Metody nadzorowane w ujednoznacznianiu sensów słów korpusu ekonomicznego

Łukasz Kobyliński

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

21 listopada 2011

Na czym polega zadanie?

Projekt NEKST – zadanie 1

- podzadanie 1.4 – przegląd i przystosowanie do języka polskiego metod uczenia nadzorowanego w zastosowaniu do automatycznego ujednoznaczniania sensów słów,
- podzadanie 1.5 – przygotowanie zasobów niezbędnych do trenowania i testowania algorytmów w Z1.4.

Dodatkowe założenia

- trenowanie i testowanie metod odbywa się na korpusie dziedzinowym – o tematyce ekonomicznej.
- ostateczna metoda ujednoznaczniania powinna wykorzystywać formalizm, wyrażający reguły ujednoznaczniania sensów słów.

Plan

1. **Zasoby**
   - Słownik haseł ekonomicznych
   - Korpus tekstów
2. **Automatyczne ujednoznacznianie**
   - Analiza danych
   - Metody uczenia maszynowego
   - Metody regułowe
   - Wykorzystanie zewnętrznych zasobów lingwistycznych do poprawy skuteczności metod uczenia maszynowego
3. **Podsumowanie wyników**

Słownik haseł ekonomicznych

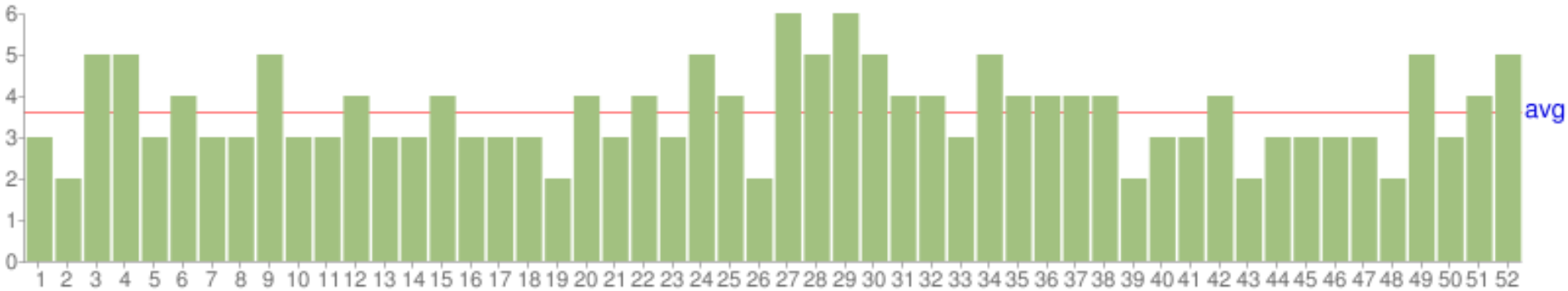
Opracowano słownik 52 haseł z dziedziny ekonomicznej

- agent[n]
- akcja[n]
- baza[n]
- cena[n]
- dochód[n]
- efekt[n]
- firma[n]
- fundusz[n]
- gospodarka[n]
- granica[n]
- inwestycja[n]
- jednostka[n]
- kontrola[n]
- koszt[n]
- linia[n]
- ochrona[n]
- opcja[n]
- pieniądz[n]
- podatek[n]
- podstawa[n]
- polityka[n]
- pomoc[n]
- postępowanie[n]
- praca[n]
- prawo[n]
- projekt[n]
- punkt[n]
- rachunek[n]
- rynek[n]
- rząd[n]
- sąd[n]
- siła[n]
- spółka[n]
- stan[n]
- stopa[n]
- stopień[n]
- system[n]
- środek[n]
- świadczenie[n]
- ubezpieczenie[n]
- udział[n]
- umowa[n]
- unia[n]
- wartość[n]
- warunek[n]
- zasada[n]
- zmiana[n]
- zysk[n]
- czarny[a]
- specjalny[a]
- wolny[a]
- złoty[a]

Słownik haseł ekonomicznych

Statystyki słownika

- 52 hasła
- najmniejsza liczba sensów: 2
- największa liczba sensów: 6
- suma liczby sensów: 188
- średnia liczba sensów: 3,62
- średnia liczba definicji każdego sensu: 2,78

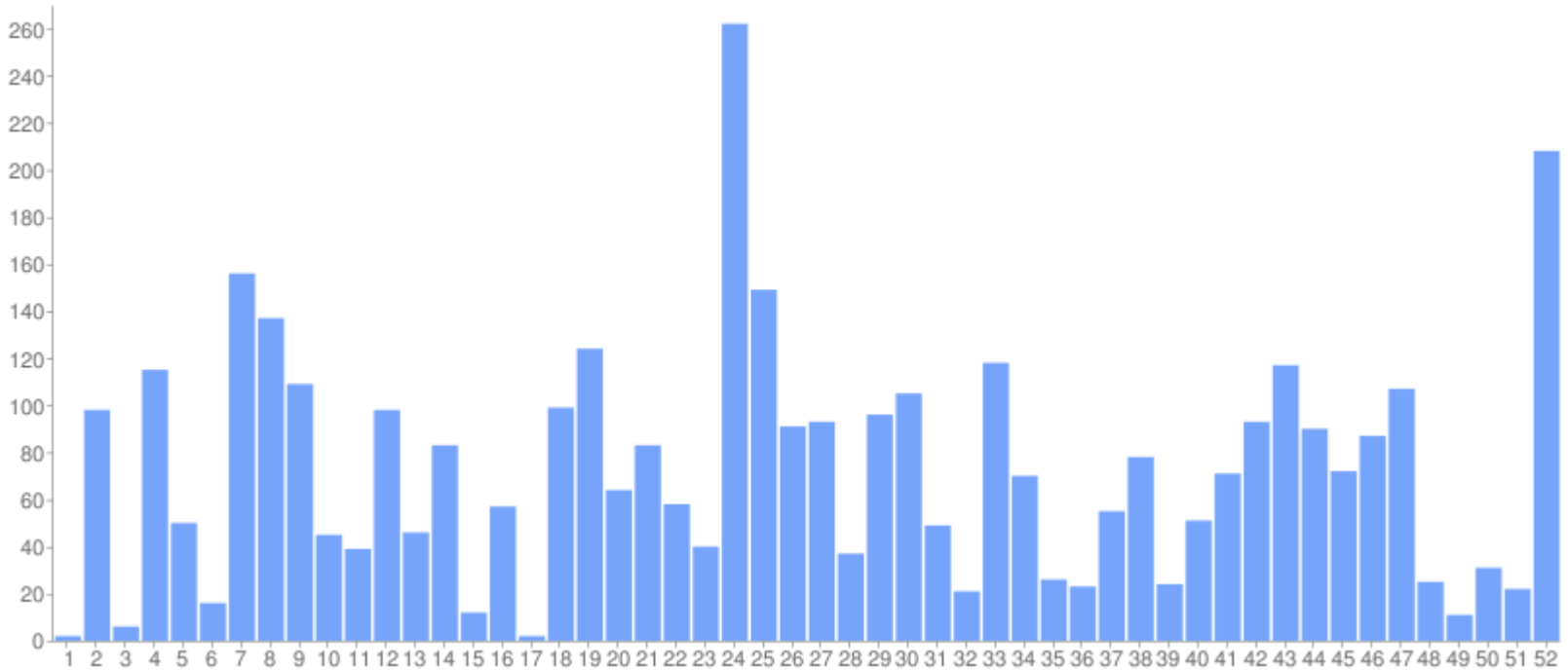


Korpus – dane źródłowe

Korpus NKJP_econo

- wybrano paragrafy, które dotyczą dziedziny ekonomicznej
- liczba segmentów: 87 816

Rozkład liczby haseł ze słownika w korpusie

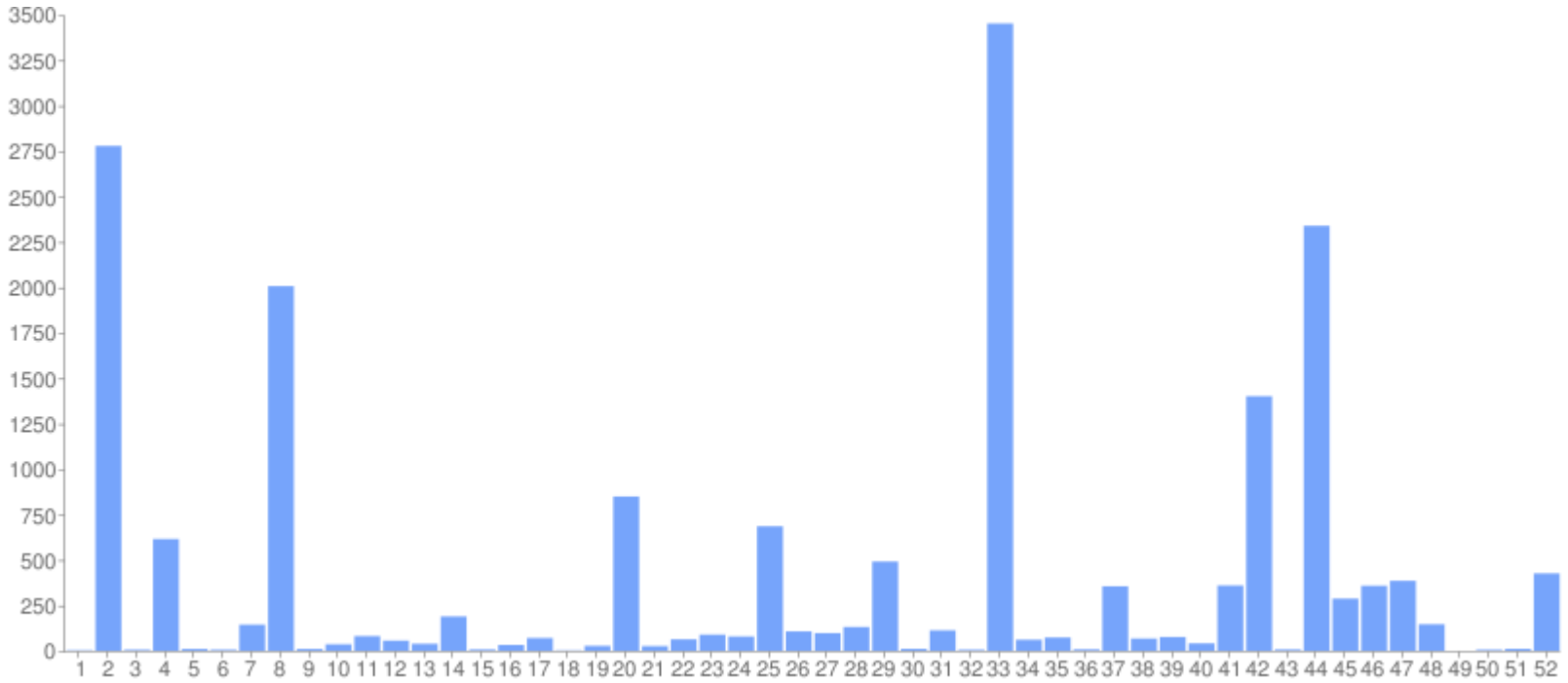


Korpus – dane źródłowe

Korpus GPW (raporty giełdowe)

- dodano anotację morfosyntaktyczną za pomocą TAKIPI 1.8
- liczba segmentów: 282 366

