PolEval 2020

Task 2. Morphosyntactic tagging of Middle, New and Modern Polish

Witold Kieraś, Marcin Woliński
(prezentuje Łukasz Kobyliński)

Znakowanie morfosyntaktyczne

Znakowanie słów w tekście odpowiednimi tagami (znacznikami) morfosyntaktycznymi.

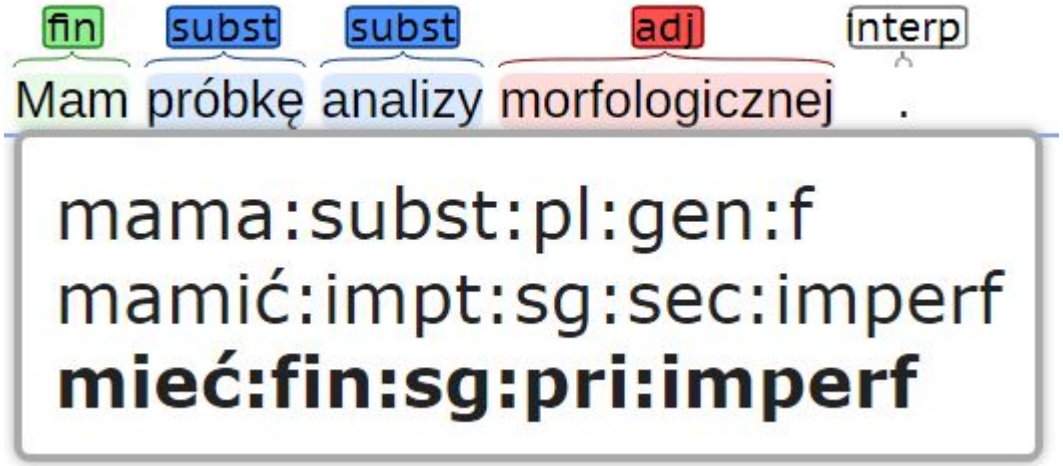
W języku polskim często: dezambiguacja morfosyntaktyczna (wskazywanie właściwego tagu ze zbioru możliwych).

Potrzebna definicja tagsetu:

- w języku angielskim: 36 - 200 tagów, np. the Penn Treebank:
  - NN - noun (singular); NNS - noun (plural);
  - VB - verb, base form; VBD - verb, past tense
- w języku polskim: 4 000 teoretycznie możliwych tagów!

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Tagset w języku polskim



Tagset jest pozycyjny

- `mam – mieć:`**`fin:sg:pri:imperf`**
  - `mieć` – lemat
  - `fin` – klasa gramatyczna (35 możliwości)
  - `sg, pri, imperf` – wymagane atrybuty gramatyczne (specyficzne dla każdej z klas)
  - dla wielu klas gramatycznych - również atrybuty opcjonalne
- … daje to łącznie ok. 4000 możliwych tagów
  (ok. 1 000 pojawia się w rzeczywistym korpusie).

Tagset w języku polskim (cd.)

Skąd tak wiele klas gramatycznych i atrybutów?

Czy *czytanie* to rzeczownik, który odmienia się przez przypadek, czy czasownik (ma aspekt: *rozpoczęto czytanie* ale nie *rozpoczęto przeczytanie*)?

Rozwiązanie: wydzielono osobną klasę gramatyczną dla form typu *czytanie*.

Klasy zostały wyodrębnione

- na podstawie cech fleksyjnych form (przez co się odmieniają, z czym uzgadniają),
- na podstawie cech składniowych (dystrybucyjnych),
- na podstawie cech semantycznych i pragmatycznych.

Tagset w języku polskim (cd.)

