Rozróżnianie sensów polskich słów za pomocą rozwinięcia metody Leska

Seminarium przetwarzania języka naturalnego

Mateusz Kopeć

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

6 lutego 2012

Plan

Plan

Word Sense Disambiguation

- Rozróżnianie Sensów Słów - Word Sense Disambiguation
- Problem: wieloznaczność słów
- Zadanie: wybór właściwego znaczenia w danym kontekście
  - **wejście**: zbiór możliwych sensów i kontekst
  - **wyjście**: odpowiedni sens

Przykład

Zamek może oznaczać (m.in.):

1. budowlę,
2. zamek do drzwi.

Przykład

1. Przed nimi widać było ogromny $zamek_1$.
2. Długo się zastanawiał, czy zamknął $zamek_2$ po wyjściu z domu.

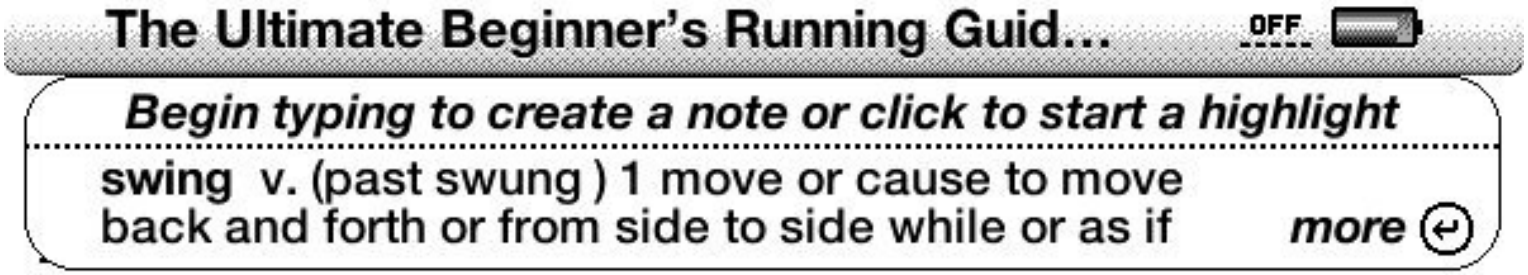
Zastosowania WSD

- Rozwiązenie przydatne w:
  - Tłumaczeniu maszynowym,
  - Question answering,
  - Wyszukiwaniu informacji,
  - Klasyfikacji tekstów,
  - itp.
- Seminaria tutejsze:
  - Mechanizm budowy streszczeń dokumentów w wyszukiwarce NEKST
  - Metody nadzorowane w ujednoznacznianiu sensów słów korpusu dziedzinowego

Przykład

Tłumaczenie: polski $\rightarrow$ angielski. Słowo *piłka*:
piłka $\rightarrow$ *ball*?
piłka $\rightarrow$ *small saw*?

Zastosowania WSD – przykład



Let's start with running form, as good form will greatly help increase your comfort and efficiency on initial runs. The basic mantra for running is, "Run comfortably, but run disciplined." This may sound contradictory, but it reminds us to keep our shoulders, chest, and neck loose while maintaining good posture; to keep our shoulders and elbows moving in a forward direction (not side-to-side) while keeping fluid in all of our movements. We should have a slight *whole body* forward lean, being careful not to bend at the waist. Our running form is broken down into two parts: the stance phase and the swing phase. Stance phase occurs when the foot comes in contact with the ground, while

Pokrycie tekstu

Możliwe są dwa podejścia do WSD:

- Rozróżnianie sensów wybranego słowa lub małej liczby słów w kontekście (*lexical sample*),
- Rozróżnianie sensów wszystkich słów w kontekście (*all-words*).

Słownik sensów

Wybór słownika sensów i jego cechy zależą od zastosowania:

- W tłumaczeniach trzeba rozróżniać sensy różnie przekładane,
- W syntezie mowy wystarczy rozróżniać homografy,
- W praktyce najczęściej podział jest gruboziarnisty.

Sposoby rozwiązywania

Klasyfikacja z nadzorem

Zalety:

+ Wysoka skuteczność.

Wady:

- Duży koszt ręcznej anotacji przykładów treningowych.