Rozkład liczby haseł ze słownika w korpusie

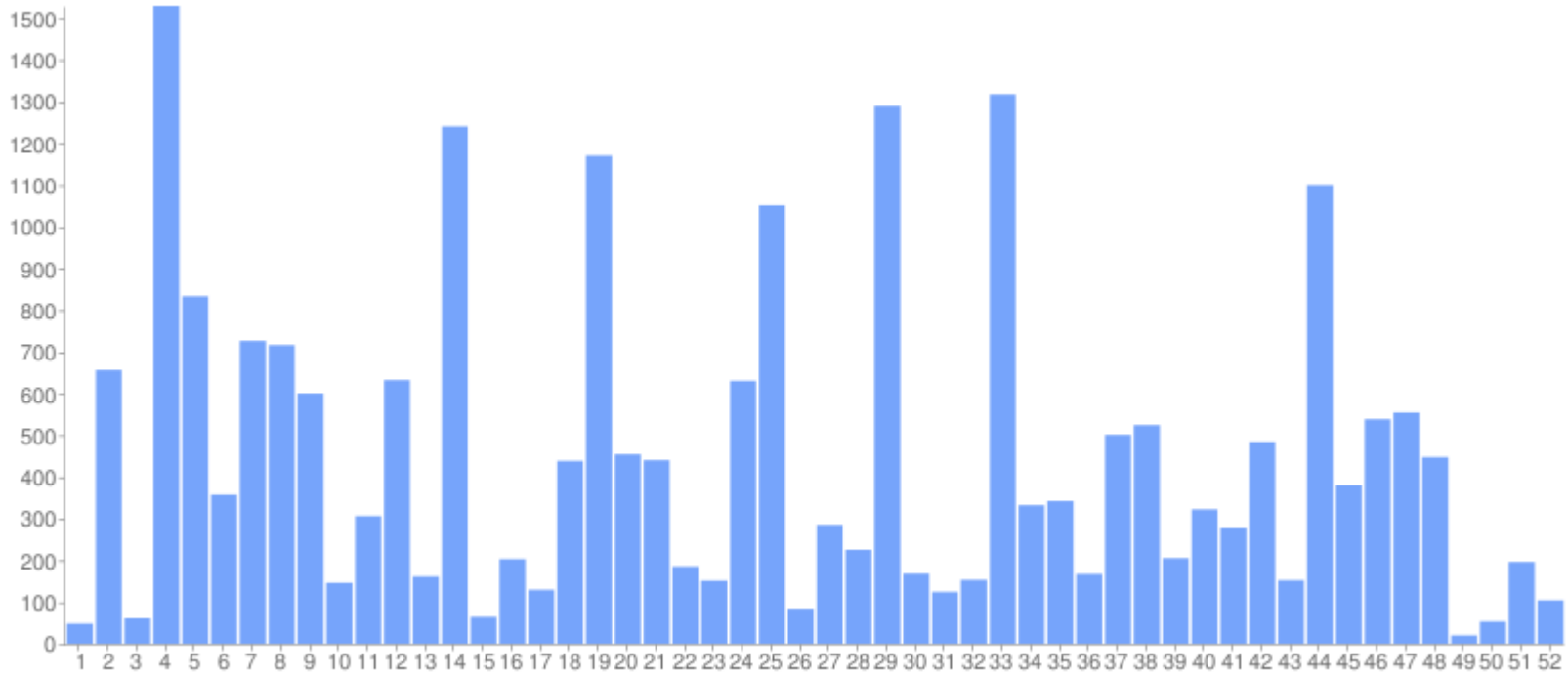


Korpus – dane źródłowe

Korpus Wiki_econo

- wybrano artykuły, które dotyczą dziedziny ekonomicznej
- dodano anotację morfosyntaktyczną za pomocą TAKIPI 1.8
- liczba segmentów: 408 221

Rozkład liczby haseł ze słownika w korpusie



Korpus Wiki_econo – pozyskanie

```
'''Rynek pierwotny''' – [[rynek kapitałowy]], na którym następuje sprzedaż nowych
papierów wartościowych bezpośrednio przez [[emitent]]a – [[akcja (prawo)|akcji]]
i [[obligacja|obligacji]] dopuszczonych do obrotu przez [[Komisja Nadzoru
Finansowego|Komisję Nadzoru Finansowego]]. Cenę papierów wartościowych w tym
wypadku ustala emitent, czyli instytucja wystawiająca [[akcja (prawo)|akcje]] lub
[[obligacja|obligacje]] we własnym imieniu. Emisja i zakup papierów wartościowych
na rynku pierwotnym odbywa się za pośrednictwem [[dom maklerski|domów
maklerskich]] lub banków prowadzących działalność maklerską.

== Zobacz też ==
* [[emisja papierów wartościowych]]
* [[oferta publiczna]]
* [[rynek wtórny]]
* [[IPO]]

[[Kategoria:Rynki finansowe]]
[[Kategoria:Giełda]]

[[ar:السوق الأولي]]
[[de:Primärmarkt]]
[[en:Primary market]]
[[fr:Marché primaire]]
[[lt:Pirminė rinka]]
[[mr:प्रायमरी मार्केट]]
```

Korpus Wiki_econo – przykład

Przykład Rynek pierwotny – rynek kapitałowy, na którym następuje sprzedaż nowych papierów wartościowych bezpośrednio przez emitenta – akcji i obligacji dopuszczonych do obrotu przez Komisję Nadzoru Finansowego. Cenę papierów wartościowych w tym wypadku ustala emitent, czyli instytucja wystawiająca akcje lub obligacje we własnym imieniu. Emisja i zakup papierów wartościowych na rynku pierwotnym odbywa się za pośrednictwem domów maklerskich lub banków prowadzących działalność maklerską.

Korpus Wiki_econo

Korpus został uzyskany poprzez:

- wybranie kategorii ekonomicznych najwyższego rzędu,
- dodanie kategorii potomnych, które również były ekonomiczne,
- pobranie wszystkich artykułów z wynikowej listy kategorii,
- usunięcie z nich fragmentów poza główną treścią artykułu (odnośniki, bibliografia, inne języki, itp.) za pomocą biblioteki jwpl,
- przetworzenie za pomocą TaKIPI 1.8,
- przekonwertowanie do formatu TEI,
- ostateczna lista kategorii obejmowała 81 kategorii i ok. 3200 stron (artykułów),
- dane pochodzą z końca kwietnia 2011.

Kategorie wyjściowe: Ekonomia, Ekonometria, Makroekonomia, Polityka cenowa, Mikroekonomia, Międzynarodowe stosunki gospodarcze, Problemy ekonomiczne, Rynki, Rachunkowość, Finanse, Prawo gospodarcze, Gospodarka, Handel, Produkcja, Usługi.

Korpus – anotacja

Przebieg anotacji

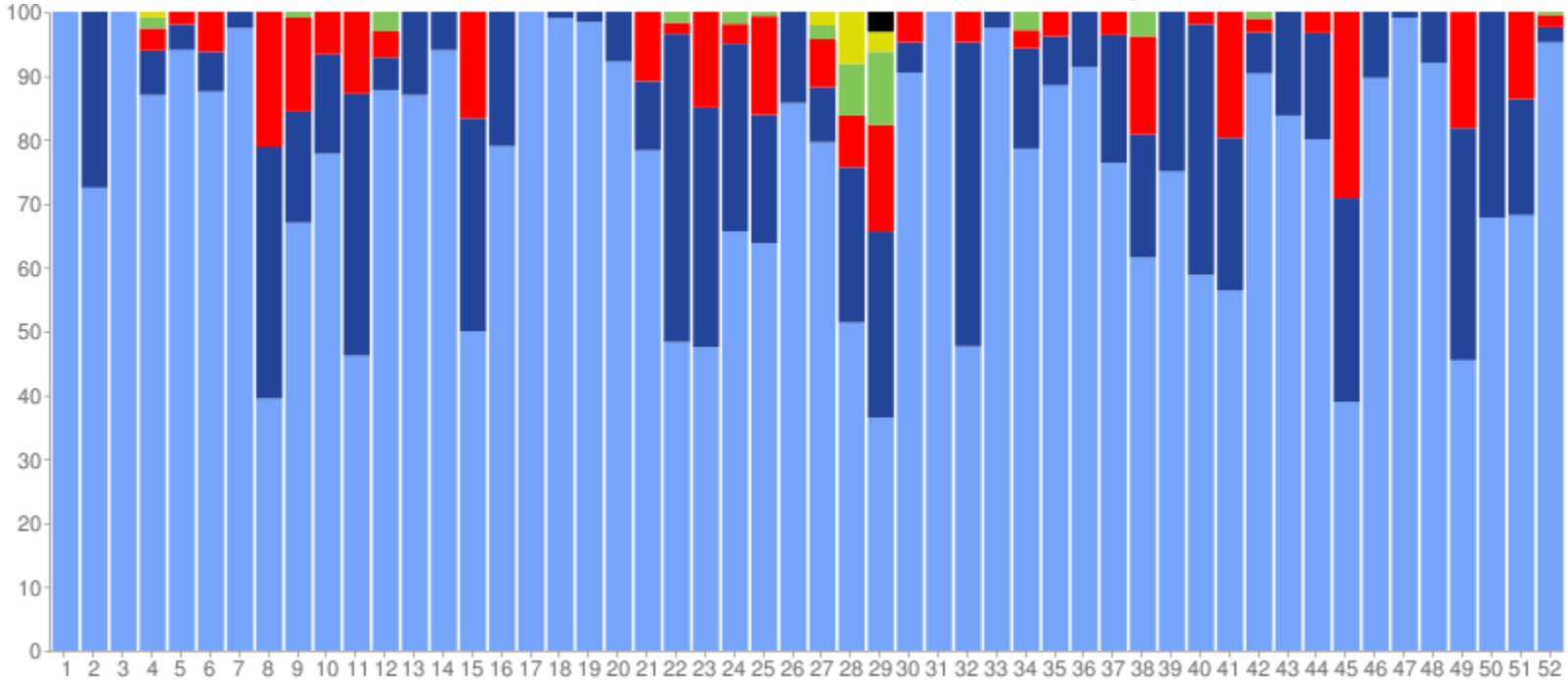
- wykorzystane narzędzie – AnotEk,
- te same transze przypisywane są do dwóch anotatorów,
- w przypadku konfliktu anotacji, transza wraca do anotatorów z prośbą o skomentowanie podjętej decyzji,
- transze skomentowane trafiają do superanotatora, który przypisuje anotację ostateczną.

Podziękowania

- Łukasz Szałkiewicz – SuperAnotator,
- Izabela Will – utworzenie pierwszej wersji słownika,
- wszyscy anotatorzy.

Wyniki anotacji korpusów

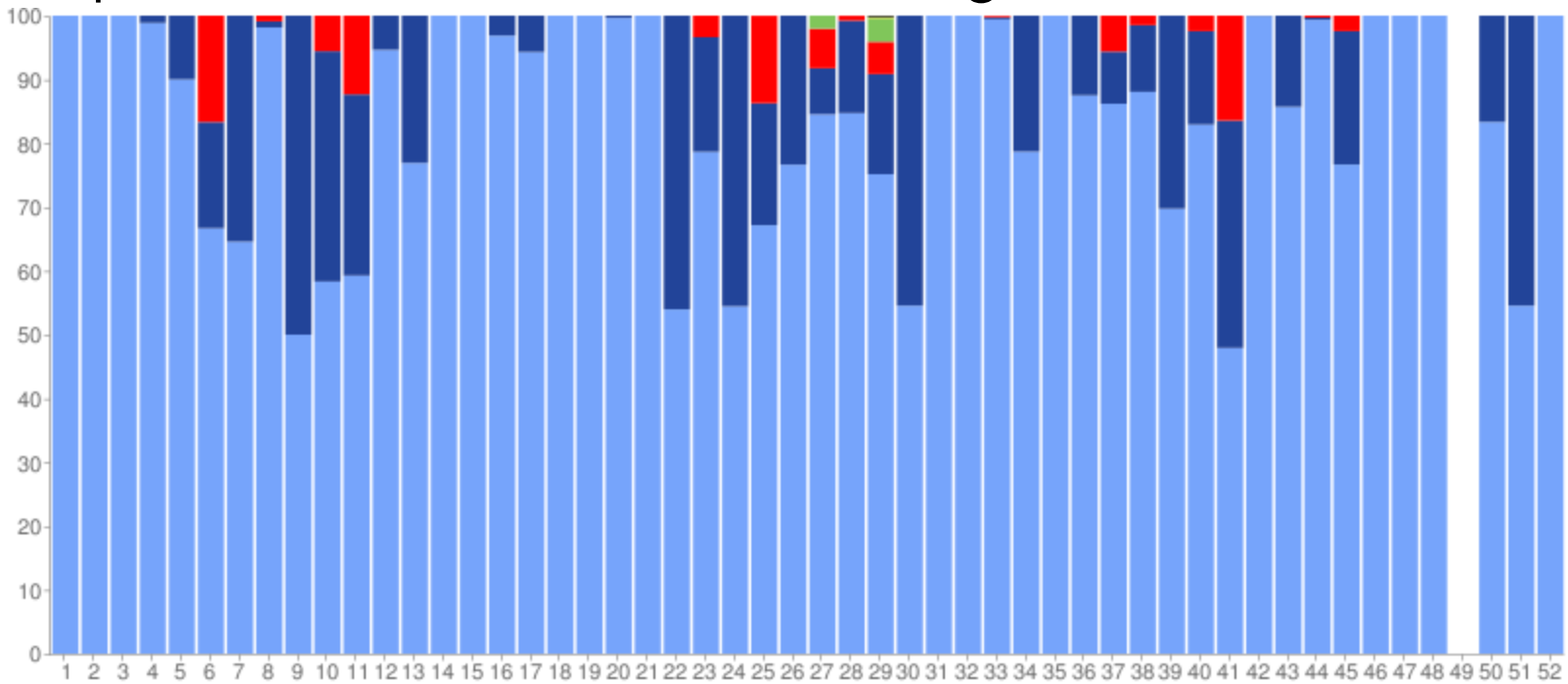
- Korpus NKJP_econo – zaanotowano 3 821 segmentów.



