

Dążenie do ideału, czyli co można poprawić w NKJP?

Łukasz Kobyliński Łukasz Szałkiewicz

Instytut Podstaw Informatyki Polskiej Akademii Nauk
ul. J. K. Ordona 21, 01-237 Warszawa

7 stycznia 2013

Informacje o projekcie

Projekt Narodowego Centrum Nauki

- „Automatyczne wykrywanie i korekcja błędów anotacyjnych w polskich korpusach językowych” (umowa nr 2011/01/N/ST6/01107)
- projekt dwuletni: styczeń 2012 – grudzień 2013
- wykonawcy: Ł. Kobyliński (kierownik), A. Przepiórkowski (opiekun), Ł. Szalkiewicz (lingwista)

Wsparcie ze strony projektu CESAR

- współfinansowanie narzędzia do ręcznej korekcji anotacji korpusu,
- współfinansowanie pracy lingwistów,
- podziękowania dla M. Lenarta za udostępnienie oprogramowania używanego we wcześniejszych eksperymentach.

Założenia i cele

Przyjęte założenia

- weryfikacja metody automatycznej detekcji i korekcji błędów przeprowadzana będzie dla morfosyntaktycznej warstwy anotacji.

Cele

- określenie skuteczności metod detekcji błędów anotacyjnych znanych dla innych języków w korpusie polskojęzycznym,
- zaproponowanie metody dostosowanej do specyfiki języka polskiego,
- zaproponowanie metody automatycznej korekcji błędów anotacyjnych w warstwie morfosyntaktycznej i jej implementacja w postaci narzędzia,
- poprawienie jak największej liczby błędów anotacyjnych w NKJP.

Sposób weryfikacji metod

Sposób weryfikacji metod

- zadanie polega na wskazaniu segmentów nieprawidłowo zaanotowanych, wśród segmentów zaanotowanych prawidłowo,
- stosujemy klasyczne miary oceny jakości klasyfikacji:
 - precyzja – jaka część wskazanych przez metodę segmentów jest rzeczywiście błędnie zaanotowanych,
 - odzysk – jaka część istniejących w korpusie błędów anotacyjnych została wskazana przez metodę.

Problem z obliczaniem miar oceny

- ręczne sprawdzenie wszystkich rezultatów zwracanych przez metody automatyczne jest czasochłonne,
 - możemy szacować po przejrzeniu pewnej próbki znalezionych segmentów,
- nie wiemy ile jest wszystkich błędów w korpusie,
 - możemy dokonać wstępnego przeglądu korpusu i utworzyć listę znanych błędów.

Plan

1 Ręczne poprawianie błędów

- Narzędzie CorpCor
- Zapytania poliqarpowe
- Statystyki poprawek

2 Zastosowanie metod automatycznych do detekcji błędów

- Przegląd metod
- Metoda n-gramowa
- Odkrywanie reguł

3 Rzut oka na automatyczną anotację pełnego NKJP

- Błędy nielingwistyczne
- Błędy lingwistyczne – tager niewinny
- Błędy lingwistyczne – wina tagera

4 Podsumowanie i dalsze prace

Narzędzie CorpCor

Narzędzie webowe, umożliwiające modyfikację anotacji morfosyntaktycznej korpusu

- możliwa równoległa praca wielu lingwistów jednocześnie,
- zmiany zapisywane w bazie danych, łącznie z historią i metadanymi,
- korpus odczytywany jest z formatu XML.

Sposób pracy

- wyszukiwanie w korpusie (integracja z Poliqarpem),
- przeglądanie korpusu (przeglądanie według tekstów, akapitów, segmentów).

Narzędzie CorpCor – wyszukiwanie

Szukaj w korpusie Przeglądaj korpus Przeglądaj zmiany Wyloguj się

[base=Koziegłów] Szukaj Edytuj zaznaczone Zaznacz wszystkie Kasuj zaznaczenie ☒ Wyświetlaj modyfikowane Składnia zapytań

☐ Lewy kontekst Dopasowanie Prawy kontekst

☐ , Barczewa, Czarnego i Koziegłów[base=Koziegłów, msd=pl:gen:n, ctag=subst] .

1-1 of 1

Anotacje Kontekst dopasowania Informacje

Id	Forma ortograficzna	Forma bazowa	MSD	CTAG
segm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst

Forma ortograficzna:
Koziegłów

Forma bazowa:
Koziegłów

MSD:
pl:gen:n

CTAG:
subst

Komentarz:
[]

Usunięty: ☐

Zastosuj Zastosuj do zaznaczonych rezultatów

Narzędzie CorpCor – historia anotacji

Szukaj w korpusie Przeglądaj korpus Przeglądaj zmiany Wyloguj się

<< A >>

Tekst ciągły Lista

15 redaktorów 13 więziennych gazetek ukazujących się w zakładach kamych całej Polski rozpoczyna dziś " Ogólnopolskie
Gazet Więziennych Pen-Press Białystok '99 " . Jest to pierwsze spotkanie tego typu w kraju . Do Białegostoku
więziennej m.in . z Warszawy , Krakowa , Barczewa , Czarnego i Kozieglów .