Klasyfikacja bez nadzoru

Zalety:

+ Niski koszt.
+ Szeroka stosowalność.

Wady:

- Mniejsza skuteczność.

Temat pracy

Pierwszy plan:

- Rozróżnianie sensów wszystkich słów w kontekście (*all-words*),
- Uczenie z minimalnym nadzorem, oparte na idei *boostrapingu*,
- Tym samym metoda możliwie ogólna.

Problemy:

- Brak testowego zbioru do ewaluacji.
- Brak odpowiedniego słownika sensów.
- Trudności w stworzeniu samodzielnie takich zasobów.

Temat pracy

Co udało się znaleźć:

- Odpowiedni dla WSD słownik dla 106 często występujących słów wieloznacznych,
- Duży (>1 mln słów) zbiór tekstów do ewaluacji, oznaczony sensami z powyższego słownika.

Drugi plan:

- Potencjalne rozróżnianie sensów wszystkich słów w kontekście (*all-words*),
- w praktyce tylko wybrane słowa, dla których istnieją wpisy w słowniku.
- Brak fazy uczenia, algorytm opiera się na porównaniu definicji sensów z kontekstem słowa.

Plan

Algorytm Leska

Jeden z pierwszych algorytmów (1986) opartych na wiedzy.
Porównywane są **definicje** wielu sensów wielu słów.

Niech słowa *W*1 i *W*2 występują we wspólnym kontekście.
Wtedy działanie jest następujące:

- dla każdego sensu *i* słowa *W*1,
  - dla każdego sensu *j* słowa *W*2,
    - oblicz *Przecięcie(i, j)*, równe liczbie słów wspólnych w definicji sensu *i* oraz *j*,
- znajdź *i* i *j*, dla których *Przecięcie(i, j)* jest największe,
- przypisz sens *i* słowu *W*1 i sens *j* słowu *W*2.

Uproszczony algorytm Leska

Uproszczony algorytm Leska porównuje **kontekst** z **definicją** dla pojedynczego słowa wieloznacznego.

Działa następująco:

- dla każdego sensu *i* słowa *W* w kontekście *c*,
  - oblicz *Przecięcie(i, c)*, które jest równe ilości wspólnych słów w definicji *i* oraz kontekście *c*,
- znajdź *i*, dla którego *Przecięcie(i, c)* jest największe,
- przypisz sens *i* słowu *W*.

Jak to działa? - przykład

Słowo: zamek

1. Obronna budowla z kamienia.
2. Urządzenie pozwalające na zamykanie czegoś, np. zamek w drzwiach.

Kontekst

... po czym zamknął **zamek** w drzwiach, kilkukrotnie się upewniając, czy z pewnością ...

Działanie algorytmu:

- Porównanie definicji 1. z kontekstem,
- Porównanie definicji 2. z kontekstem,
- Wybór bardziej podobnej definicji sensu.

Rozszerzony algorytm Leska

Rozszerzony algorytm Leska porównuje **kontekst** z **definicją** dla pojedynczego słowa wieloznacznego.

Różni się wyłącznie tym, że używa bardziej wyszukanej funkcji podobieństwa zamiast zwykłego przecięcia tekstowego.
Działa następująco:

- dla każdego sensu *i* słowa *W* w kontekście *c*,
  - oblicz *Podobieństwo(i, c)*,
- znajdź *i*, dla którego *Podobieństwo(i, c)* jest największe,
- przypisz sens *i* słowu *W*.

Obliczanie podobieństwa kontekstu i definicji

W skrócie:

1. Z kontekstu jest tworzony wektor:
   - współrzędne – słowa z kontekstu,
   - wartości – pewne wagi tychże słów.
2. Z definicji sensu jest tworzony analogiczny wektor.
3. Wektory są porównywane poprzez np. ich przemnożenie przez siebie.

Obliczanie podobieństwa kontekstu i definicji I

Parametryzacja tego obliczenia:

- **Rozmiar definicji** – krótka, długa.
- **Rozmiar kontekstu** – 1-50.
- **Lematyzacja** – tak/nie.
- **Binaryzacja** – tak/nie.
- **IDF** – tak/nie.
- **IGF** – tak/nie.
- **Dolne wartości skrajne** – między 0 a 1.
- **Górne wartości skrajne** – między 0 a 1.
- **Normalizacja definicji** – tak/nie.
- **Normalizacja kontekstu** – brak/liniowa/kwadratowa.