- Korpus GPW – zaanotowano 18 719 segmentów.
- Korpus WIKI_econo – zaanotowano 23 269 segmentów.
- Łącznie – zaanotowano 45 809 segmentów.

Wyniki anotacji korpusów

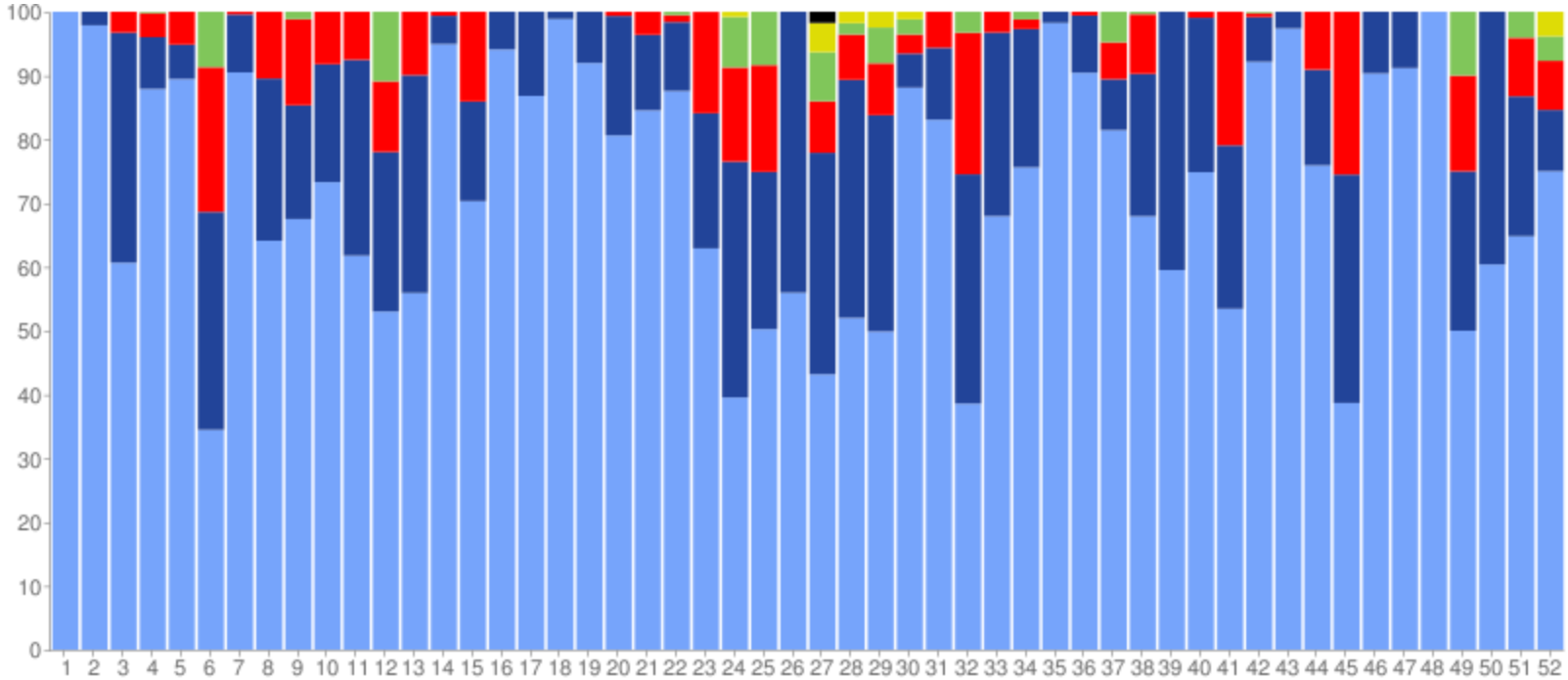
- Korpus NKJP_econo – zaanotowano 3 821 segmentów.
- Korpus GPW – zaanotowano 18 719 segmentów.



- Korpus WIKI_econo – zaanotowano 23 269 segmentów.
- Łącznie – zaanotowano 45 809 segmentów.

Wyniki anotacji korpusów

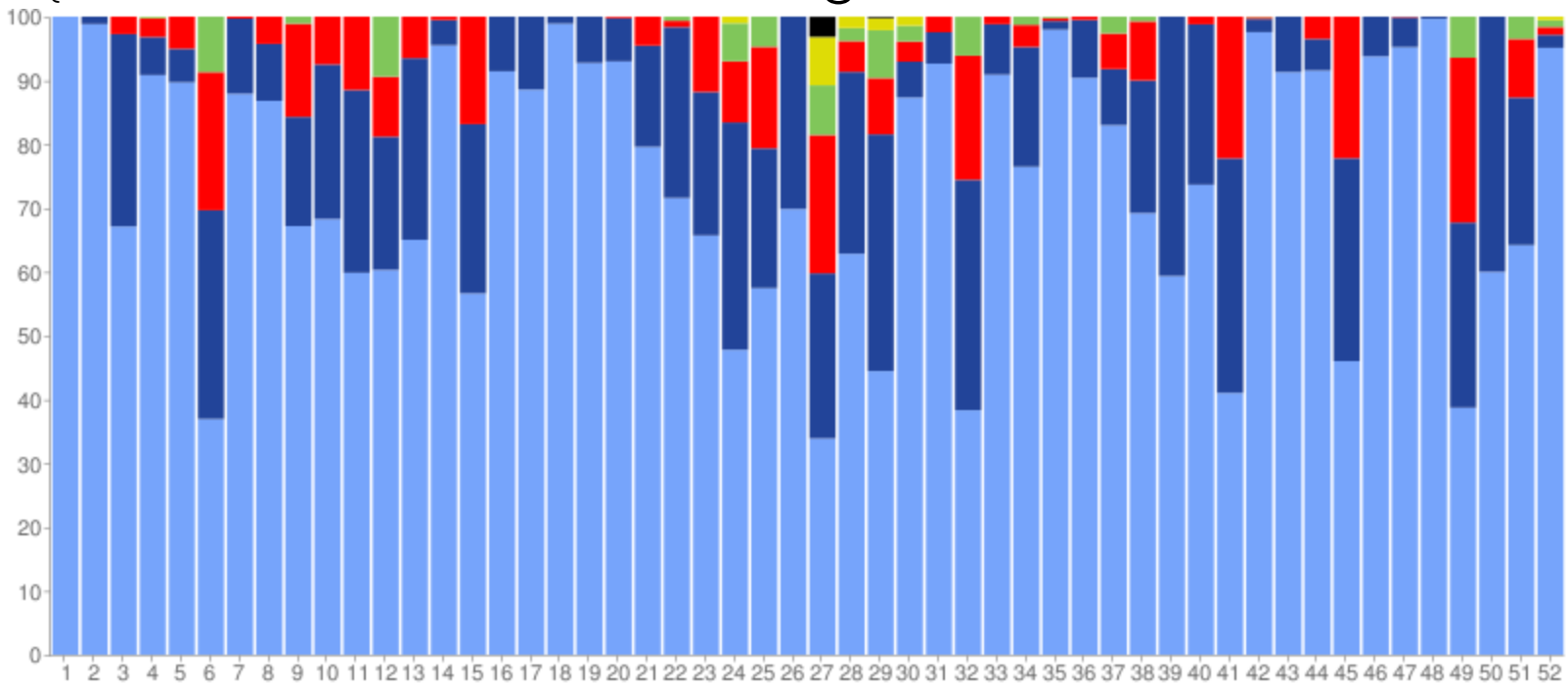
- Korpus NKJP_econo – zaanotowano 3 821 segmentów.
- Korpus GPW – zaanotowano 18 719 segmentów.
- Korpus WIKI_econo – zaanotowano 23 269 segmentów.



- Łącznie – zaanotowano 45 809 segmentów.

Wyniki anotacji korpusów

- Korpus NKJP_econo – zaanotowano 3 821 segmentów.
- Korpus GPW – zaanotowano 18 719 segmentów.
- Korpus WIKI_econo – zaanotowano 23 269 segmentów.
- Łącznie – zaanotowano 45 809 segmentów.



Wyniki anotacji korpusów (2)

Oczekiwane rezultaty metod automatycznych – pomiędzy dolnym a górnym ograniczeniem

- ograniczenie dolne – MFS (Most Frequent Sense),
- ograniczenie górne – ITA (Inter-Annotator Agreement).

Statystyki korpusu

- Korpus NKJP_econo – MFS = 77,65%, ITA = 91,97%.
- Korpus GPW – MFS = 94,31%, ITA = 96,82%.
- Korpus WIKI_econo – MFS = 74,76%, ITA = 90,58%.
- Łącznie – MFS = 81,80%, ITA = 93,25%.

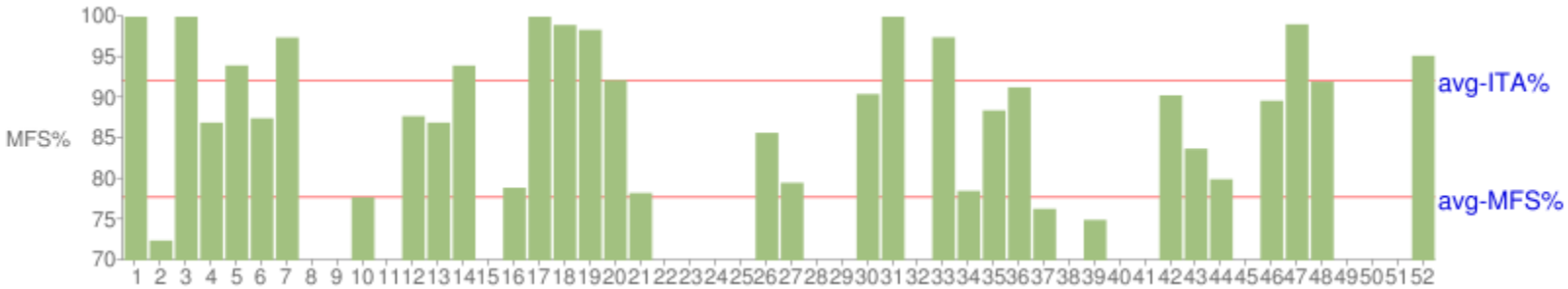
Wyniki anotacji korpusów (2)

Oczekiwane rezultaty metod automatycznych – pomiędzy dolnym a górnym ograniczeniem

- ograniczenie dolne – MFS (Most Frequent Sense),
- ograniczenie górne – ITA (Inter-Annotator Agreement).

Statystyki korpusu

- Korpus NKJP_econo – MFS = 77,65%, ITA = 91,97%.



- Korpus GPW – MFS = 94,31%, ITA = 96,82%.
- Korpus WIKI_econo – MFS = 74,76%, ITA = 90,58%.
- Łącznie – MFS = 81,80%, ITA = 93,25%.