Forma ortograficzna:
Kozieglów

Forma bazowa:
Kozieglowy

MSD:
pl:gen:n

CTAG:
subst

Komentarz:
Usunięty: ☐

Zapisz Historia edycji Anuluj

Historia anotacji

Użytkownik	Data	Forma ortograficzna	Forma bazowa	MSD	CTAG	Komentarz	Zapytanie
gwt	2012 Sep 19 18:01:29	Kozieglów	Kozieglowy	pl:gen:n	subst		[base=Kozieglów]
XML		Kozieglów	Kozieglowy	pl:gen:n	subst		

Zamknij

Narzędzie CorpCor – rozszerzenia

Możliwe rozszerzenia w przyszłości

- inteligentne podpowiedzi – dobierane biorąc pod uwagę kontekst segmentu,
- uczenie systemu on-line – zwiększanie jakości automatycznej korekty anotacji,
- wsparcie dla innych formatów korpusów.

Narzędzie CorpCor – odpowiedzi anotacji

[Szukaj w korpusie](#) [Przeglądaj korpus](#) [Przeglądaj zmiany](#) [Wyloguj się](#)

<< A >>

Tekst ciągły Lista

Id	Forma ortograficzna	Forma bazowa	MSD	CTAG
segm_1.1-seg	Po	po	acc	prep
segm_1.2-seg	raz	raz	sg:acc:m3	subst
segm_1.3-seg	pierwszy	pierwszy	sg:acc:m3:pos	adj
segm_1.4-seg	archiwum	archiwum	sg:nom:n	subst
segm_1.5-seg	byłego	były	sg:gen:n:pos	adj
segm_1.6-seg	KGB	KGB	sg:gen:n	subst
segm_1.7-seg	przekazało	przekazać	sg:n:perf	praet

Forma ortograficzna:

archiwum

Forma bazowa:

archiwum

MSD:

sg:nom:n

C

m1

m2

m3

K f tarz:

L n ty: ☐

Zapisz

Historia edycji

Anuluj

Jak znajdowano błędy



Nie, to nie mój

Szukam dalej...

Jak znajdowano błędy

- komentarze anotatorów
- Instrukcja znakowania
- zapytania regułowe
- zgłoszenia błędów
- literówki i nadmierne segmenty

Bardzo mocne reguły

- a/i [base=także&pos=qub] **qub -> conj**
- [base=jakoś&pos=adv°!=pos] **adv -> adv:pos**
- [orth=mimo/i&pos!=prep] [pos=interp]? że
mimo:. -> mimo:prep:gen
- [orth=z/i&pos!="brev|subst"&vocalicity!="wok|nwok"]
trzeba dodać parametr: nwok

Mocne reguły

- [orth=„ponad|około|koło”/i&pos!=qub] [pos=num]
prep:. → qub
- [pos=ppas] [orth=jako&pos!=prep] [case=nom]
jako:. → jako:prep:nom
- [base=„szelki|wierni|kozaczki|oszczędności|media|dane|środki|prace|roboty|negocjacje|zakupy|siły”] **nie-PT**

Do sprawdzenia

- [base=im&pos!=conj] []{1,20} [orth=tym] **im:comp -> im:conj**
- między/i [orth=innymi&gender!=f] . -> f
- [orth=niedaleko/i&pos!="prep|qub"] [pos!=interp] **adv -> prep:gen**
- [orth=za | orth=z] [orth="młodu|widna|bliska|daleka|dawna|grubsza|lekka|lewa|nagła|osobna|ostrożna|prawa|ruska|rzadka|wolna|wysoka"&pos!=adjp]
 . -> **adjp uwaga: plus zmiana base**
- [orth=milion/i&pos!=subst]

Błędy poprawione w korpusie NKJP1m

zmiana	liczba zmian
poprawa tagu	1052
poprawa literówki	1046
poprawa base	239
poprawa base, poprawa tagu	198
poprawa tagu, poprawa base	85
poprawa literówki, zmiana base	16
poprawa base, poprawa literówki	4
poprawa literówki, zmiana tagu	4
poprawa literówki, poprawa base	3
poprawa literówki, poprawa tagu	3
poprawa literówki, zmiana base, zmiana tagu	2
poprawa literówki, zmiana base i tagu	1
inne	39
SUMA	2692