Obliczanie podobieństwa kontekstu i definicji II

- **Rozszerzenie** – 7 możliwości:
  - brak rozszerzenia,
  - naiwne, za pomocą synonimów ze słownika `synonimy.ux.pl`,
  - ostrożne, za pomocą synonimów ze słownika `synonimy.ux.pl`,
  - naiwne, za pomocą synonimów ze Słowosieci, pobieranych z synsetów zawierających słowo bazowe,
  - ostrożne, za pomocą synonimów ze Słowosieci, pobieranych z synsetów zawierających słowo bazowe,
  - naiwne, za pomocą synonimów ze Słowosieci, pobieranych z synsetów zawierających słowo bazowe oraz synsetów znajdujących się w określonej relacji (hiperonimii, części, bliskoznaczności i fuzzynimii synsetów) z nimi,
  - za pomocą słów znajdujących się najbliżej wg miary korpusowego podobieństwa znaczeniowego.

Obliczanie podobieństwa kontekstu i definicji III

- **Porównanie wektorów**
  - produkt kartezjański
  - cosinus
  - odległość euklidesowa (odwrócona)
  - współczynnik Jaccarda
  - współczynnik Pearsona (przesunięty)
  - uśredniona dywergencja Kullbacka-Leiblera (odwrócona)

Plan

Miary ewaluacji

- Zadanie polega na wybraniu dla każdego wystąpienia słowa ze słownika jednego pasującego sensu.
- Najprostsza miara – skuteczność – **ACC**.
- Obliczania jako stosunek ilości dobrych odpowiedzi do wszystkich odpowiedzi.

- Ale czy to odpowiednia miara?
- Przykład: *ciało* – 127 razy w 1. sensie, 6 razy w drugim.

Dodatkowe miary ewaluacji

- Miara **IMPROVE**:
  - Oznacza liczbę słów wieloznacznych, dla których metoda pokonała MFS.
- Miara **RARE**:
  - Promuje poprawne rozpoznawanie rzadszych sensów.
  - Waga odpowiedzi: $1 + \ln \frac{1}{\textit{częstość danego sensu}}$.
  - W przypadku *ciała*: częstszy sens ma wagę ok. 1, rzadki ok. 4.

