Uczenie nienadzorowane w wydobywaniu znaczeń leksykalnych słów

Bartosz Broda

Politechnika Wrocławska
Samsung Research Institute Poland

20 maja 2013

Politechnika
Wrocławska

Plan prezentacji

Przykład wieloznaczności

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów po jakimś terenie
2. książka podająca wiadomości z historii, geografii danego regionu
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik**-2 po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. książka podająca wiadomości z historii, geografii danego regionu
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik**-2 po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. książka podająca wiadomości z historii, geografii danego regionu
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik**-2 po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. książka podająca wiadomości z historii, geografii danego regionu
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik**-2 po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. książka podająca wiadomości z historii, geografii danego regionu
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik-2** po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. **książka podająca wiadomości z historii, geografii danego regionu**
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Przykład niejednoznaczności

Aleksander – nasz przewodnik-1 po Petersburgu, profesor, humanista, erudyta... fanatyk zbiorów...

To może **przewodnik-2** po najpiękniejszych polskich miastach?

Izolowane od siebie chłodzone segmenty zachowują się jak zły przewodnik-3 elektryczny, a dobry przewodnik-3 cieplny.

Przewodnik Słownik Języka Polskiego PWN

1. osoba, która zawodowo oprowadza turystów
2. **książka podająca wiadomości z historii, geografii danego regionu**
3. ciało dobrze przewodzące prąd elektryczny w warunkach normalnych

Pojęcia podstawowe

Rozstrzyganie niejednoznaczności leksykalnej słów

- ustalenie znaczenia każdego słowa w kontekście
- ustalenie, które znaczenie słowa zostało użyte w danym kontekście – problem klasyfikacji
- ang. *Word Sense Disambiguation*

Rozróżnianie znaczeń słów

- ustalenie jakie znaczenia występują w tekście
- podział wystąpień słów niejednoznacznych na rozłączne znaczeniowo grupy – problem grupowania
- ang. *Word Sense Discrimination*

E. Agirre and P. Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2006
R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009

Zastosowania

- Maszynowe tłumaczenie
  - Wybór odpowiedniego tłumaczenia
  - Rozwiązanie – denouement, solution, outcome, delivery
- Wyszukiwanie informacji
  - Którego „rozwiązania” użytkownik szuka?
- Ekstrakcja informacji
  - W wypadku zbierania informacji o spółkach giełdowych – interesujące są rozwiązania spółek a nie np. porody czy rozwiązania łamigłówek
- Badania językowe
  - Wspomaganie tworzenia słowników

Skala problemu

	Monosemiczne	Polisemiczne		Stopień polisemii	
	słowa i znaczenia	słowa	znaczenia	wszystkie	wieloznaczne
WordNet					
Rzeczowniki	101 863	15 935	44 449	1,24	2,79
Czasowniki	6 277	5 252	18 770	2,17	3,57
Przymiotniki	16 503	4 976	14 399	1,40	2,71
Słowosieć 1.6					
Rzeczowniki	50 298	16 851	46 166	1,44	2,74
Czasowniki	10 149	7 167	21 440	1,82	2,99
Przymiotniki	2 634	912	2 384	1.82	2,61

Tabela: Stopień wieloznaczności dla języka angielskiego i polskiego na przykładzie elektronicznych tezaurusów

- Korpus Instytutu Podstaw Informatyki PAN — ponad 250 milionów słów (rozszerzony do 1,8 mld)
- Narodowego Korpusu Języka Polskiego (1 mld słów)

Typowe podejścia

- Oparte o wiedzę
  - ręcznie napisane reguły
  - porównanie definicji słownikowej z kontekstem
  - heurystyki
- Metody wykorzystujące uczenie nadzorowane
  - indukcja drzew decyzyjnych
  - uczenie pamięciowe
- W oparciu o przesłanki wielojęzyczne (ang. *translational equivalence*) — wykorzystanie zasobów wielojęzycznych
- Metody wykorzystujące uczenie nienadzorowane — wykorzystanie dużych korpusów nieoznaczonych ręcznie