Poprawa tagu – 1052

- *Łupem padła złota biżuteria (...) a także trzy lub cztery złote monety Republiki Południowej Afryki*
także:qub → także:conj
- *Ceny oscylują w granicach kilku tysięcy złotych, choć zdarzyło się już, że parokrotnie przekroczyły granicę 10000 zł.*
10000:pl:gen:m3:congr → 10000:pl:gen:m2:congr
- *Naszym zadaniem było jedynie znalezienie i wyremontowanie pomieszczenia.*
jedynie:adv:pos → jedynie:qub
- — *a ino że to to to już zrobione tam już rok czy ze dwa*
z:prep:acc:woc → z:qub:wok

Poprawa tagu, poprawa base; poprawa base, poprawa tagu – 283

- (...) *rozpoczął się dolina, gdzie przycupnięty zagajnik, wysokie trawy i gęste partie krzaków dawały szansę skutecznej ucieczki, gdyby zaszła taka potrzeba*
przycupnąć:ppas:sg:nom:m3:perf:aff →
przycupnięty:adj:sg:nom:m3:pos
- *Ich celem był sklep spożywczy w Niedalinie.*
Niedalin:subst:sg:loc:m3 → **Niedalino:subst:sg:loc:n**
- (...) *w liście, który Szaron wysłał do Arafata w odpowiedzi na gratulacje z okazji objęcia władzy.*
gratulacja:pl:acc:f → **gratulacje:pl:acc:n**
- *byłem dziwnie pewien, że Magda by mnie nie obudziła*
pewien:adj:sg:nom:m1:pos → **pewny:adjc**

Poprawa base – 239

- *W rozgrywanym systemem KO konkursie Schmitt przegrał z Hiroya Saitohem.*
Hiroya:subst:sg:inst:m1 Saitohem:subst:sg:inst:m1 →
Hiroya:subst:sg:inst:m1 Saitoh:subst:sg:inst:m1
- *Swoje dzieła wystawiał m.in. w Zakopanem, Wrocławiu Wiesbaden oraz Karslsruhe.*
*Karslsruhe:subst:sg:loc:n → **Karlsruhe:subst:sg:loc:n***
- *Pod Choszcznem, na prostym odcinku drogi między Suliborzem a Bytowem...*
*Bytowem:subst:sg:inst:m3 → **Bytów:subst:sg:inst:m3***
- *PS Minęły znów dwa dni.*
*post scriptum:brev:npun → **postscriptum:brev:npun***

Poprawa literówki – 1046

- *Omawialisiny* hałaśliwie szkolne sprawy
omawiać:fin:pl:pri:imperf
- *Honecker został zwolniony przez sąd w Berlinie* za względów humanitarnych.
ze:prep:gen:wok
- — *Nie jestem* zadna królewna — *burknęła*.
żaden:adj:sg:nom:f:pos
- *Tu* im *dobrze odrobił pracę domową i zdania do tego* skeczy *dobrał* zgodnie ze znaczeniem słów.
skecz:subst:sg:gen:m3

Poprawa literówki + zmiana base i/lub tagu – 23

- *A tysy? 0, demoniczny tysy, za którym (...)*
*0:interj → **O:interj***
- *Polsce egzorcycści chcą, by przy Episkopacie Polski został powołany koordynator ich działalności.*
*Polska:subst:sg:loc:f → **polski:pl:nom:m1:pos***
- — *Na różnych etach swojego życia parałem się różnymi zajęciami.*
*eta:subst:pl:loc:f → **etap:subst:pl:loc:m3***

Poprawa literówki + poprawa base i/lub tagu – 10

- Po zawarciu nowego pokoju w połowie trzynastego stulecia Świętopełk nową, piątą z rzędu wojnę rozpoczął.

Świętopełk:subst:sg:nom:m1 → **Świętopełk:subst:sg:nom:m1**

- (...) a rząd zadłużył się na kolejne pół miliarda dolarów (...)

pół:num:pl:acc:m3:rec → pół:**subst:sg:acc:n**

Przegląd podejść do automatycznej detekcji błędów



ołówek automatyczny

1 wciśnięcie za mało, 2 wciśnięcia za dużo

Przegląd podejść do automatycznej detekcji błędów

Zadanie rozumiane jako problem detekcji anomalii (obserwacji odstających)

- uczenie tagera statystycznego na podstawie zaanotowanego korpusu i porównanie wyniku jego działania na tym samym korpusie z anotacją faktyczną [van Halteren, 2000],
- reprezentowanie anotacji morfosyntaktycznej w korpusie za pomocą rozkładu prawdopodobieństwa i wyszukiwanie elementów, które należą do tego rozkładu z niewielkim prawdopodobieństwem [Eskin, 2000],
- wyszukiwanie nieprawidłowych kombinacji tagów morfosyntaktycznych [Kveton and Oliva, 2002],
- wykorzystanie maszyny wektorów podpierających (SVM) dla każdej wartości tagu do znalezienia kontekstów, które z dużym prawdopodobieństwem są błędem anotacyjnym [Nakagawa and Matsumoto, 2002],