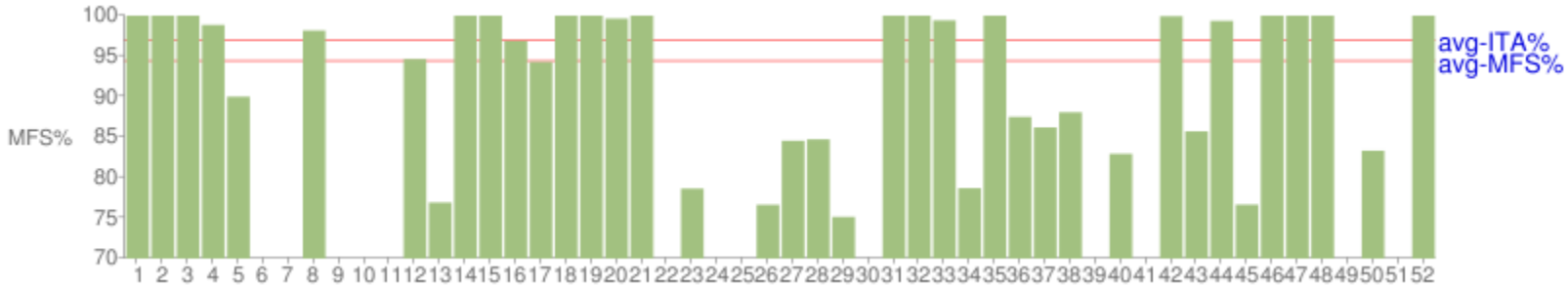
Wyniki anotacji korpusów (2)

Oczekiwane rezultaty metod automatycznych – pomiędzy dolnym a górnym ograniczeniem

- ograniczenie dolne – MFS (Most Frequent Sense),
- ograniczenie górne – ITA (Inter-Annotator Agreement).

Statystyki korpusu

- Korpus NKJP_econo – MFS = 77,65%, ITA = 91,97%.
- Korpus GPW – MFS = 94,31%, ITA = 96,82%.



- Korpus WIKI_econo – MFS = 74,76%, ITA = 90,58%.
- Łącznie – MFS = 81,80%, ITA = 93,25%.

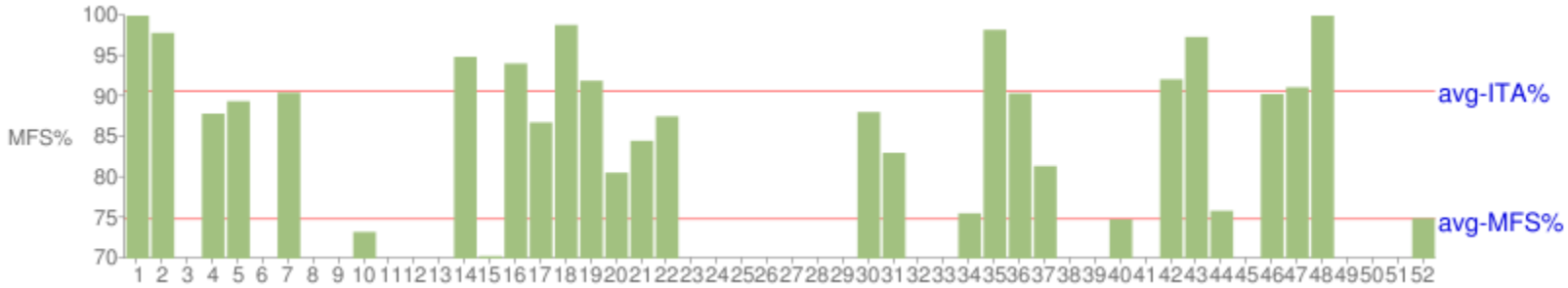
Wyniki anotacji korpusów (2)

Oczekiwane rezultaty metod automatycznych – pomiędzy dolnym a górnym ograniczeniem

- ograniczenie dolne – MFS (Most Frequent Sense),
- ograniczenie górne – ITA (Inter-Annotator Agreement).

Statystyki korpusu

- Korpus NKJP_econo – MFS = 77,65%, ITA = 91,97%.
- Korpus GPW – MFS = 94,31%, ITA = 96,82%.
- Korpus WIKI_econo – MFS = 74,76%, ITA = 90,58%.



- Łącznie – MFS = 81,80%, ITA = 93,25%.

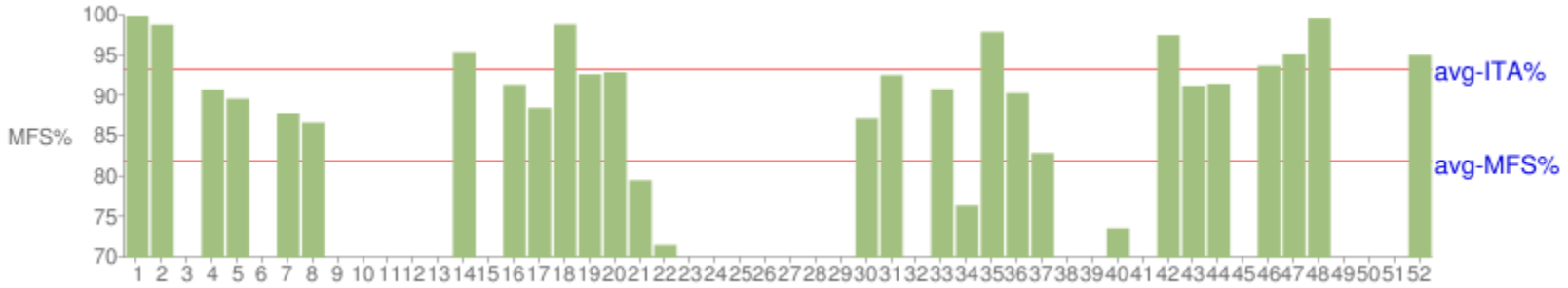
Wyniki anotacji korpusów (2)

Oczekiwane rezultaty metod automatycznych – pomiędzy dolnym a górnym ograniczeniem

- ograniczenie dolne – MFS (Most Frequent Sense),
- ograniczenie górne – ITA (Inter-Annotator Agreement).

Statystyki korpusu

- Korpus NKJP_econo – MFS = 77,65%, ITA = 91,97%.
- Korpus GPW – MFS = 94,31%, ITA = 96,82%.
- Korpus WIKI_econo – MFS = 74,76%, ITA = 90,58%.
- Łącznie – MFS = 81,80%, ITA = 93,25%.



Metody uczenia maszynowego

Przebadanie metod uczenia maszynowego, uzyskujących najlepsze wyniki dla języka angielskiego

- wykorzystujemy implementacje znanych metod uczenia maszynowego i narzędzia WSDDE do przeprowadzenia eksperymentów,
- budujemy klasyfikatory dla każdego ujednoznacznianego hasła,
- określamy skuteczność klasyfikacji w podejściu 10CV (dziesięciokrotna walidacja krzyżowa),
- dla każdego hasła przeglądamy przestrzeń metod określenia istotności atrybutów, metod klasyfikacji oraz parametrów reprezentacji danych (np. wielkości okna).

Rezultat

- istotna poprawa wyniku względem MFS (wyboru najczęstszego sensu).

Zastosowanie metod uczenia maszynowego – reprezentacja danych

W jaki sposób reprezentować ujednoznaczniane hasło w pewnym kontekście za pomocą wektora cech liczbowych?

Przykład

Płacąc np. złotówkę za kilogram sprowadzonego mięsa, przetwarzając go wraz z innym obniżyły sobie koszty surowca, a więc zwiększyły swoje dochody.

- występowanie słów w dużym oknie wokół hasła,

płacić	cena	złotówka	moralność	kilogram	przetwarzać
1	0	1	0	1	1

- występowanie słów na pewnej pozycji, w niewielkiej odległości od hasła,
- występowanie form gramatycznych w niewielkiej odległości od hasła,
- forma gramatyczna hasła.

Zastosowanie metod uczenia maszynowego – reprezentacja danych

W jaki sposób reprezentować ujednoznaczniane hasło w pewnym kontekście za pomocą wektora cech liczbowych?

Przykład

Płacąc np. złotówkę za kilogram sprowadzonego mięsa, przetwarzając go wraz z innym obniżyły sobie koszty surowca, a więc zwiększyły swoje dochody.

- występowanie słów w dużym oknie wokół hasła,
- występowanie słów na pewnej pozycji, w niewielkiej odległości od hasła,

obniżyć-2	obniżyć-1	siebie-1	surowiec+1	praca+1
1	0	1	1	0

- występowanie form gramatycznych w niewielkiej odległości od hasła,
- forma gramatyczna hasła.

Zastosowanie metod uczenia maszynowego – reprezentacja danych

W jaki sposób reprezentować ujednoznaczniane hasło w pewnym kontekście za pomocą wektora cech liczbowych?

Przykład

Płacąc np. złotówkę za kilogram sprowadzonego mięsa, przetwarzając go wraz z innym obniżyły sobie koszty surowca, a więc zwiększyły swoje dochody.

- występowanie słów w dużym oknie wokół hasła,
- występowanie słów na pewnej pozycji, w niewielkiej odległości od hasła,
- występowanie form gramatycznych w niewielkiej odległości od hasła,

praet-2	subst-1	adj-1	subst+1
1	0	0	1

- forma gramatyczna hasła.

Zastosowanie metod uczenia maszynowego – reprezentacja danych

W jaki sposób reprezentować ujednoznaczniane hasło w pewnym kontekście za pomocą wektora cech liczbowych?

Przykład

Płacąc np. złotówkę za kilogram sprowadzonego mięsa, przetwarzając go wraz z innym obniżyły sobie koszty surowca, a więc zwiększyły swoje dochody.

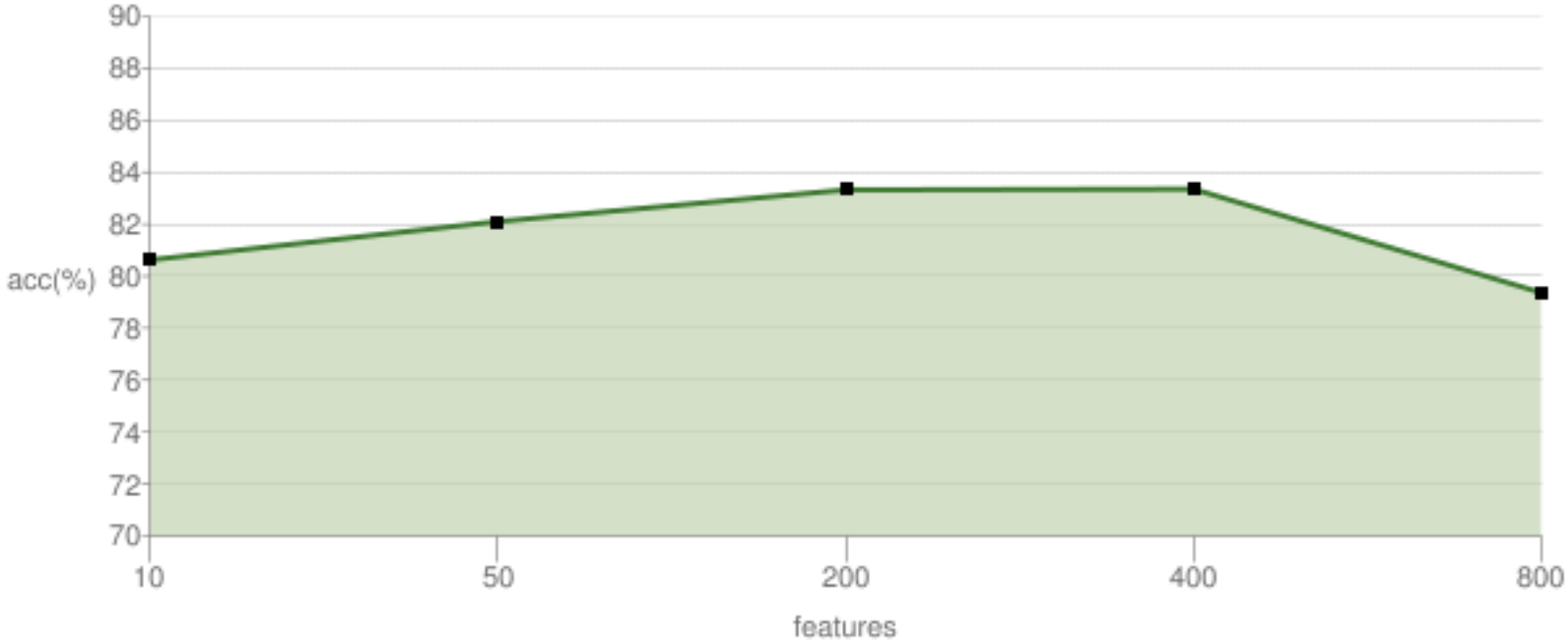
- występowanie słów w dużym oknie wokół hasła,
- występowanie słów na pewnej pozycji, w niewielkiej odległości od hasła,
- występowanie form gramatycznych w niewielkiej odległości od hasła,
- forma gramatyczna hasła.