Typowe podejścia

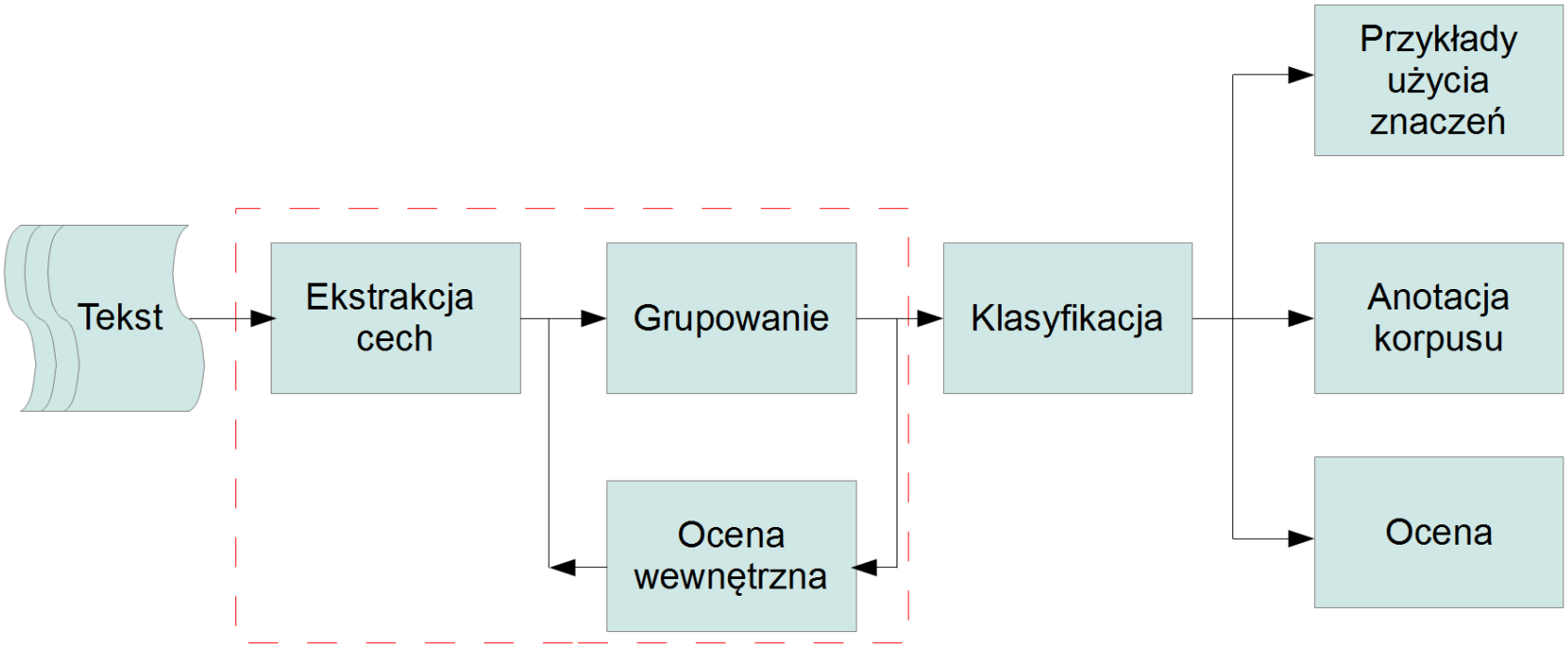
- Oparte o wiedzę
  - ręcznie napisane reguły
  - porównanie definicji słownikowej z kontekstem
  - heurystyki
- Metody wykorzystujące uczenie nadzorowane
  - indukcja drzew decyzyjnych
  - uczenie pamięciowe
- W oparciu o przesłanki wielojęzyczne (ang. *translational equivalence*) — wykorzystanie zasobów wielojęzycznych
- Metody wykorzystujące uczenie nienadzorowane — wykorzystanie dużych korpusów nieoznaczonych ręcznie