Przegląd podejść do automatycznej detekcji błędów

Zadanie rozumiane jako problem detekcji anomalii (obserwacji odstających)

- znajdowanie n-gramów, które zawierają w sobie leksemy anotowane w różny sposób w poszczególnych wystąpieniach w korpusie [Dickinson and Meurers, 2003],
- znajdowanie reguł asocjacyjnych pomiędzy anotacjami segmentów występujących w podobnych kontekstach [Novak and Razimova, 2009],
- łączenie wielu metod [Loftsson, 2009].

Metoda n-gramowa

Dickinson and Meurers, 2003

- anotacja segmentu zależy od kontekstu, w jakim segment się znajduje,
- błędy anotacyjne to niekonsekwentnie zaanotowane segmenty w podobnych kontekstach.

Variation n-gram

- n-gram, zawierający jeden lub wiele segmentów o różnych anotacjach morfosyntaktycznych, na przykład (długość 10):
 - Zamykam dyskusję. Do głosowania [głosować:ger:sg:n:imperf:n] nad uchwałą Senatu przystąpimy *jutro rano*.
 - Zamykam dyskusję. Do głosowania [głosowanie:subst:sg:gen:n] nad uchwałą Senatu przystąpimy *w bloku głosowań*.

Schemat odkrywania

Algorytm wzorowany na metodzie apriori

- wygeneruj kandydatów: znajdź wszystkie unigramy w korpusie,
- powtarzaj, dopóki zbiór kandydatów jest niepusty:
 - ogranicz zbiór kandydatów do n-gramów, które mają niejednakową anotację w korpusie,
 - wygeneruj kandydatów: rozszerz n-gramy do $(n+1)$ -gramów, uwzględniając segmenty po prawej i po lewej stronie dotychczasowego n-gramu.

- Zamykam dyskusję. Do głosowania nad uchwałą Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana pośła Czesława Sobierajskiego.

- Zamykam dyskusję. Do głosowania nad uchwałą Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana pośła Roberta Smolenia.

- Zamykam dyskusję. Do **głosowania** nad uchwałą Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana pośła Czesława Sobierajskiego.

- Zamykam dyskusję. Do **głosowania** nad uchwałą Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana pośła Roberta Smolenia.

- Zamykam dyskusję. Do **głosowania nad** uchwałą Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana pośła Czesława Sobierajskiego.

- Zamykam dyskusję. Do **głosowania nad** uchwałą Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana pośła Roberta Smolenia.

- Zamykam dyskusję. **Do głosowania** nad uchwałą Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana pośła Czesława Sobierajskiego.

- Zamykam dyskusję. **Do głosowania** nad uchwałą Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana pośła Roberta Smolenia.

- Zamykam dyskusję. **Do głosowania nad** uchwałą Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana pośła Czesława Sobierajskiego.

- Zamykam dyskusję. **Do głosowania nad** uchwałą Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana pośła Roberta Smolenia.

- Zamykam dyskusję. **Do głosowania nad uchwałą** Senatu przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Ułaszczania i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana posła Czesława Sobierajskiego.

- Zamykam dyskusję. **Do głosowania nad uchwałą** Senatu przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana posła Roberta Smolenia.

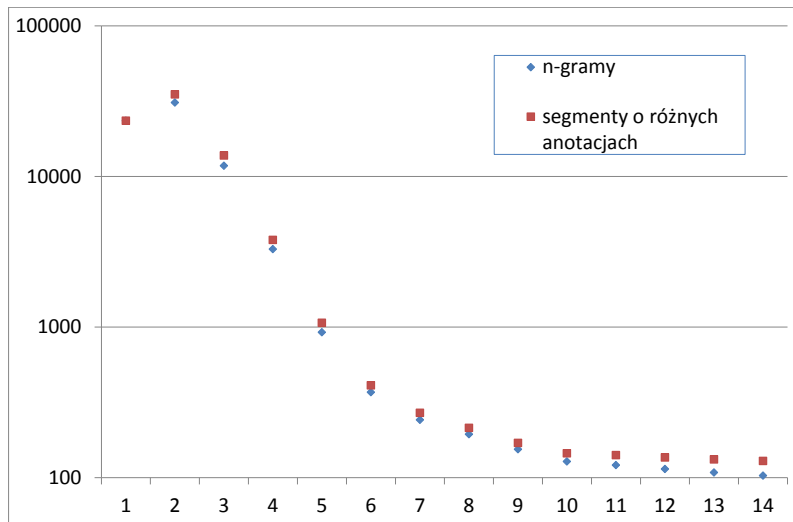
- Zamykam **dyskusję. Do głosowania nad uchwałą Senatu** przystąpimy jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Uwłaszczenia i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana posła Czesława Sobierajskiego.