Leksem	Fleksem	Symbol	Przykład
Rzeczownik	rzeczownik	subst	*kot, profesorowie*
	forma deprecjatywna	depr	*profesory*
Liczebnik	liczebnik główny	num	*sześć, dużo*
	liczebnik zbiorowy	numcol	*sześcioro, trojga*
Przymiotnik	przymiotnik	adj	*polski*
	przymiotnik przyprzymiotnikowy	adja	*polsko*
	przymiotnik poprzyimkowy	adjp	*polsku*
	przymiotnik predykatywny	adjc	*wesół*
Przysłówek		adv	*bardziej, kiedy*
Zaimek	nietrzecioosobowy	ppron12	*ja, tobie*
	trzecioosobowy	ppron3	*on, jemu*
	SIEBIE	siebie	*sobą*
Czasownik	forma nieprzeszła	fin	*jadam*
	forma przyszła czasownika BYĆ	bedzie	*będę*
	aglutynant czasownika BYĆ	aglt	*-śmy*
	pseudoimiesłów	praet	*jadał*
	rozkaźnik	impt	*jadaj*
	bezosobnik	imps	*jadano*
	bezokolicznik	inf	*jadać*
	imiesłów przysłówkowy współczesny	pcon	*jadając*
	imiesłów przysłówkowy uprzedni	pant	*zjadłszy*
	odsłownik	ger	*jadanie*
	imiesłów przymiotnikowy czynny	pact	*jadający*
	imiesłów przymiotnikowy bierny	ppas	*jadany*

Czasownik typu WINIEN (forma teraźniejsza)		winien	*winna, powinni*
Predykatyw		pred	*trzeba, słychać*
Przyimek		prep	*pod, we*
spójnik	współrzędny	conj	*oraz, lub*
	podrzędny	comp	*że, aby*
Wykrzyknik		interj	*ach, psiakrew*
Burkinostka		burk	*omacku, trochu*
Kublik		qub	*nie, -ż, również*
Skrót		brev	*dr, np*
Ciało obce		xxx	*errare, humanum*
Interpunkcja		interp	*;, ., (,]*

Zadanie historycznofleksyjne PolEval

Na danych reprezentujących 400 lat rozwoju języka polskiego obejmujących:

- XVII i XVIII wiek — okres średniopolski — https://korba.edu.pl/,
- XIX wiek — okres nowopolski — http://korpus19.nlp.ipipan.waw.pl/,
- okres współczesny — http://nkjp.pl/.

Opis fleksji historycznej wymagał dostosowania tagsetu — w każdym z korpusów stosowany jest nieco inny opis, zostały one ujednolicone na potrzeby tego zadania.

Celem zadania jest ujednoznacznienie fleksyjne zaprezentowanych danych i dodanie interpretacji dla słów nieznanych analizatorowi, a więc takie samo zadanie, jakie wykonują tagery tekstów współczesnych.

Zadanie historycznofleksyjne PolEval

Wydaje się nam, że interesującym elementem jest to, że dane nie są jednorodne, opisują trzy ściśle związane ze sobą języki.

Udostępnione dane zawierają informację o czasie powstania poszczególnych tekstów (niekiedy przybliżonym).

Spodziewamy się, że najlepsze tagery będą korzystały z tej informacji.

Tagset historyczny (wybrane różnice)

- Trzy wartości liczby: pojedyncza, podwójna, mnoga, np.
  - *Dwie*$_{\text{DUAL}}$ *żabie*$_{\text{DUAL}}$ *upragnione*$_{\text{DUAL}}$ *po polach biegały*
- Uwzględniono tzw. formy krótkie lub niezłożone przymiotników paralelne do używanych współcześnie form „długich”, czyli „złożonych”:
  - *równy* adj:sg:nom:m:pos — *równien* adjb:sg:nom:m:pos
  - *dawnego* adj:sg:gen:m.n:pos — *dawna* adjb:sg:gen:m.n:pos
  - *polskiemu* adj:sg:dat:m:pos — *polsku* adjb:sg:dat:m.n:pos
  - *piękną* adj:sg:acc:f:pos — *pięknę* adjb:sg:acc:f:pos

  W języku współczesnym ich pozostałością są formacje typu *z dawna* i *po polsku*.

Tagset historyczny (wybrane różnice)

- W okresie średniopolskim dopiero kształtuje się rodzaj męskoosobowy, dlatego w korpusach historycznych stosowane jest inne oznaczenia rodzajów męskich. W danych PolEval przyjęto dla wszystkich okresów oznaczenie m — męski, a bardziej szczegółowe oznaczenia manim1 i manim2 — tylko w kontekstach różnicujących odpowiednie formy.