Inspiracja dla zaproponowanej metody

- Podejście zostało oparte na sposobie pracy leksykografów[1]:
  - Zbieranie przesłanek z korpusów tekstów w postaci przykładów użycia danego słowa
  - Intuicyjny podział zebranych danych na grupy
  - Analiza grup w poszukiwaniu wspólnych właściwości przykładów użycia słowa w grupach
  - Sformułowanie definicji dla znaczeń
- Lexicographer-Controlled Semi-automatic word sense Disambiguation (**LexCSD**)[2]:
  - Zbieranie przykładów użycia
  - Grupowanie
  - Opcjonalne opisywanie grup
  - Trenowanie klasyfikatorów
  - Przygotowanie reprezentatywnych przykładów użycia

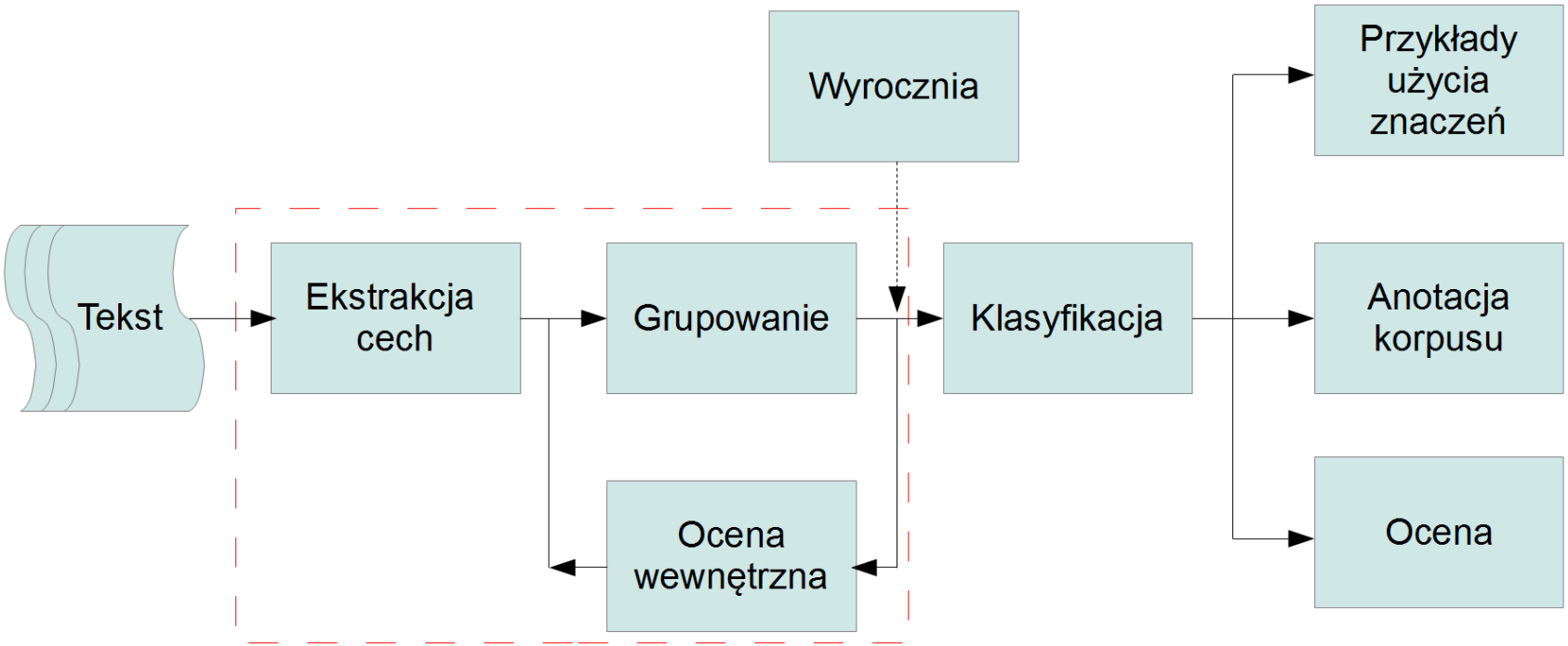
1. A. Kilgarriff. The hard parts of lexicography. *International Journal of Lexicography*, 11:51–54, 1997
2. B. Broda and M. Piasecki. Evaluating LexCSD in a Large Scale Experiment. *Control and Cybernetics*, 40(2), 2011