- Zamykam **dyskusję. Do głosowania nad uchwałą Senatu** przystąpimy w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana posła Roberta Smolenia.

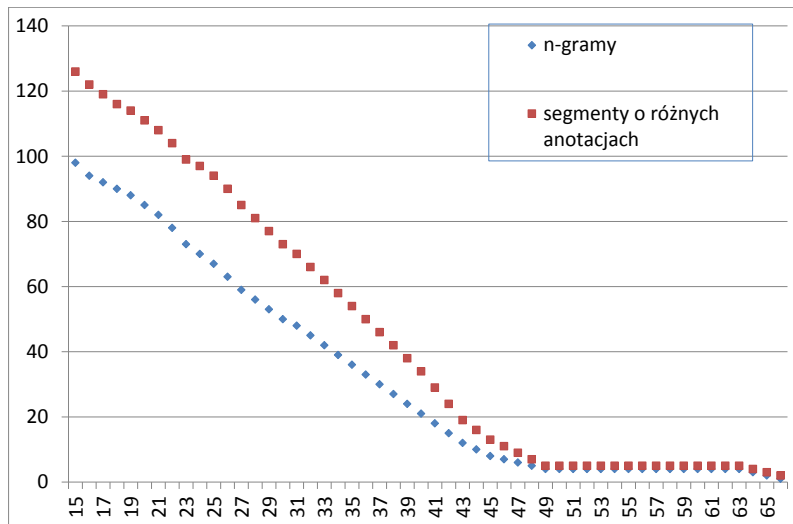
- **Zamykam dyskusję. Do głosowania nad uchwałą Senatu przystąpimy** jutro rano. Przystępujemy do rozpatrzenia punktu 8. porządku dziennego: Sprawozdanie Komisji Gospodarki oraz Komisji Skarbu Państwa, Uwłaszczenia i Prywatyzacji o stanowisku Senatu w sprawie ustawy o giełdach towarowych (druki nr 2231 i 2261). Proszę o zabranie głosu sprawozdawcę komisji pana posła Czesława Sobierajskiego.

- **Zamykam dyskusję. Do głosowania nad uchwałą Senatu przystąpimy** w bloku głosowań. Przystępujemy do rozpatrzenia punktu 24. porządku dziennego: Sprawozdanie Komisji Europejskiej oraz Komisji Administracji i Spraw Wewnętrznych o stanowisku Senatu w sprawie ustawy Ordynacja wyborcza do Parlamentu Europejskiego (druki nr 2416 i 2436). O zabranie głosu proszę sprawozdawcę komisji pana posła Roberta Smolenia.

Znalezione n-gramy



Znalezione n-gramy

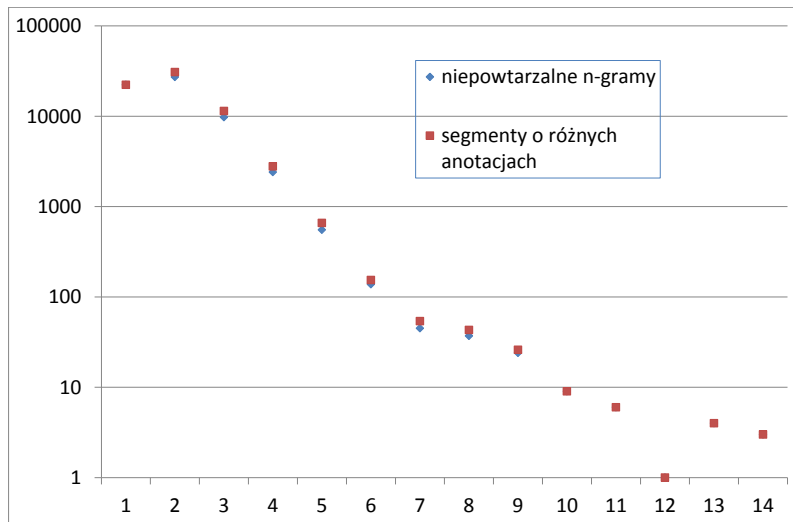


Obserwacje

Obserwacje

- krótsze n-gramy są często powtórzeniami fragmentów kontekstów znalezionych przez dłuższe n-gramy,
- im dłuższy n-gram, tym większe prawdopodobieństwo, że znajduje się w nim błędna anotacja (szerszy wspólny kontekst),
- podejrzane segmenty na krawędziach n-gramów są najczęściej fałszywymi zgłoszeniami (kontekst po lewej lub po prawej jest inny),
- bogata fleksja w języku polskim powoduje, że budowanie n-gramów na podstawie form ortograficznych jest mniej skuteczne, niż dla języka angielskiego.

