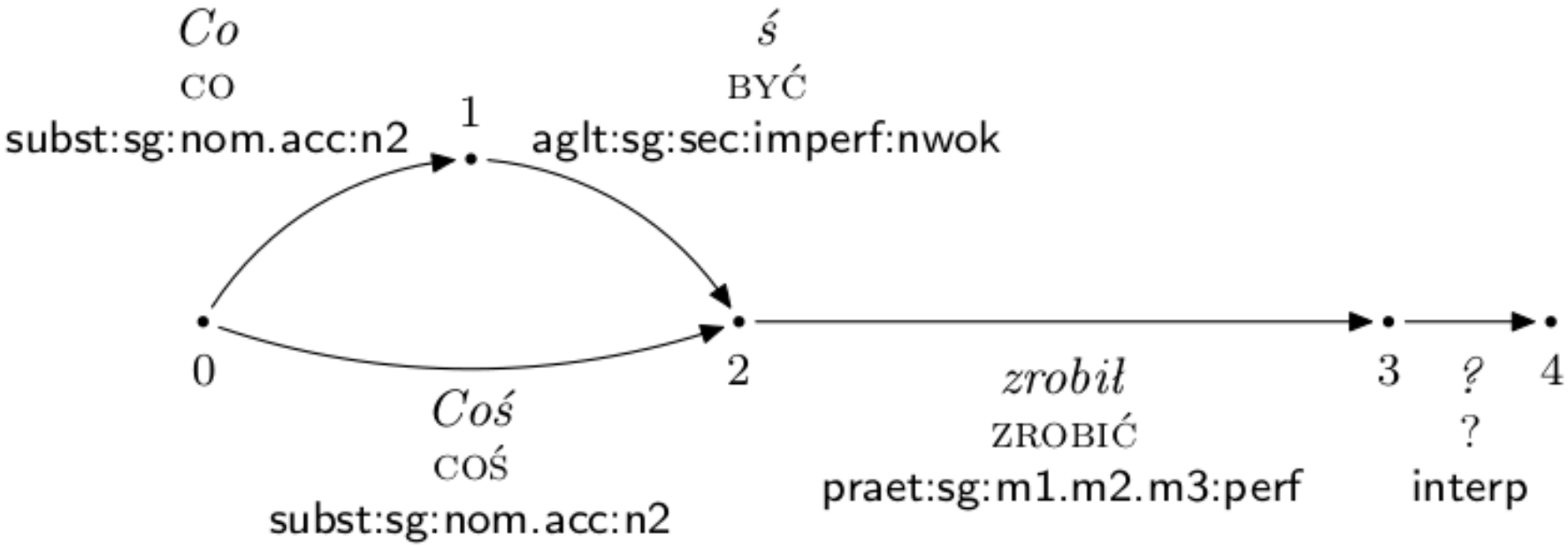
  subst:sg:gen:f- subst:sg:gen:n- subst:sg:nom:n- subst:sg:acc:m1-

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

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
gdybym	40
udziałem	40
żebym	39
żebyś	34

komuś	31
gdybyśmy	19
tom	19
żebyśmy	12
gdybyś	7
działam	6
jam	5
doń	5
oścież	5
gdybyście	4
tyłem	4
musiałem	3
podziałem	3
piekłem	3

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

Problem niejednoznaczności segmentacji

Możliwe rozwiązania: tagset pośredni Pomysł od Adama Radziszewskiego

Problem niejednoznaczności segmentacji

Możliwe rozwiązania: dostosowanie metody uczenia maszynowego do reprezentacji grafowej Pomysł od Jakuba Waszczuka - CRFy

Poziom pewności ujednoznaczniania morfosyntaktycznego

Concraft - marginals

Problem lematyzacji

Tagery morfostynaktyczne innych języków

jakie rozwiązania są możliwe - jakie były na świecie

TODO

jak wzrasta jakość tagowania wraz ze wzrostem ilości danych

TODO

Tagery języka polskiego – analiza jakościowa

Dyskusja i rekomendacje

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 SGJP, używany w słowniku Morfologicznym i w Morfeuszu.

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
```

Struktura danych treningowych w NKJP1M

TODO

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.

Bibliografia II


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.


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.

Bibliografia III


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.


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.

Bibliografia IV

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.