Schemat autorskiej metody LexCSD



B. Broda and M. Piasecki. Evaluating LexCSD in a Large Scale Experiment. *Control and Cybernetics*, 40(2), 2011

Schemat autorskiej metody LexCSD



B. Broda and M. Piasecki. Evaluating LexCSD in a Large Scale Experiment. *Control and Cybernetics*, 40(2), 2011

Korpus ręcznie oznaczony znaczeniami

- Korpusik US II PWr
  `http://nlp.pwr.wroc.pl/webann/browser/`
- Pierwszy dostępny korpus języka polskiego oznaczony znaczeniami leksykalnymi
- Oparty na Korpusie IPI PAN
- 13 słów o różnym stopniu polisemii, znaczenia drobnoziarniste ze Słowosieci
- Różnorodność tematyczna tekstów
- 2 osoby anotujące cały tekst
- Wysoka zgodność anotatorów, $\kappa = 0.88$
- 1344 przykłady użycia słów

B. Broda, M. Maziarz, and M. Piasecki. Evaluating lexcsd — a weakly-supervised method on improved semantically annotated corpus in a large scale experiment. In S. T. Wierzchoń M. A. Kłopotek, A. Przepiórkowski and K. Trojanowski, editors, *Proceedings of Intelligent Information Systems*, 2010

Statystyki korpusu

Słowo	Znaczenia	Anotacje	Przykłady	κ
agent	6	1/9/3/47/10	70	0.80
automat	7	1/24/30/4/46	105	0.97
dziób	6	28/13/31/9	81	0.98
język	7	3/23/49	75	0.97
klasa	14	15/6/12/11/14/31/10/8/1/10/1	119	0.80
linia	14	13/3/2/2/4/2/11/13/4/3/1/2/21	81	0.72
pole	11	1/1/23/25/46	96	0.86
policja	3	17/25/22	64	0.73
powód	3	136/122	258	0.98
sztuka	10	12/10/2/11/41/19	95	0.84
zamek	6	18/19/36/19	92	1.00
zbiór	7	32/7/8/31/9	87	0.87
zespół	7	10/4/28/58/1/20	121	0.95

Sposób oceny

- Określenie punktów odniesienia
  - Dolna granica: heurystyka wyboru najczęstszego znaczenia, ang. *Most Frequent Sense, MFS*
  - Górna granica: wynik metod nadzorowanych
- Metody nadzorowane testowane z wykorzystaniem walidacji krzyżowej
- Ocena metod przy pomocy dokładności i pokrycia:
  - Dokładność $D_i = \frac{h_i}{h_i+m_i}$, gdzie h_i to liczba poprawnie podjętych decyzji dla $i-tego$ znaczenia, a m_i to liczba niepoprawnych decyzji dla $i-tego$ znaczenia
  - Pokrycie: $P_i = \frac{h_i+m_i}{h_i+m_i+s_i}$, gdzie s_i oznacza liczbę niepodjętych decyzji przez metodę
  - Miara $F_1 = \frac{2D \times P}{D+P}$