Niepowtarzalne n-gramy



Niepowtarzalne n-gramy – weryfikacja

Przecięcie ze zbiorem poprawek ręcznie wprowadzonych

- 2692 poprawek, z czego 1332 dotyczyło poprawy tagu

minN	segmenty	% korpusu	TP	FP	precyzja	odzysk	F
2	226548	18,64%	901	225	0,40%	67,64%	0,79%
3	54970	4,52%	398	38	0,72%	29,88%	1,41%
4	10448	0,86%	97	3	0,93%	7,28%	1,65%
5	2513	0,21%	24	0	0,96%	1,80%	1,25%
6	873	0,07%	12	0	1,37%	0,90%	1,09%

Usunięcie segmentów na skraju n-gramów – weryfikacja

Przecięcie ze zbiorem poprawek ręcznie wprowadzonych

- 2692 poprawek, z czego 1332 dotyczyło poprawy tagu

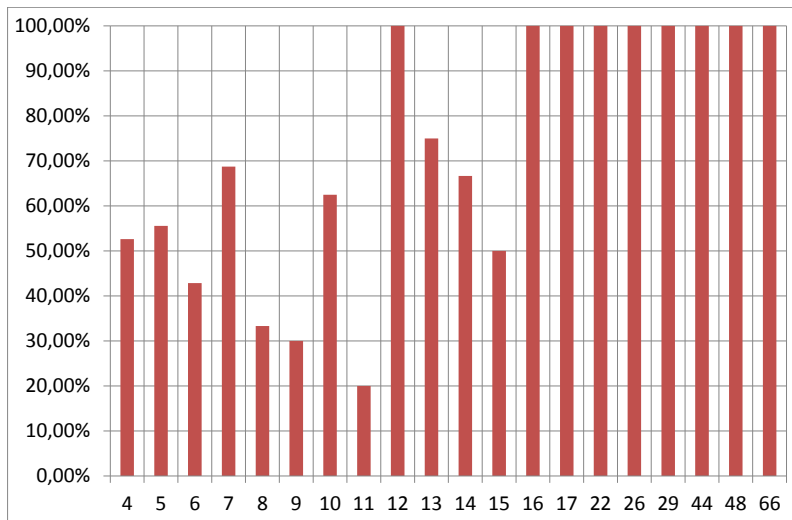
minN	segmenty	% korpusu	TP	FP	precyzja	odzysk	F
3	18855	1,55%	203	10	1,08%	15,24%	2,01%
4	4870	0,40%	73	2	1,50%	5,48%	2,35%
5	1605	0,13%	23	0	1,43%	1,73%	1,57%
6	678	0,06%	11	0	1,62%	0,83%	1,09%

Usunięcie segmentów na skraju n-gramów – weryfikacja

Bezpośrednie sprawdzenie rezultatów generowanych przez metodę

N	wątpliwe konteksty	sprawdzone	zawierające błędy	precyzja
4	1192	19	10	52,63%
5	373	9	5	55,56%
6	104	21	9	42,86%
7	32	16	11	68,75%
8	24	15	5	33,33%
9	23	20	6	30,00%
10	8	8	5	62,50%
11	5	5	1	20,00%
12	1	1	1	100,00%
>12	23	23	19	82,61%
	1785	137	72	52,55%

Precyzja dla n-gramów poszczególnych długości



Relaksacja wymagań na n-gram

Staramy się zwiększyć odzysk

- bardzo podobne konteksty mogą różnić się pojedynczymi znakami, co uniemożliwia ich identyfikację,
- dla języka angielskiego znajdowanych jest więcej n-gramów ze względu na mniejszą różnorodność form ortograficznych.

Relaksacja wymagań na n-gram

Podjęcia do relaksacji

- rozszerzać konteksty niezależnie od formy ortograficznej niektórych typów,
 - interp,
 - brev,
 - num,
 - numcol.

Park krajobrazowy z pocz. **XIX [sg]** w. z kopcem i głazem narzutowym ku pamięci braci Niemojewskich, członków Rządu Narodowego w powstaniu listopadowym.

Wieś na szlaku bursztynowym. We wsi sklepiona kapliczka wieżowa z poł. **XIX [pl]** w. z rzeźbą ludową św. Jana Nepomucena.

Relaksacja wymagań na n-gram

Podjęcia do relaksacji

- rozszerzać konteksty niezależnie od formy ortograficznej niektórych typów,
 - interp,
 - brev,
 - num,
 - numcol.

System komunikacyjny Warszawy sypie się jak **wiadukt w Al. Jerozolimskich** [adj:pl:loc:f:pos].

Podstawowym kryterium, decydującym o wyborze firmy, która wyremontuje **wiadukt w Al. Jerozolimskich** [adj:pl:loc:n:pos] , okazała się cena.