Dane testowe

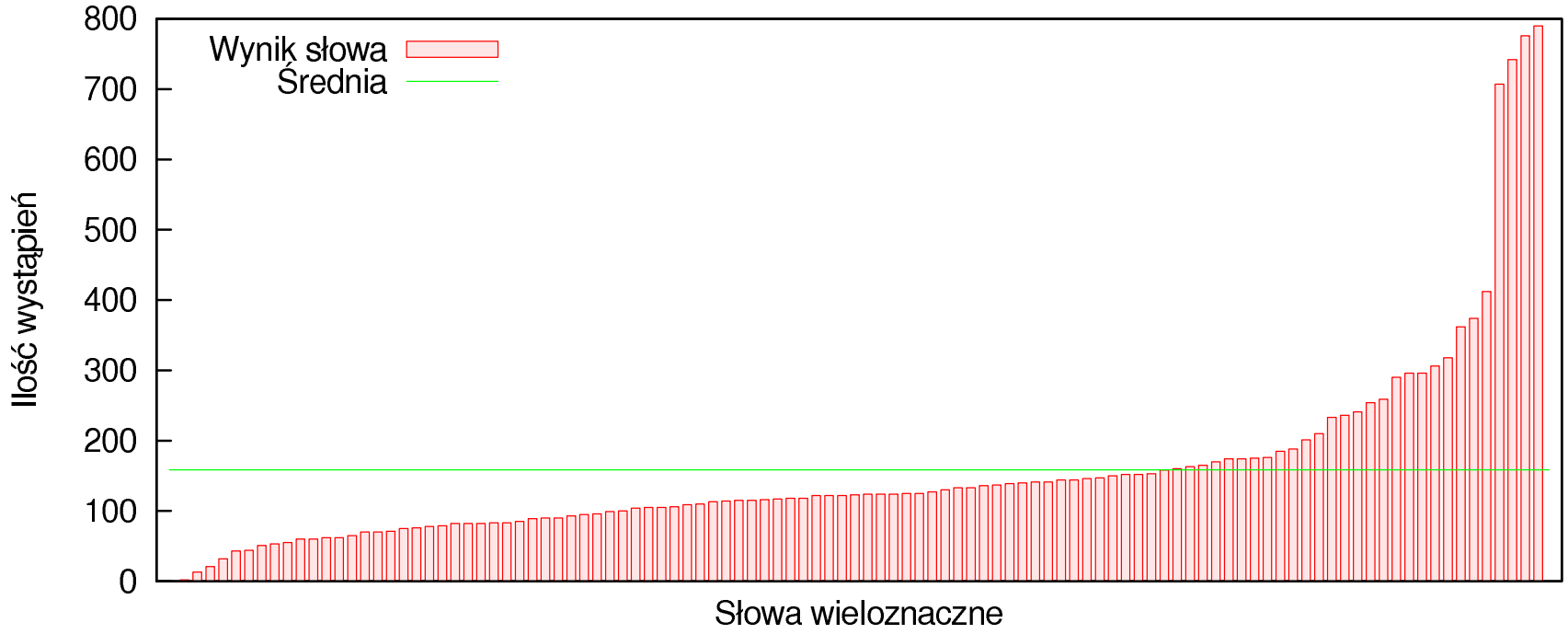
- Dane z ręcznie anotowanej części NKJP,
- tym samym zrównoważone.
- Korpus podzielony na dwie części (na poziomie tekstów): rozwojową i testową.

Statystyki użytych tekstów

	Cały korpus	Część rozwojowa	Część testowa
Liczba tekstów	3889	1944	1945
Liczba segmentów	1217822	592420	625402

Słownik sensów

- Słownik stworzony na potrzeby NKJP.
- 106 słów wieloznacznych, od 2 do 7 sensów na słowo.
- 50 rzeczowników, 48 czasowników oraz 8 przymiotników.
- Średnia ilość sensów/słowo: 2,85.