subst	pl	dat	acc
1	1	0	1

Eksperymenty wstępne

Istotnie wpływają na wynik klasyfikacji (leksem praca):

- liczba wykorzystanych atrybutów (długość wektora cech),

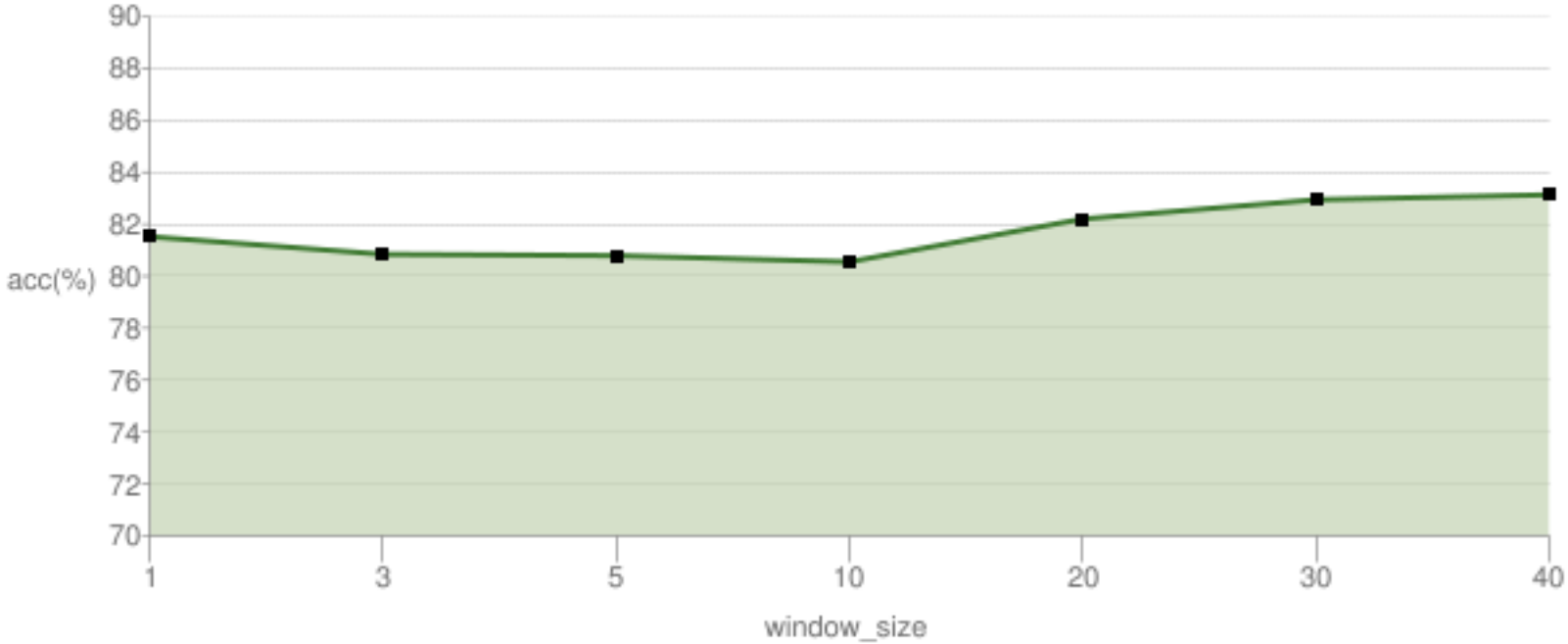


- szerokość okna, w którym badamy występowanie słów.

Eksperymenty wstępne

Istotnie wpływają na wynik klasyfikacji (leksem praca):

- liczba wykorzystanych atrybutów (długość wektora cech),
- szerokość okna, w którym badamy występowanie słów.



Wnioski z eksperymentów na pojedynczych hasłach

Wnioski ogólne

- dla każdego hasła optymalny dobór atrybutów i ich liczby jest inny,
- metody trzeba uczyć indywidualnie dla każdego hasła,
- zwiększanie liczby atrybutów poprawia wyniki, do pewnej wartości granicznej,
- dobre rezultaty dają metody bayesowskie.

Najlepsze uśrednionione wyniki uzyskano:

- wykorzystując formy bazowe słów w analizowanym oknie,
- biorąc pod uwagę występowanie słów oraz form gramatycznych bezpośrednio przed i po haśle ujednoznacznianym,
- uwzględniając interpretację morfosyntaktyczną hasła ujednoznacznianego.

Wnioski z eksperymentów na pojedynczych hasłach

Wnioski ogólne

- dla każdego hasła optymalny dobór atrybutów i ich liczby jest inny,
- metody trzeba uczyć indywidualnie dla każdego hasła,
- zwiększanie liczby atrybutów poprawia wyniki, do pewnej wartości granicznej,
- dobre rezultaty dają metody bayesowskie.

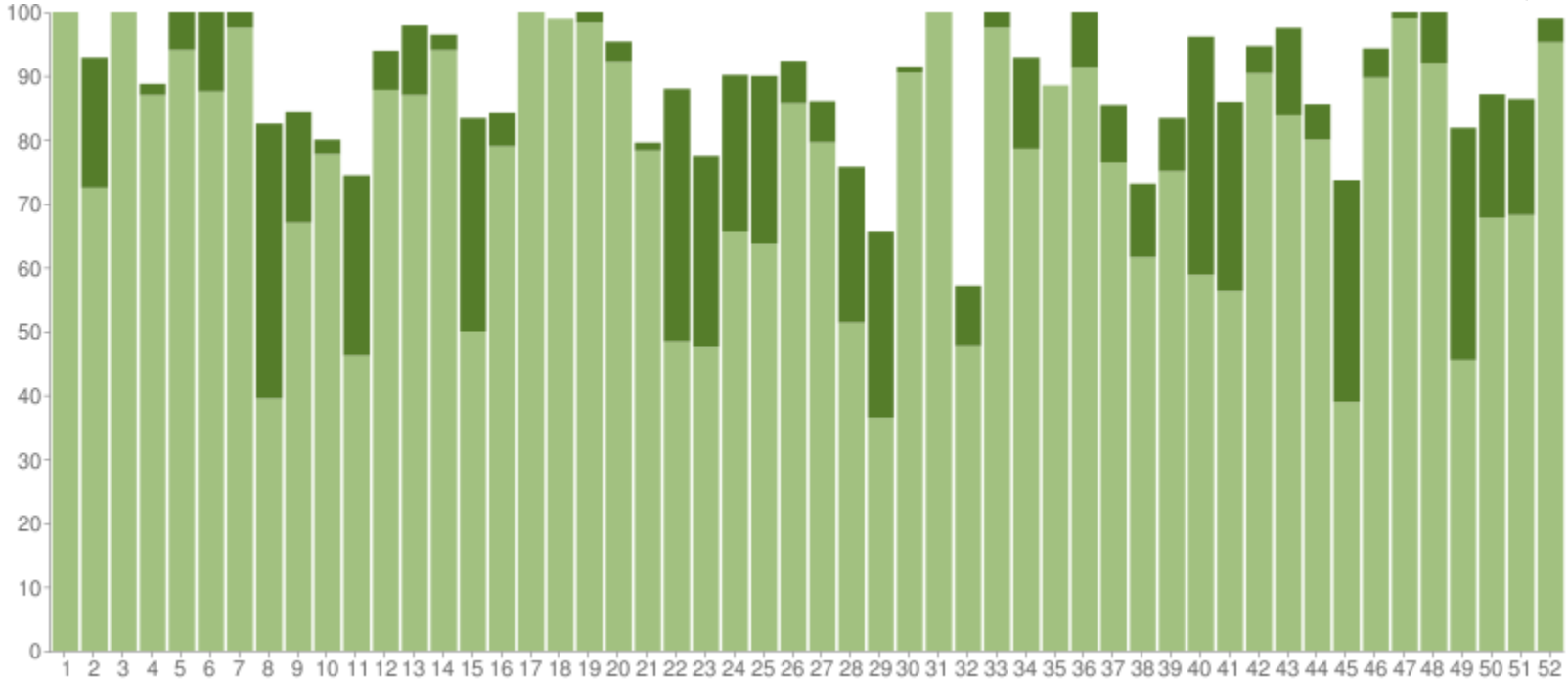
Najlepsze uśrednionione wyniki uzyskano:

- wykorzystując formy bazowe słów w analizowanym oknie,
- biorąc pod uwagę występowanie słów oraz form gramatycznych bezpośrednio przed i po haśle ujednoznacznianym,
- uwzględniając interpretację morfosyntaktyczną hasła ujednoznacznianego.

Wyniki

Otrzymane wyniki skuteczności klasyfikacji (vs MFS)

- korpus NKJP_econo – acc = 90,37%, dla 46 haseł poprawa wyniku

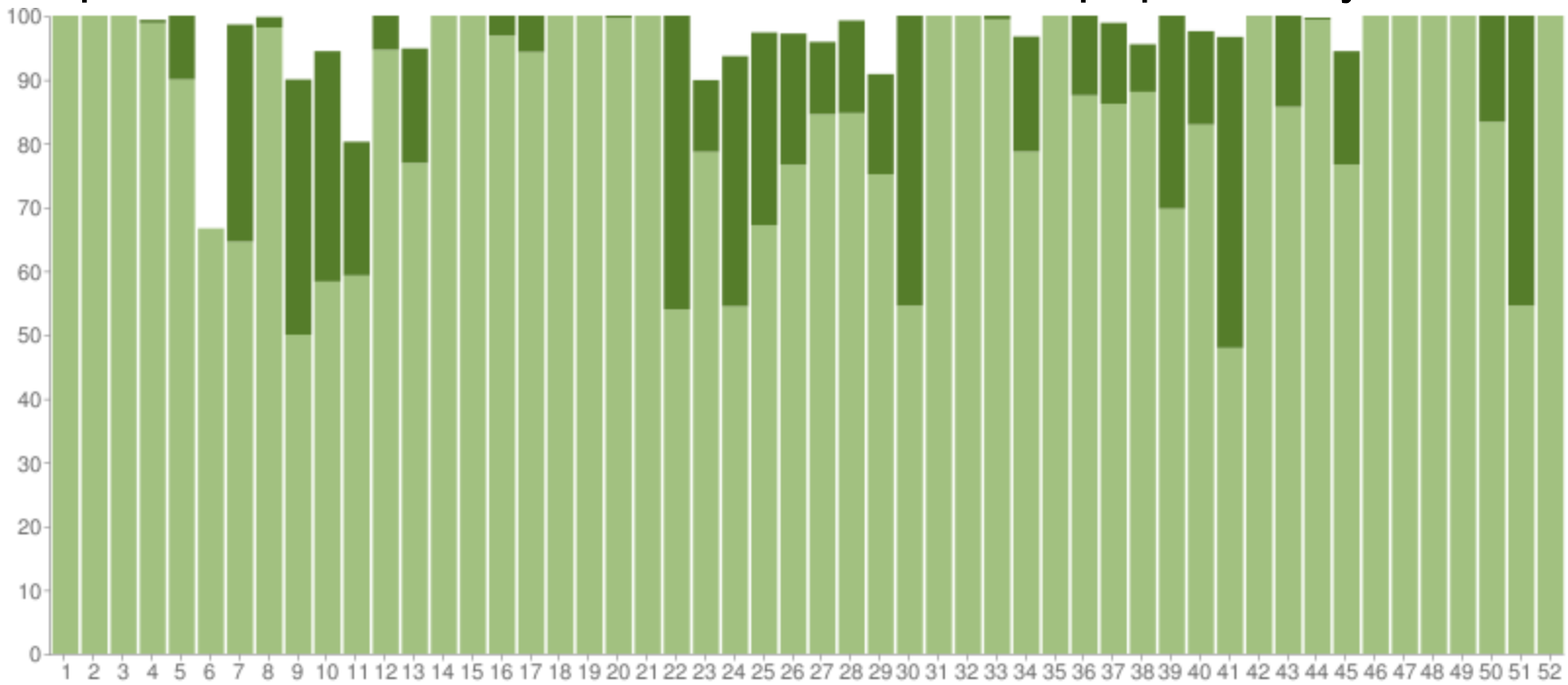


- korpus GPW – acc = 99,13%, dla 36 haseł poprawa wyniku
- korpus Wiki_econo – acc = 90,38%, dla 49 haseł poprawa wyniku