Relaksacja wymagań na n-gram

Podejścia do relaksacji

- ignorować formę ortograficzną, badać konteksty tylko na podstawie tagów morfosyntaktycznych.

29 września ok. godz. 22. 30 przy ul. Sokolskiej w Radlinie
[subst:sg:loc:m3] **nieznani** [znać:ppas] sprawcy włamali się do
pomieszczeń Miejskiego Ośrodka Kultury „Sokolnia”.

Policjanci po otrzymaniu informacji o zaistniałym zdarzeniu zatrzymali
wszystkich sprawców rozboju i odzyskali skradzione przedmioty.
Postępowanie prowadzi KPP Ciechanów. Dwa dni wcześniej, 7 bm. ok.
godz. 22. 30 przy ul. Pułtuskiej w Ciechanowie [subst:sg:loc:m3] **nieznani**
[nieznany:adj] sprawcy pobili 15-latkę.

Przecięcie ze zbiorem poprawek ręcznie wprowadzonych (n-gramy uogólnione)

Uogólnienie interp, brev, num, numcol

minN	TP	FP	segmenty	precyzja	odzysk	F
4	90	2	8939	1,01%	6,76%	1,75%
5	32	0	2878	1,11%	2,40%	1,52%
6	16	0	1107	1,45%	1,20%	1,31%

Uogólnienie wszystkich typów

minN	TP	FP	segmenty	precyzja	odzysk	F
4	257	30	28499	0,90%	19,29%	1,72%
5	98	9	9547	1,03%	7,36%	1,80%
6	36	0	2762	1,30%	2,70%	1,76%

Obserwacje i wnioski

Obserwacje

- uogólnienie n-gramów prowadzi do zwiększenia odzysku kosztem precyzji,
- zaletą n-gramów uogólnionych jest zarówno możliwość znalezienia nowych kontekstów wątpliwych, jak i wydłużenie kontekstów znalezionych zwykłymi n-gramami,
- n-gramy uogólnione pozwalają odzyskać istotne n-gramy, usunięte po zastosowaniu heurystyki ignorującej segmenty na krawędzi n-gramów.

Wnioski

- najlepiej wykorzystać wszystkie typy n-gramów,
- wyeliminować powtarzające się konteksty, a pozostałe przeglądać od najdłuższych, do najkrótszych.

Odkrywanie reguł

Podejście do problemu „od drugiej strony”

- zamiast szukać błędów, szukamy silnych (ale nie pewnych!) reguł, które obowiązują w anotowanym korpusie,
- podejrzane są przypadki, dla których istnieje reguła o pewności rzędu 99% i istotnym wsparciu.

Odkrywamy reguły asocjacyjne pomiędzy atrybutami segmentów

- wybieramy losową próbkę segmentów z korpusu (np. 10%),
- próbka może zawierać pojedyncze segmenty (błędy anotacji niezależne od kontekstu) lub kilka segmentów (uwzględnienie kontekstu),
- szukamy reguł o wsparciu rzędu 0,1% i zaufaniu powyżej 99%, ale poniżej 100%.

Odkrywanie reguł – przykład

- `base_0=przez 500 ==> msd_0=acc:nwok 497`
false positive
- `base_0=ja 770 ==> ctag_0=ppron12 765`
życie.. – to ja [Jagódka:subst:sg:acc:f] Jagódkę też leczy..
- `base_0=my 276 ==> ctag_0=ppron12 274`
, o którym mówił „ my [my:subst:sg:nom:n] portfolio”.
- `ctag_0=aglt 290 ==> base_0=być 288`
[pos=aglt&base!=być] – często base=śmy zamiast być
- `base_0=no msd_0= 446 ==> ctag_0=qub 442`
: wow, oh, no [no:interj] ! i damned. 32
- `base_0=tak ctag_0=adv 118 ==> msd_0=pos 117`
dla niektórych nie ma stopnia [degree!=pos]

Wykorzystanie informacji z narzędzia CorpCor do zwiększenia jakości wyników

Podczas pracy lingwisty przechowujemy nie tylko wynik poprawek, ale też zapytania, jakie zostały wykorzystane do znalezienia kontekstów

- podczas rozwijania metody regułowej możemy porównywać reguły znalezione automatycznie z wprowadzonymi ręcznie,
- jeśli n-gram lub reguła sugeruje błędy w kontekstach, wśród których jest kontekst poprawiony przez lingwistę:
 - jeśli nie było to zapytanie ogólne, to podbijamy istotność pozostałych sugerowanych kontekstów z tej samej reguły / n-gramu,
 - podbijamy istotność kontekstów wygenerowanych z n-gramów krótszych, które wskazują inne konteksty,
 - możemy zasugerować wybraną przez lingwistę anotację jako najbardziej prawdopodobną (normalnie: najczęstsza anotacja).