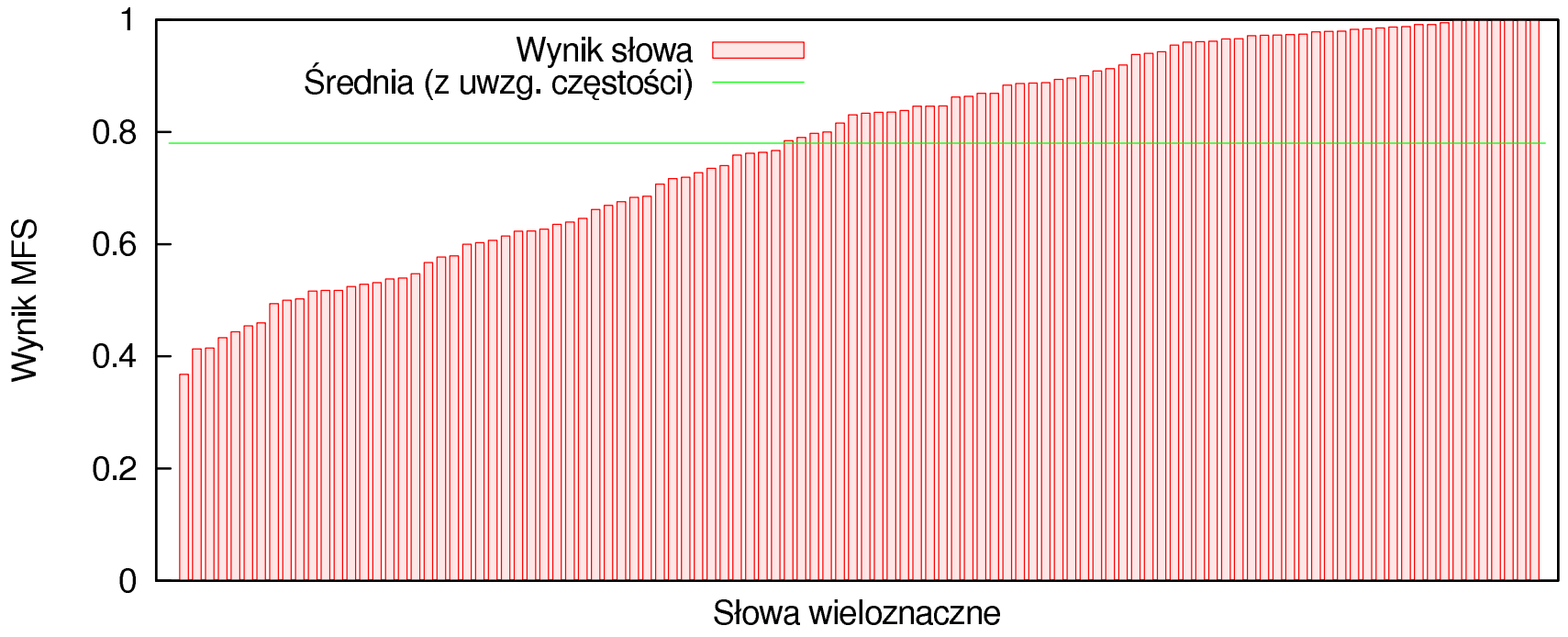


Słownik sensów

Statystyki wystąpień słów wieloznacznych w części rozwojowej

Części mowy	Liczba wyst. w słowniku	Liczba wyst. w korpusie	MFS	Random
Rzeczowniki	50	8096	0,809	0,397
Czasowniki	48	7054	0,753	0,313
Przymiotniki	8	1653	0,796	0,362
Wszystkie	106	16803	0,784	0,358

Słownik sensów



Przestrzeń poszukiwań

Parametr	Przyjmowane wartości	
	Liczba	Wartości
Rozmiar definicji	2	pełna, krótka
Rozmiar kontekstu	7	1, 2, 5, 10, 20, 30, 50
Lematyzacja	2	tak, nie
Binaryzacja	2	tak, nie
IGF	2	tak, nie
IDF	2	tak, nie
Dolne wartości skrajne	5	0,00, 0,01, 0,05, 0,10, 0,20
Górne wartości skrajne	5	0,80, 0,90, 0,95, 0,99, 1,00
Normalizacja definicji	2	tak, nie
Normalizacja kontekstu	3	brak, liniowa, kwadratowa
Rozszerzenie	7	brak, słowosieć-naiwne, słowosieć-ostrożne, synonimy.ux-naiwne, synonimy.ux-ostrożne, słowosieć-relacje, słowosieć-podobieństwo semantyczne
Porównanie wektorów	6	product, cosine, euclidean, jaccard, pearson, kullback

Przeprowadzone eksperymenty

6 zestawów:

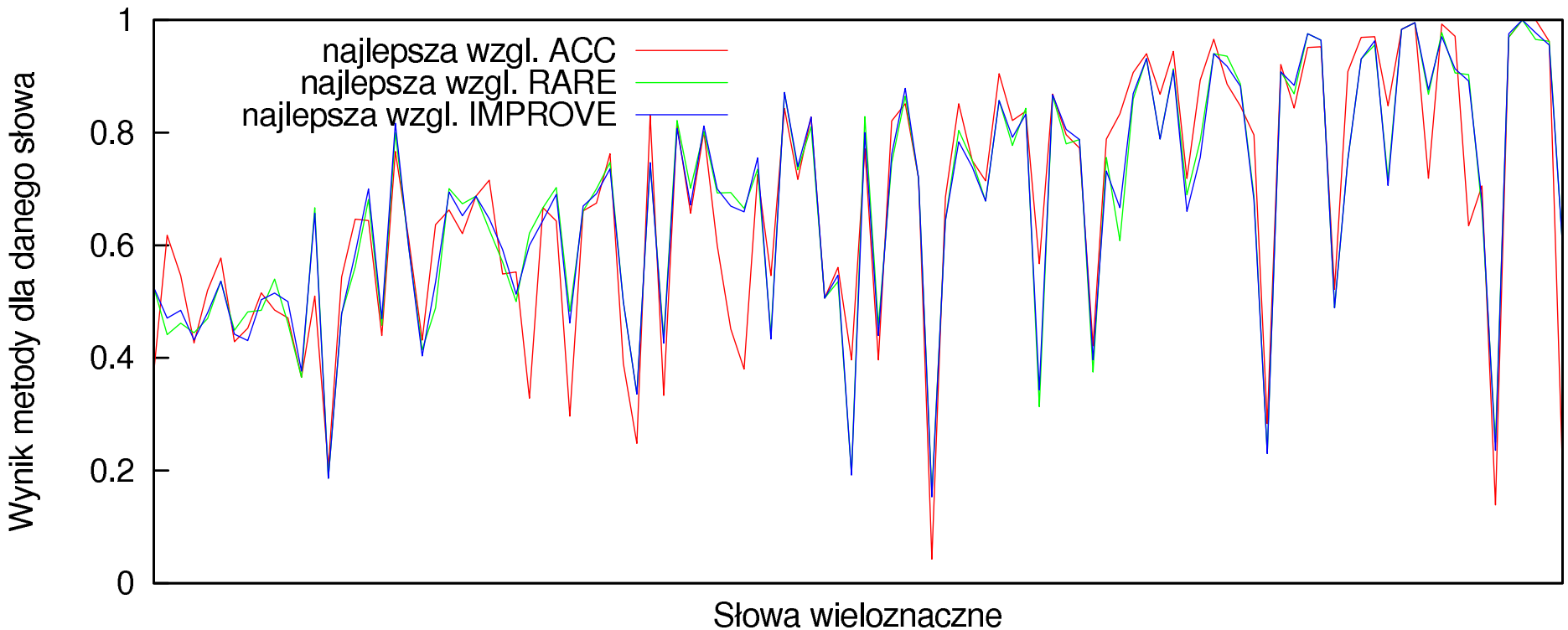
- Zestaw I - rozmiar definicji, kontekstu, lematyzacja - 112 metod,
- Zestaw II - IDF, IGF, binaryzacja - 64 metody,
- Zestaw III - normalizacja definicji, kontekstu, wielkość kontekstu - 672 metody,
- Zestaw IV - rozszerzanie kontekstu - 84 metody,
- Zestaw V - przecięcie wektorów - 96 metod,
- Zestaw VI - odrzucanie skrajnych wartości - 150 metod.

Najlepsze metody

Parametr	Najlepsza względem:		
	ACC	RARE	IMPROVE
Rozmiar definicji	pełna		
Rozmiar kontekstu	30		
Lematyzacja	tak		
Binaryzacja	nie		
IGF	tak		
IDF	tak		
Dolne wartości skrajne	0,01		
Górne wartości skrajne	1		0,99
Normalizacja definicji	nie	tak	
Normalizacja kontekstu	liniowa	kwadratowa	
Rozszerzenie	słowosieć-podobieństwo semantyczne		
Porównanie wektorów	product	jaccard	

Skuteczność wybranych trzech najlepszych metod

Miara	ACC	RARE	IMPROVE
Wynik na części rozwojowej	0,658	15307,483	29
Wynik na części testowej	0,661	15895,4	27
Wynik na całym korpusie	0,66	31253,203 (/2 = 15626,602)	31



10 najłatwiejszych słów

1. najlepsza względem ACC
2. najlepsza względem RARE
3. najlepsza względem IMPROVE

Słowo	Liczba	Rozkład	MFS	Random	1	2	3
punkt	163	60/53/27/23	0,368	0,250	0,325	**0,386**	**0,392**
stanąć	75	31/28/16	0,413	0,333	**0,533**	**0,546**	**0,546**
składać	82	34/31/9/5/2/1	0,414	0,166	**0,439**	**0,463**	**0,463**
zwrócić	90	39/37/14/0	0,433	0,250	**0,588**	**0,500**	**0,500**
strona	374	166/108/100	0,443	0,333	**0,459**	**0,483**	**0,473**
grać	110	50/30/25/5	0,454	0,250	**0,672**	**0,600**	**0,518**
góra	137	63/52/13/6/3	0,459	0,200	**0,540**	**0,518**	**0,510**
stosunek	83	41/38/4	0,493	0,333	**0,566**	**0,566**	**0,614**
brać	152	76/56/10/6/2/1/1	0,500	0,142	0,493	0,421	0,407
stać	412	207/183/19/3	0,502	0,250	0,466	0,502	**0,509**