Wyniki

Otrzymane wyniki skuteczności klasyfikacji (vs MFS)

- korpus NKJP_econo – acc = 90,37%, dla 46 haseł poprawa wyniku
- korpus GPW – acc = 99,13%, dla 36 haseł poprawa wyniku

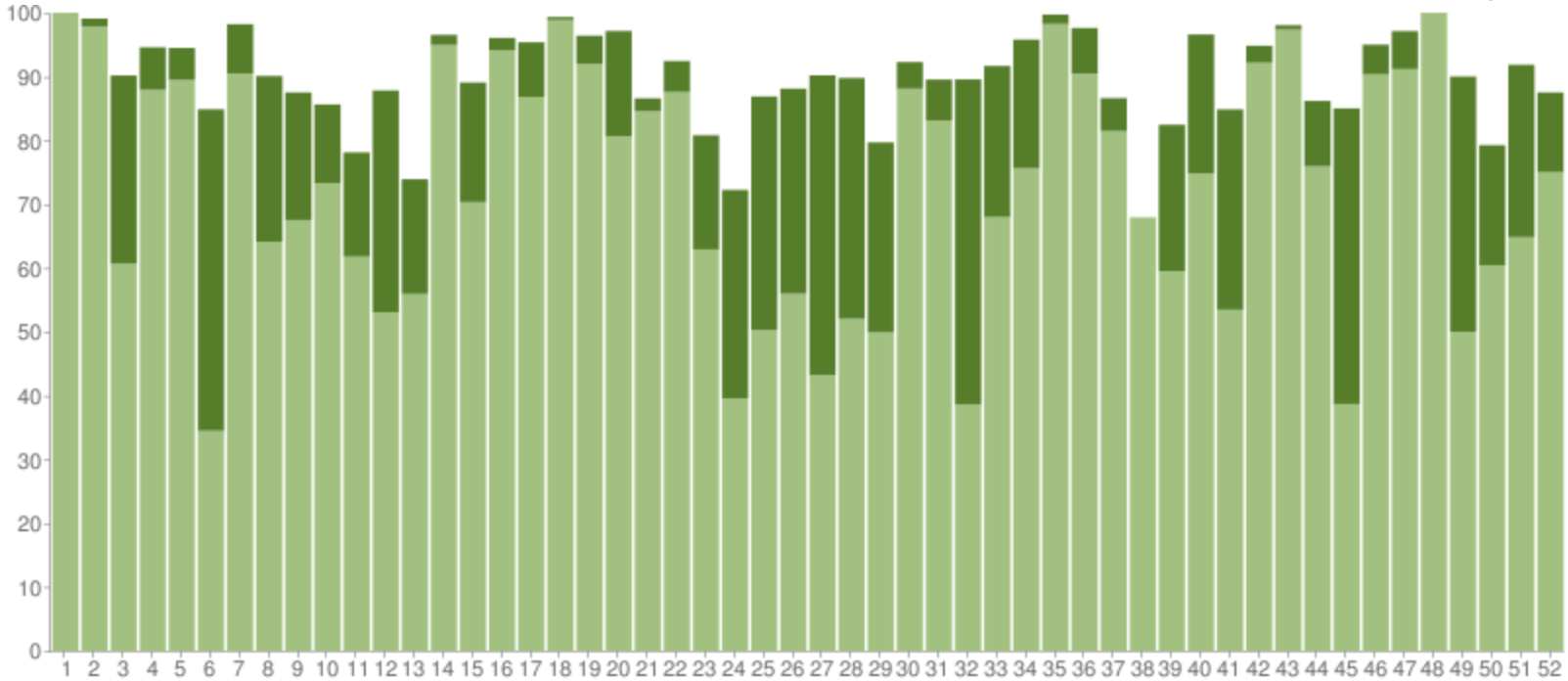


- korpus Wiki_econo – acc = 90,38%, dla 49 haseł poprawa wyniku

Wyniki

Otrzymane wyniki skuteczności klasyfikacji (vs MFS)

- korpus NKJP_econo – acc = 90,37%, dla 46 haseł poprawa wyniku
- korpus GPW – acc = 99,13%, dla 36 haseł poprawa wyniku
- korpus Wiki_econo – acc = 90,38%, dla 49 haseł poprawa wyniku



Wnioski z eksperymentów

Skuteczność metod klasyfikacji

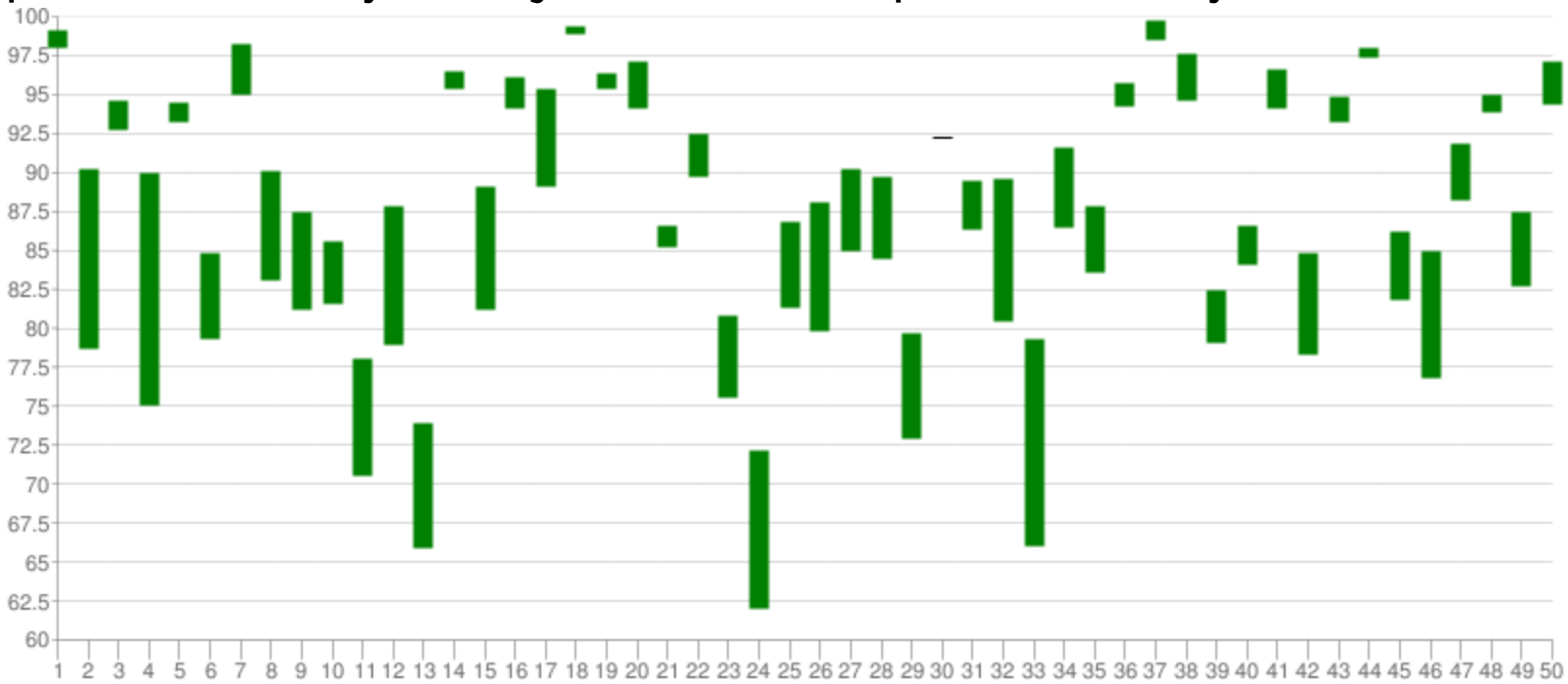
- uzyskana dokładność dla każdej z metod,

- przedział maksymalnej skuteczności przetestowanych metod,
- liczba zwycięstw poszczególnych metod.

Wnioski z eksperymentów

Skuteczność metod klasyfikacji

- uzyskana dokładność dla każdej z metod,
- przedział maksymalnej skuteczności przetestowanych metod,

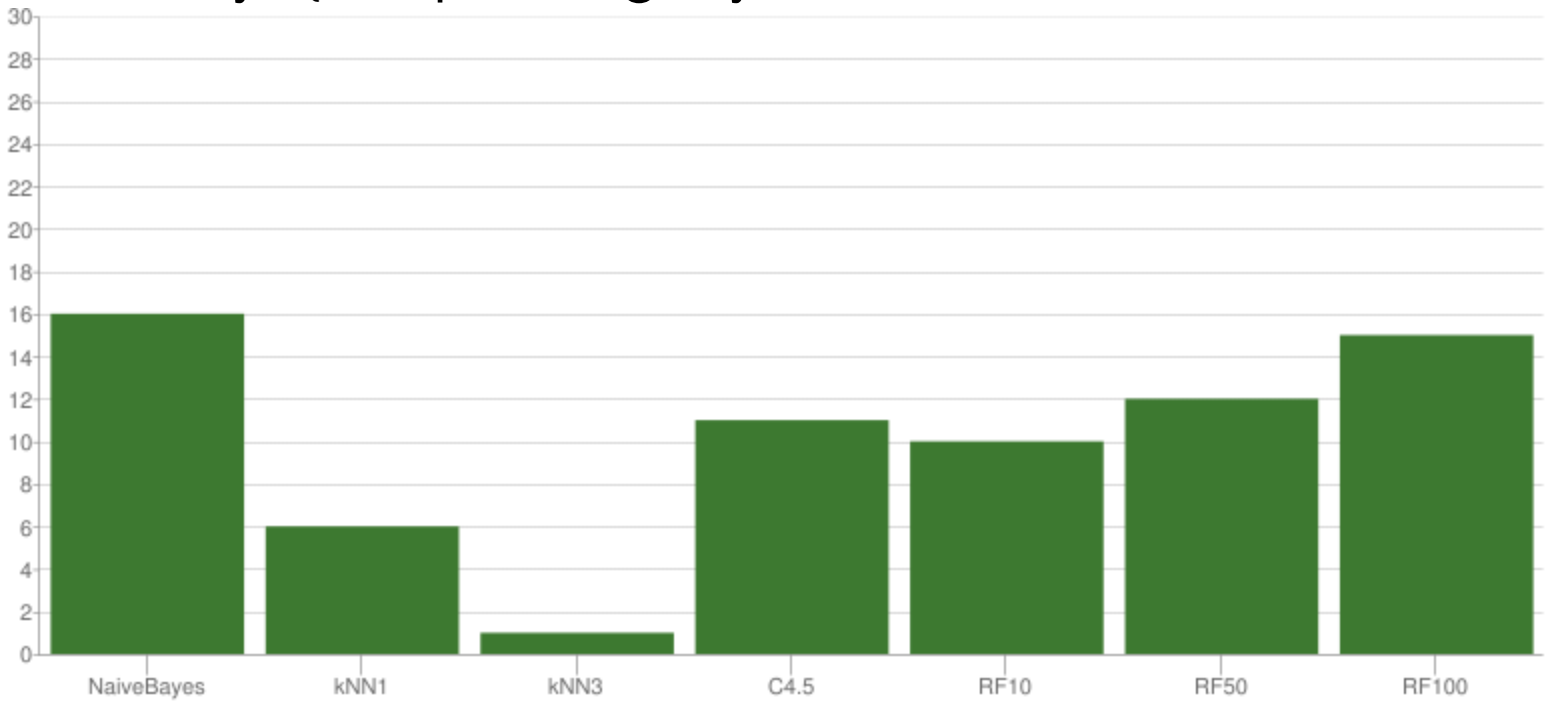


- liczba zwycięstw poszczególnych metod.