Rzut oka na automatyczną anotację pełnego NKJP



DEMOTYWATORY.PL

Ctrl + F

tego brakuje mi w marcecie...

Przygotowanie tekstów

metadane

- Władysław Reymont a Władysław Stanisław Reymont
- autorzy – nicki internetowe

pozorne różnice

- Czy taka demonstracja mogłaby się odbyć | bez zezwolenia i aprobaty władz uczelni? Na pewno nie. Oto jakie szkoły mamy | w tym piepszonym PRL-bis
- Czy taka demonstracja mogłaby się odbyć bez zezwolenia i aprobaty władz uczelni? Na pewno nie. Oto jakie szkoły mamy w tym piepszonym PRL-bis

Literówki w tekstach

[base=hurtować]

- (...) działa tam zdaje się hurtowania papieru.
[hurtować:ger:sg:gen:n:imperf:aff]
- Mężczyźni nie zważając jednak na ochronę budynku, wkroczyli do hurtowani i grożąc jej pracownikom użyciem broni zażądali wydania pieniędzy. [hurtować:ppas:pl:nom:m1:imperf:aff]

Zasady znakowania

słownik analizatora

- *zbyt* – kublik (adv)
- *mniej* – liczebnik (adv:com)
- *obowiązany* – przymiotnik (subst)
- *w lewo* – przymiotnik poprzyimkowy adjp (subst)
- Mamy nadzieję, że nie tylko wobec siatkarek, ale także (...)

adresy mailowe, adresy stron

- news:biig72\$muu\$1@kastor.ds.pg.gda.pl [Kastor:subst:sg:nom:m1]

inne reguły

Błędny leksem

gdera

- Pani Hela nadal dobrotliwie gdera. [gdera:subst:sg:nom:m1]
- Narzeka i gdera jak zwykle, ale widzę, że frajdę ma w oczach. [gdera:subst:sg:nom:m1]

łup

- gdzie złoczyńcy próbowali spieniężyć łup [łupa:subst:pl:gen:f]
- Zanim oddadzą swój łup w skupie [łupa:subst:pl:gen:f]

czasowniki z pierwszej setki listy frekwencyjnej

- mać, mamć, nizać, chybać, sposobić, mieść, udziać

kup

- I kuszą: wstąp, kup, przecież to czas świąt, wielkiego kupowania, prezentów. [kupa:subst:pl:gen:f]

Błędny leksem

[base=chachać]

- mówi Mateusz Chachaj, rzecznik prasowy Fundacji Polsko-Niemieckie Pojednanie [chachać:impt:sg:sec:imperf]

[base=wiktor]

- Bardzo szybko z zawodami rozstali się natomiast Michał Bałdys, Wiktor Biały, Michał Dziąbek [wiktor:subst:sg:nom:m1]

[base=pecyna]

- a miejsca trzecie Adamowi Sielskiemu z ZSS w Lublińcu i Adamowi Pecynie [pecyna:subst:sg:loc:f]

Błędna charakterystyka gramatyczna

[case=voc]

- Były panie montażystki [pan:subst:sg:voc:m1]
- Za męstwo w wojnie polsko–bolszewickiej 1920 r. został odznaczony [Polska:subst:sg:voc:f]
- w dalszej części lotu zbombardowała niemieckie linie okopów [niemiecki:adj:sg:voc:n:pos] [lin:subst:sg:voc:m2]

Automatyczne poprawianie błędów w pełnym NKJP



Takie drzwi...
...to tylko w Polsce

Automatyczne poprawianie błędów w pełnym NKJP

Inne zadanie, niż poprawianie podkorpusu ręcznie anotowanego

- nie można zastosować metody nienadzorowanej, wyszukującej niekonsekwencji w anotacji (bo anotacja automatyczna),
- można dążyć do wyeliminowania jak największej liczby błędów z NKJP 1M i ponownie wyuczyć tager, który posłuży do nowej anotacji pełnego korpusu,
- można wyszukiwać częste sekwencje (n-gramy i wzorce sekwencyjne) tagów i form podstawowych w pełnym korpusie i weryfikować je z podkorpusem 1M,
- można odkrywać silne reguły w jednym korpusie i weryfikować, czy są spełnione również w drugim.

Podsumowanie

Podsumowanie

- do pełnego wyeliminowania błędów z obu korpusów potrzebne jest jednocześnie zastosowanie metod automatycznych i praca lingwistyczna,
- poprawianie błędów w podkorpusie 1M i pełnym korpusie NKJP to dwa odrębne zadania,
- kluczem wydaje się być odkrywanie częstych zależności i obserwacji odstających w jednym korpusie i kontrastowanie ich z drugim.