Podejścia do znakowania morfosyntaktycznego

Uczenie maszynowe, czy reguły lingwistyczne

- ręczne reguły tworzone przez lingwistów
- metody mieszane, np. tager TaKIPI (jęz. polski) - łączą podejście nadzorowane z regułami przygotowanymi przez lingwistów
- metody nienadzorowane - HMM (ukryte modele Markowa) i algorytm Expectation Minimization (EM)
- metody nadzorowane - oparte na algorytmach ML, np. tager Brilla (rule-based); oparte na Conditional Random Fields (CRF) - uwzględniają sekwencyjność języka; wykorzystujące sieci neuronowe

Krótka historia tagerów dla języka polskiego

Dotychczasowe podejścia:

- 2007: znakowanie oparte na regułach (TaKiPI), ok. 88% dokładności,
- 2010: tager Brilla dostosowany do j. polskiego (PANTERA), ok. 89% dokładności,
- 2012: uczenie pamięciowe (WMBT), ok. 90% dokładności,
- 2012: warunkowe pola losowe (WCRFT, Concraft), ok. 91% dokładności,

Tymczasem dla języka angielskiego dokładność przekracza 97%

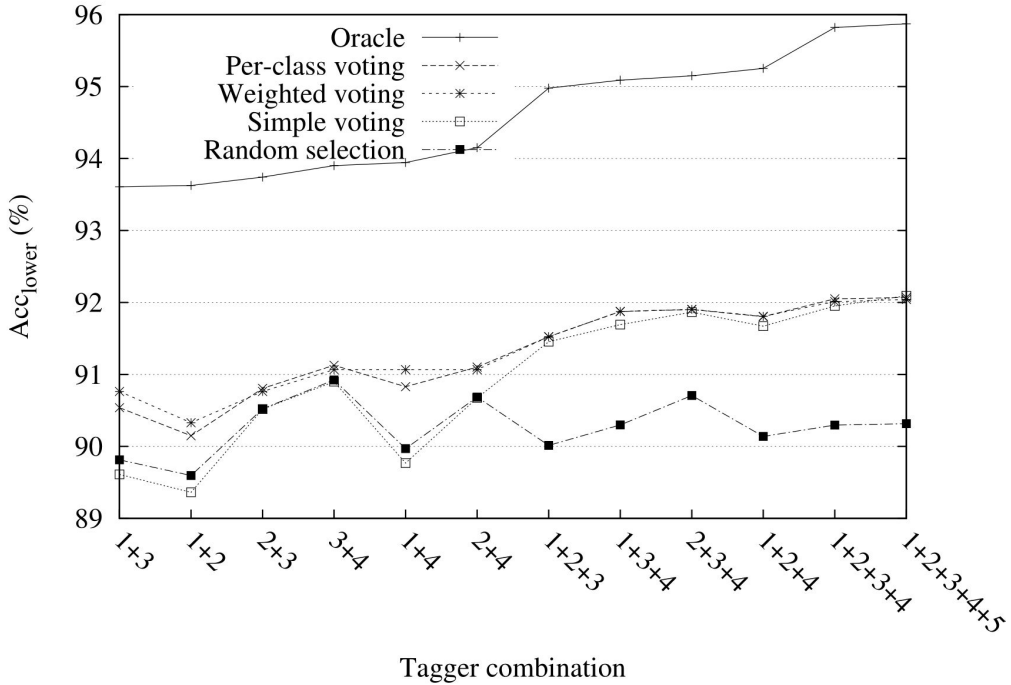
Znakowanie a klasyfikacja

Skoro tagery są klasyfikatorami - możemy wykorzystać ensembling do poprawy jakości znakowania.

2014: PoliTa,

ok. 92% dokładności znakowania

8% błędnych znakowań to 80 milionów błędów w korpusie o rozmiarze 1 miliarda słów!



A może meta-uczenie?

Nauczmy nowy klasyfikator na podstawie wyników działania pojedynczych tagerów:

- który z klasyfikatorów wybrać w danym kontekście?