10 najtrudniejszych słów

1. najlepsza względem ACC
2. najlepsza względem RARE
3. najlepsza względem IMPROVE

Słowo	Liczba	Rozkład	MFS	Random	1	2	3
rola	115	114/1/0	0,991	0,333	0,973	0,965	0,973
kultura	118	117/1	0,991	0,500	0,974	0,940	0,940
zasada	188	187/1	0,994	0,500	0,978	0,962	0,962
nastąpić	65	65/0	1,000	0,500	0,169	0,323	0,353
powstać	122	122/0/0	1,000	0,333	0,237	0,163	0,172
podstawa	144	144/0	1,000	0,500	0,965	0,958	0,965
piłka	43	43/0	1,000	0,500	1,000	1,000	1,000
członek	170	170/0/0	1,000	0,333	0,835	0,882	0,876
czuć	176	176/0	1,000	0,500	0,994	0,982	0,988
prowadzenie	2	2/0/0	1,000	0,333	0,500	0,500	0,500

Plan

Trudne przypadki

- Powstać – sensy:
  1. zacząć istnieć w rzeczywistości lub w świadomości (122 wystąpienia),
  2. zmienić pozycję na stojącą (0 wystąpień),
  3. sprzeciwić się lub wystąpić zbrojnie (0 wystąpień).
- Nastąpić – sensy:
  1. wejść na coś stopą, atakować (65 wystąpień),
  2. zdarzyć się (0 wystąpień).

Trudne przypadki cd.

- Udać – sensy:
  1. zakończyć się powodzeniem (150 wystąpień),
  2. pójść, pojechać (21 wystąpień),
  3. udawać: sprawiać wrażenie, że coś robimy (3 wystąpienia).
- Zostać – sensy:
  1. ulec (do tworzenia strony biernej) (633 wystąpienia),
  2. stać się kimś (64 wystąpienia),
  3. być/trwać w jakimś stanie/miejscu, nie zmienić się (51 wystąpień),
  4. ostać się z całości (28 wystąpień).

Trudne przypadki cd.

Metoda	Rzeczowniki	Czasowniki	Przymiotniki
Najlepsza względem ACC	0,726	0,574	0,676

Porównanie z metodami z nadzorem

- W danych testowych MFS – niemal 80%.
- Zatem trywialny system z nadzorem osiąga taką skuteczność!
- Specyfika: oddzielna metoda dla każdego słowa.
- Skuteczność: 91,46% na podkorpusie NKJP, z walidacją krzyżową.

Porównanie z metodami bez nadzoru

- Brak prób rozróżniania sensu polskich słów metodami bez nadzoru.
- Semeval 2007:
  - *All-words* dla języka angielskiego.
  - „Gruboziarniste sensy” poprzez pogrupowanie sensów *Wordnetu*.
  - 5377 wystąpień, 30% wystąpień słów monosemicznych.
  - 3,06 sensu/słowo (w *Wordnecie: 6,18*), 78,89% – MFS, 52,43% – Random.
  - Najlepszy system oparty na wiedzy: ok. **60%** F1.
- Semeval 2010:
  - *All-words* dziedzinowe dla języka angielskiego.
  - Drobniejszy podział na sensy: 23,2% – Random.
  - Najlepszy system oparty na wiedzy: ok. **50%** F1.

Możliwe rozszerzenia systemu

- Metoda jest niestabilna – wymaga analizy trudnych przypadków.
- Pożądane było zagwarantowanie skuteczności prawie zawsze powyżej heurystyki **Random**.
- Rozszerzenie metody na klasyfikację w sposób zależny – dla wielu słów wieloznacznych w kontekście.
- Wymaga to jednak stworzenia kosztownych zasobów: słownika oraz korpusu.

To już wszystko

Dziękuję za uwagę!