Wnioski z eksperymentów

Skuteczność metod klasyfikacji

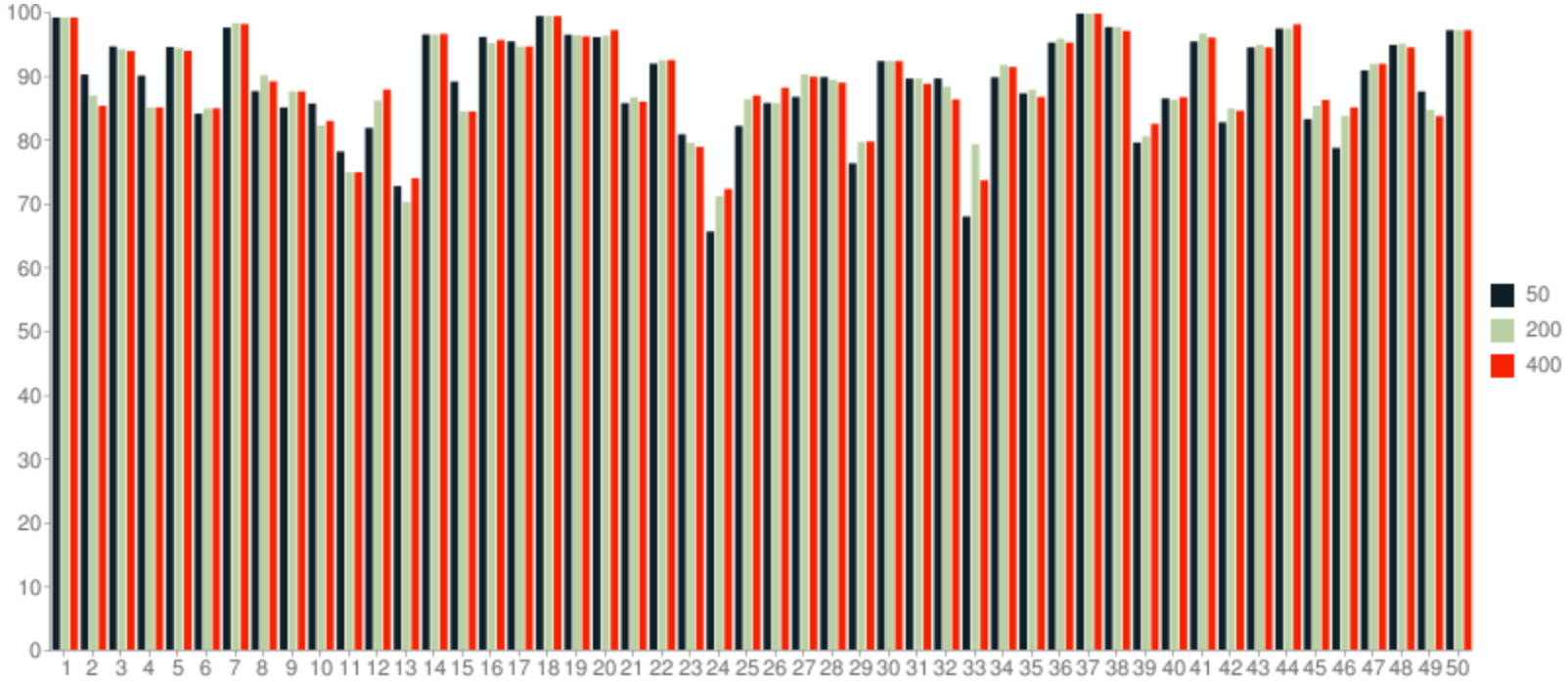
- uzyskana dokładność dla każdej z metod,
- przedział maksymalnej skuteczności przetestowanych metod,
- liczba zwycięstw poszczególnych metod.



Wnioski z eksperymentów

Wpływ liczby atrybutów w ujęciu ogólnym:

- liczba wykorzystanych atrybutów (długość wektora cech),

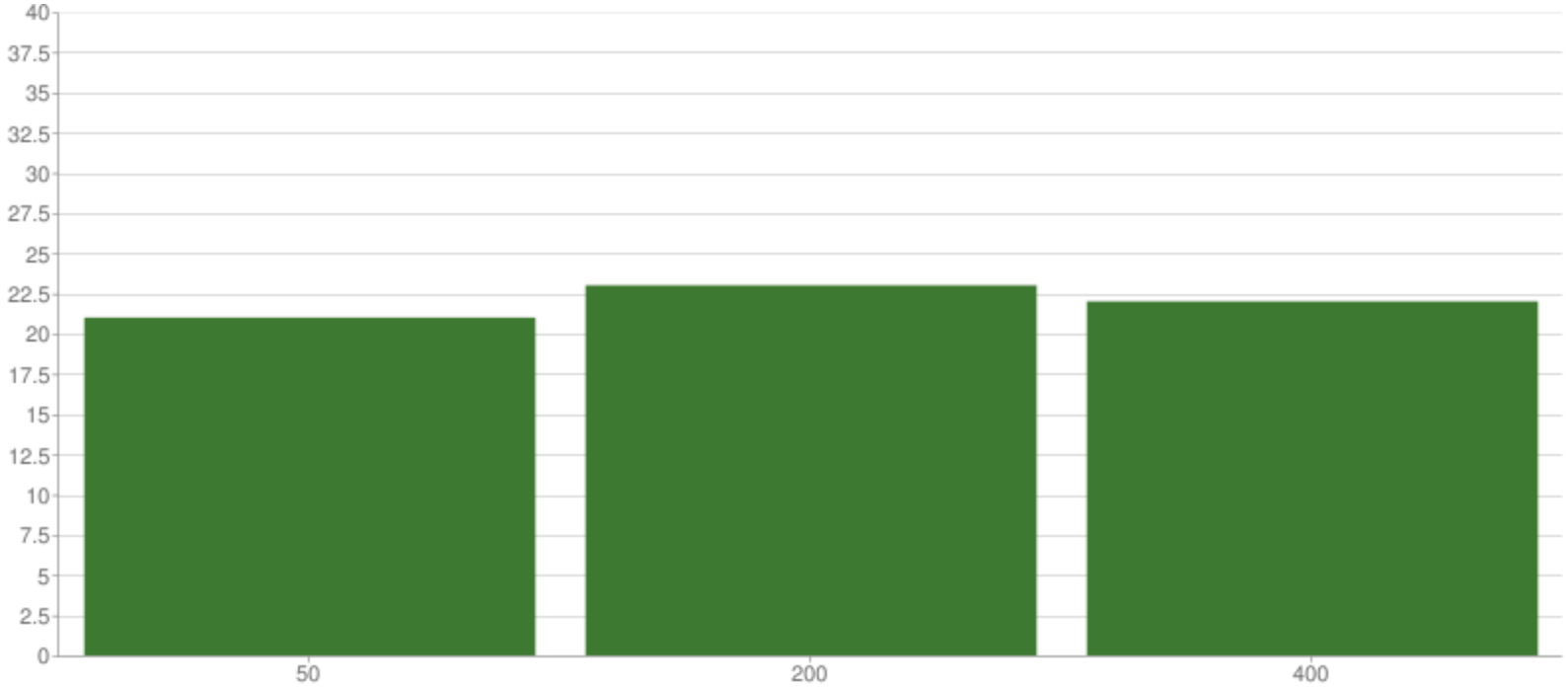


- liczba zwycięstw metody, wykorzystującej poszczególne liczby cech.

Wnioski z eksperymentów

Wpływ liczby atrybutów w ujęciu ogólnym:

- liczba wykorzystanych atrybutów (długość wektora cech),
- liczba zwycięstw metody, wykorzystującej poszczególne liczby cech.



Metody regułowe

Metody wykorzystujące język regułowy

- w projekcie NEKST zaplanowano wykorzystanie formalizmu w postaci logiki pierwszego rzędu, umożliwiającego wyrażanie reguł klasyfikacyjnych, służących do ujednoznaczniania sensów słów.

Wykorzystanie reguł asocjacyjnych do klasyfikacji

- wśród przykładów treningowych szukamy reguł asocjacyjnych, których następnikiem jest etykieta klasy (numer sensu),
- zbiór znalezionych reguł, posortowanych względem zaufania (odsetek pasujących do poprzednika reguły przykładów w zbiorze treningowym, dla których reguła jest prawdziwa) stanowi klasyfikator,
- podczas klasyfikacji przykładu ze zbioru testowego przeglądamy reguły w klasyfikatorze od pierwszej do ostatniej i przypisujemy mu klasę pierwszej pasującej reguły.

Rozszerzenie metody regułowej

Uwzględnianie krotności wystąpień poszczególnych leksemów w kontekście

- zwiększenie skuteczności metody regułowej poprzez odkrywanie klasyfikacyjnych reguł asocjacyjnych z elementami powtarzającymi się.

korpus	MFS	ITA	CAR	rCAR
NKJP_econo	77,65%	91,97%	84,14%	85,16%
GPW	94,31%	96,82%	97,26%	97,86%

Przykład

pl_KFG=0 pos+1_SFG2=0 noun-1_SFG2=1 noun+1_SFG2=0
→ ZNACZENIE=praca.2

miejsce_TFG=2 noun-1_SFG2=1 noun+1_SFG2=0
→ ZNACZENIE=praca.2

Wykorzystanie zewnętrznych zasobów lingwistycznych do poprawy skuteczności metod uczenia maszynowego

Problem – algorytmiczna metoda dezambiguacji sensów jest zbyt silnie związana z konkretnymi leksemami

- decyzja podejmowana jest na podstawie dokładnie tych słów, które występują w tekście,
- brak konkretnego leksemu w zbiorze treningowym uniemożliwia podjęcie prawidłowej decyzji w zbiorze testowym,
- człowiek analizuje kontekst i uogólnia występujące w nim słowa, aby stwierdzić jaki jest jego sens.

Przykład

- Cena jednego kilograma mąki wynosi...
- Cena 1 kg kaszy wynosi...
- Jaka była wówczas cena funta pszenicy?

Wykorzystanie zewnętrznych zasobów lingwistycznych do poprawy skuteczności metod uczenia maszynowego

Pomysł – wykorzystać funkcję podobieństwa semantycznego do rozszerzenia kontekstów słów dezambiguowanych o leksemy podobne

- dążymy do tego, żeby konkretne leksemy zamienić na ogólne pojęcia,
- funkcje podobieństwa semantycznego umożliwiają „wygenerowanie" nowych kontekstów na podstawie istniejących w zbiorze treningowym, co przekłada się na potencjalnie lepsze wytrenowanie metod uczenia maszynowego.

Funkcje podobieństwa semantycznego

Funkcje odzworowujące pary leksemów w liczbę rzeczywistą

- $W \times W \rightarrow \mathbb{R}$,
- liczba określa stopień podobieństwa pomiędzy leksemami.

Funkcje oparte na zasoboach ustrukturalizowanych

- niezbędny zasób typu WordNet z dużą liczbą relacji,
- stopień podobieństwa między leksemami określany na podstawie stopnia powiązania w grafie WordNetu, np. jako odwrotność odległości najkrótszej ścieżki pomiędzy nimi.

Funkcje oparte na korpusie tekstowym

- stopień podobieństwa między leksemami określany na podstawie częstości ich współwystępowania w korpusie,
- dla języka polskiego: RWF (Piasecki et al., 2007): korpus IPI PAN, korpus Rzeczpospolitej oraz dokumentów internetowych.