Statystyki wyboru tagerów w zbiorze testowym:

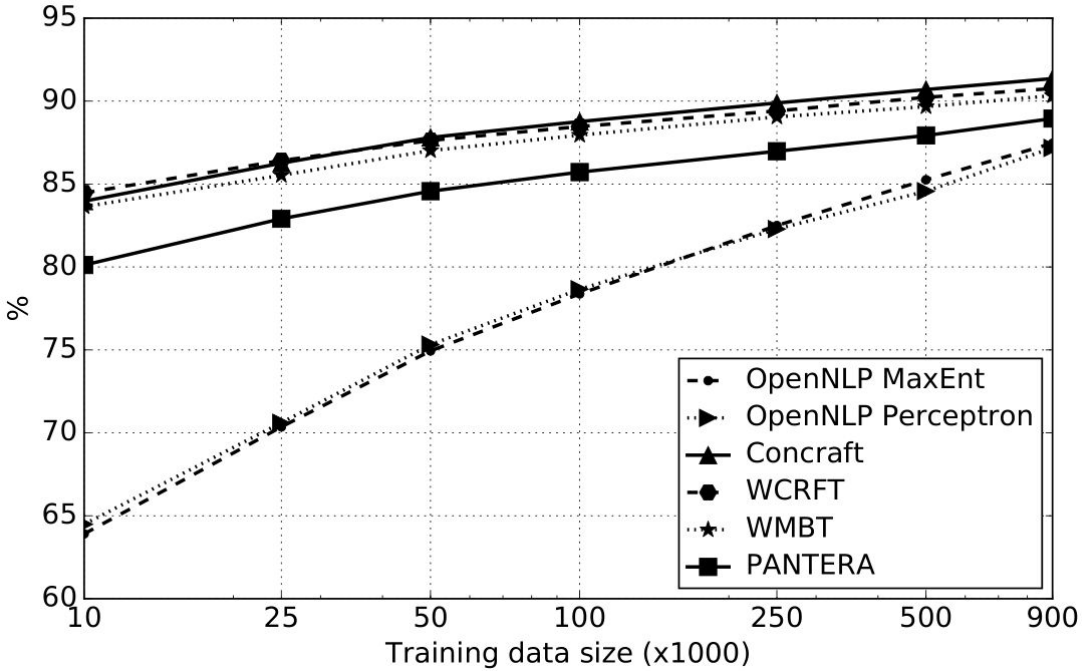
	pantera	wmbt	wcrft	concraft
Total	0.235	4.527	20.491	74.747
Known words	0.294	3.885	22.585	73.235
Unknown words	0.0	7.062	12.193	80.745

Tagger	5-2 split		
	Acc_{lower}	Acc^{K}_{lower}	Acc^{U}_{lower}
pantera	88.3646	91.9884	7.0421
wmbt	90.1567	92.0960	46.6339
wcrft	90.6354	92.7668	42.8029
concraft	91.1535	93.1072	47.3090
polita	91.7264	93.7206	46.9726
Meta libsvm	92.0797	93.9697	49.6637
Meta xgboost	92.1040	93.9842	49.8268

Wpływ rozmiaru danych uczących

Rozmiar (zaanotowanego) zbioru uczącego ma bardzo duży wpływ na jakość znakowania.

… ale tworzenie takiego zbioru jest bardzo kosztowne.



Stan dla języka polskiego - wyniki PolEval 2017

System name	Acc (%)	deep network	hand-crafted features	character-level embeddings	word-level embeddings
Toygger	94.6343	*yes*	*yes*	*no*	*yes*
KRNNT_AB	93.8083	*yes*	*yes*	*no*	*no*
NeuroParser	93.6109	*yes*	*no*	*yes*	*no*
AvgPer_Forced	90.9134	*no*	*yes*	*no*	*no*
Concraft	91.6115	*no*	*yes*	*no*	*no*
WCRFT	91.1693	*no*	*yes*	*no*	*no*
WMBT	90.6722	*no*	*yes*	*no*	*no*

http://poleval.pl