Korpusowa funkcja podobieństwa semantycznego

kilogram

podobieństwo	jednostka leksykalna
0.299	kg
0.287	kilo
0.241	tona
0.206	gram
0.196	funt
0.182	porcja
0.165	ilość
0.160	dekagram
0.150	zapas
0.144	litr
0.139	garść
0.132	import
0.127	eksport
0.127	kawałek
0.126	produkcja
0.121	deka
0.119	spożycie
0.119	łyżka
0.116	worek
0.107	uncja

http://plwordnet.pwr.wroc.pl/wordnet/msr/kilogram

Wykorzystanie funkcji do rozszerzenia wektora cech

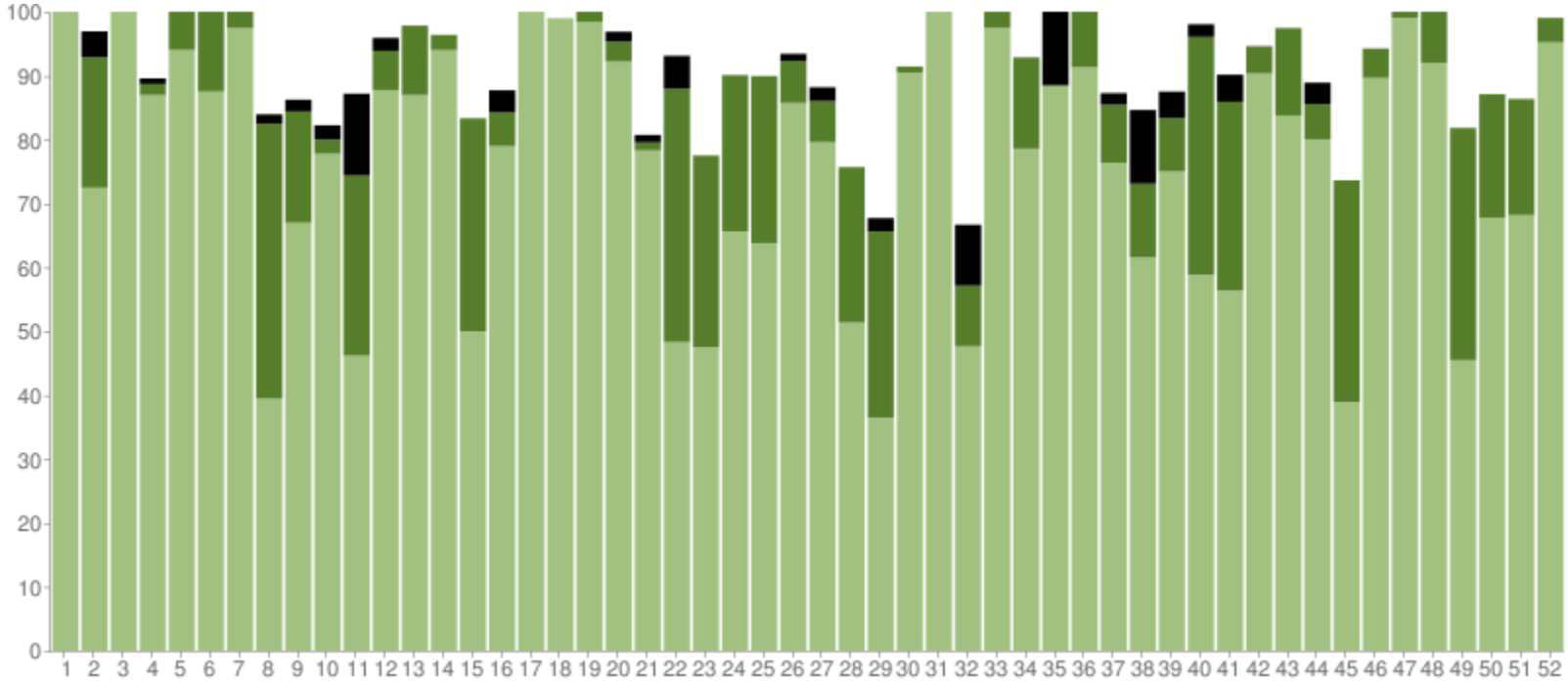
płacić	cena	złotówka	moralność	kilogram	przetwarzać
1	0	1	0	1	1

...	kilogram	kg	kilo	tona	gram	przetwarzać
...	1	1	1	1	1	0

Wyniki

Otrzymane wyniki skuteczności klasyfikacji metodami uczenia nadzorowanego z użyciem FPS (vs bez użycia FPS i MFS)

- korpus NKJP – acc = 91,73%, dla 22 haseł poprawa wyniku

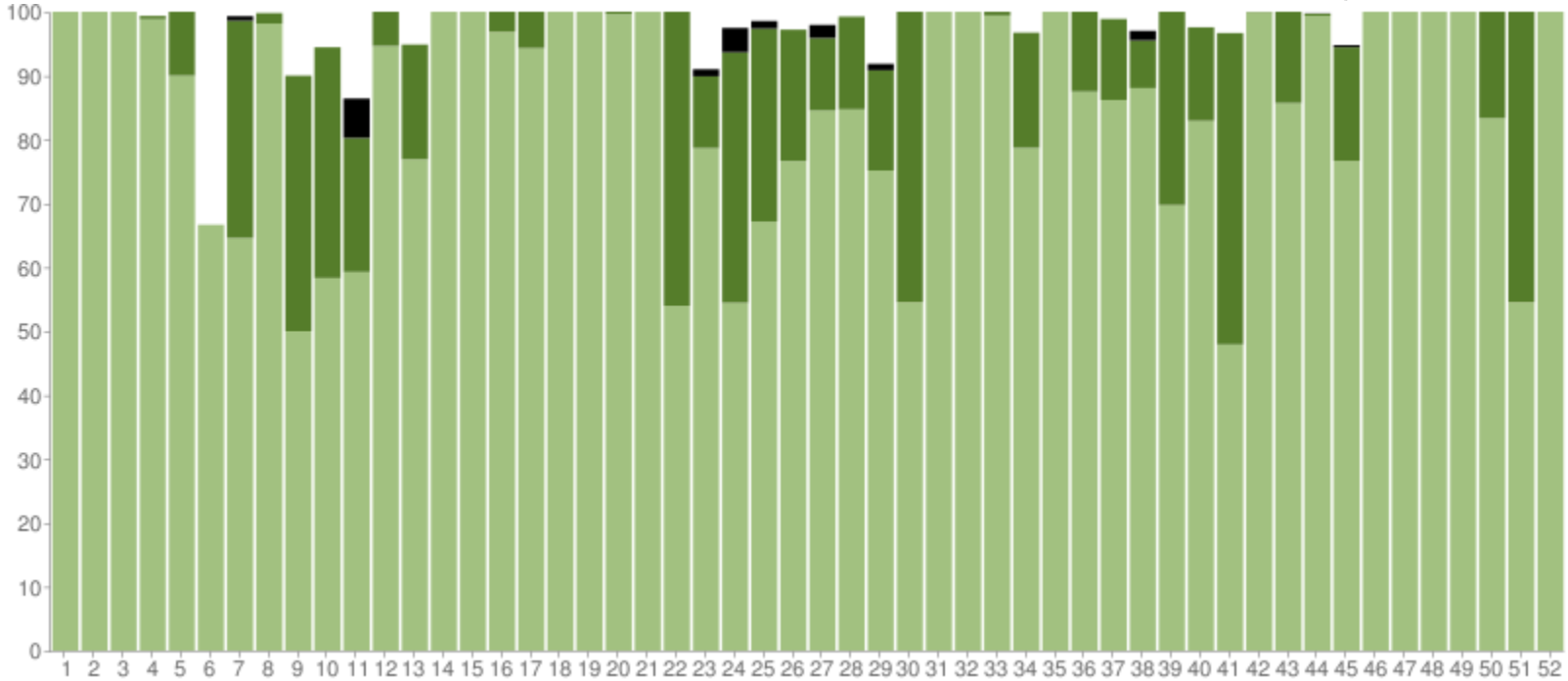


- korpus GPW – acc = 99,28%, dla 10 haseł poprawa wyniku
- połączone korpusy – acc = 97,52%, dla 19 haseł poprawa wyniku

Wyniki

Otrzymane wyniki skuteczności klasyfikacji metodami uczenia nadzorowanego z użyciem FPS (vs bez użycia FPS i MFS)

- korpus NKJP – acc = 91,73%, dla 22 haseł poprawa wyniku
- korpus GPW – acc = 99,28%, dla 10 haseł poprawa wyniku

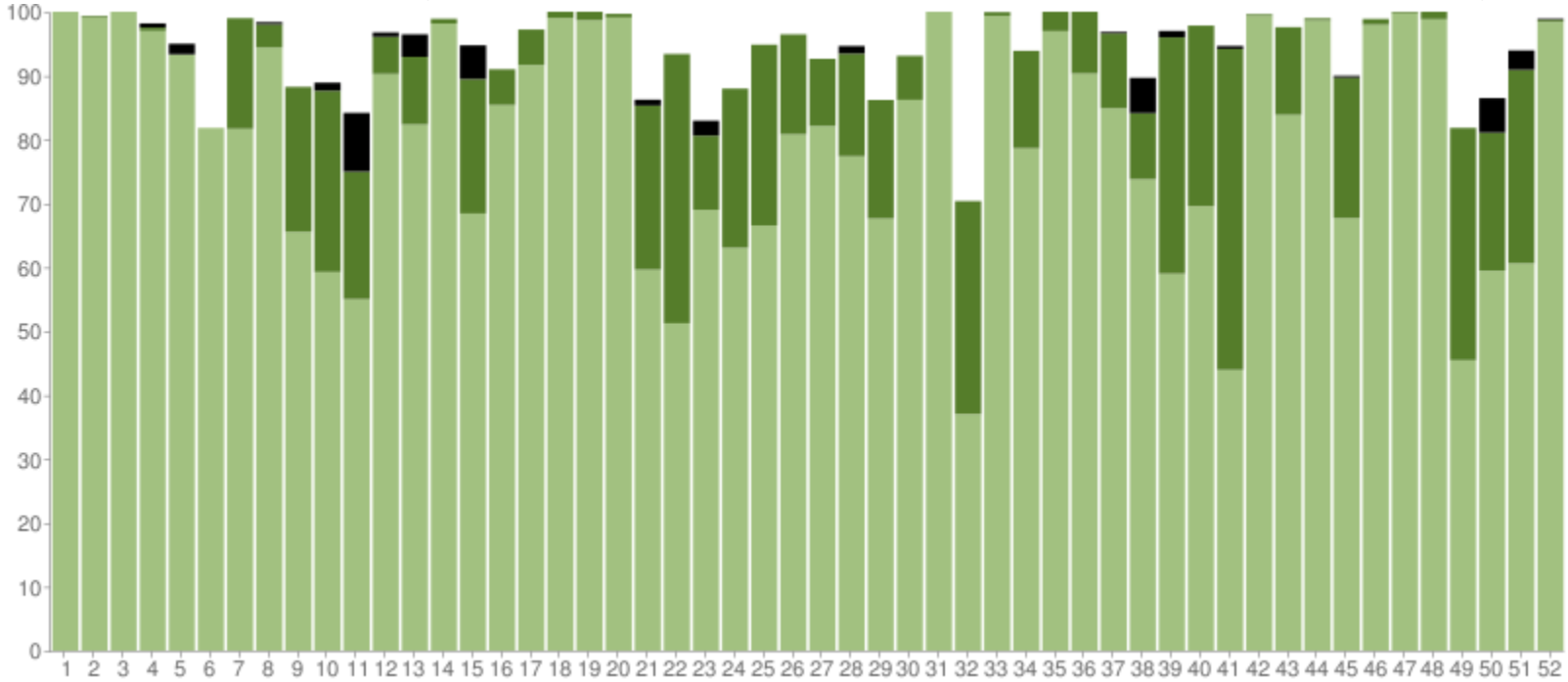


- połączone korpusy – acc = 97,52%, dla 19 haseł poprawa wyniku

Wyniki

Otrzymane wyniki skuteczności klasyfikacji metodami uczenia nadzorowanego z użyciem FPS (vs bez użycia FPS i MFS)

- korpus NKJP – acc = 91,73%, dla 22 haseł poprawa wyniku
- korpus GPW – acc = 99,28%, dla 10 haseł poprawa wyniku
- połączone korpusy – acc = 97,52%, dla 19 haseł poprawa wyniku



Podsumowanie wyników

Wytworzone oprogramowanie

- edytor słownika sensów,
- narzędzie do anotacji korpusu tekstowego sensami ze słownika,
- narzędzie do dezambiguacji (do końca roku).

Wytworzone zasoby tekstowe

- słownik sensów,
- korpus tekstowy z dziedziny ekonomicznej, anotowany tekstami słów, składający się z podkorpusów NKJP_econo, GPW i Wiki_econo.

Podsumowanie wyników

Otrzymane wyniki eksperymentalne

- metoda bazowa – przypisanie słowom wieloznacznym najczęstszego sensu,
- uczenie maszynowe (1) – statystyczne metody klasyfikacji na korpusie podzielonym zgodniem z metodyką 10CV,
- uczenie maszynowe (2) – klasyfikator regułowy z rozszerzeniem o elementy powtarzające się.
- uczenie maszynowe (3) – wykorzystanie zasobów lingwistycznych do poprawy skuteczności dezambiguacji.