Badania wielkoskalowe metody LexCSD

- Cel eksperymentu: sprawdzenie jakości działania metody uczonej na dużych zbiorach tekstów
- Proste cechy leksykalne, podejście worka słów
- LexCSD uczone na dużych zbiorach tekstów (ok. 570 milionów tokenów)
  - Korpus IPI PAN
  - Korpus Rzeczpospolitej
  - Korpus Internetowy
- Test LexCSD z wykorzystaniem Korpusika US II PWr
- Badane metody klasyfikacji: **tablice decyzyjne (DT)**, **maszyny wektorów wspierających (SVM)**, losowy las drzew (RF), AdaBoost (AB), metoda najbliższego sąsiada (1-NN), drzewa decyzyjne C4.5, naiwny klasyfikator bayesowski (NB)

Wyniki eksperymentu

Słowo	MFS	DT	SVM	LexCSD DT		LexCSD SVM	
	D[%]	D[%]	D[%]	D[%]	P[%]	D[%]	P[%]
agent	67,14	67,14	71,43	65,71	100	55,17	82,86
automat	43,81	71,43	79,05	**57,58**	94,27	**60,67**	84,76
dziób	38,27	59,26	88,89	46,91	100	51,43	86,42
język	65,33	77,33	77,33	**67,57**	98,67	**73.91**	61,33
klasa	26,05	42,86	63,87	50,79	52,94	39,8	82,35
linia	25,93	37,04	45,68	29,63	66,67	16,22	88,89
pole	47,92	75	68,75	64,44	46,88	69,44	75
policja	39,06	40,62	48,44	0	3,13	35	31,25
powód	52,71	88,37	89,53	48,58	95,74	38,1	16,28
sztuka	43,16	46,32	54,74	58,82	17,89	48,75	84,29
zamek	39,13	57,61	70,65	36,47	92,74	52,73	59,78
zbiór	36,78	57,47	74,71	70,73	47,13	76,56	73,56
zespół	47,93	65,29	77,69	0	98,35	5,71	58,85
w. avg.	**44,57**	64,06	**72,92**	45,94	74,18	**48,09**	62,2

B. Broda and M. Piasecki. Evaluating LexCSD in a Large Scale Experiment. *Control and Cybernetics*, 40(2), 2011

Podsumowanie wyników eksperymentu

Tabela: Miara F_1 dla średnich ważonych z dokładności i pokrycia LexCSD

	DT	**SVM**	**RF**	**AB**	**1-NN**	**C4.5**	**NB**
F_1	56,74	54,24	57,92	44,70	47,30	51,23	48,87

- Dokładność niższa od metod nadzorowanych, jednak wyższa od punktu odniesienia
- Pięć słów z którymi metoda miała największe problemy
  - *język*, *policja*, *powód* – jedno dominujące znaczenia w korpusie
  - *sztuka* – ograniczony zestaw cech na etapie klasyfikacji
  - *zespół* – zupełnie inny rozkład znaczeń w dużym korpusie

Opis problemu

- O jakości grupowania decyduje w dużym stopniu liczba grup
  - Zbyt mało grup – łączenie znaczeń
  - Zbyt dużo grup – podział pojedynczych znaczeń na podstawie subtelnych różnic w rozkładzie cech
- Problem jest istotny w wielu zadaniach
  - Word Sense Induction (SemEval)
  - Wspomagana komputerowo leksykografia (np. rozbudowa Słowosieci)
- Liczba znaczeń powinna być dostosowana pod kątem dostępnych korpusów tekstów

Istniejące podejścia 1/2

- Statystyczne prawa językowe
- Na przykład: prawo Zipf'a, prawo Menzeratha-Altmana
- Wykorzystują proste przesłanki takie jak częstość słowa czy jego długość
- Bardzo dobrze opisują dane
- Słaba moc predykcyjna

Istniejące podejścia 2/2

- Rodzina metod grupowania, które są w stanie określić liczbę grup
  - Na przykład: Affinity Propagation
  - Estymacja liczby grup jest ściśle związana z daną metodą grupowania
  - Zazwyczaj należy zoptymalizować inne parametry samej metody
- Funkcje (reguły) zatrzymania
  - Dane są grupowane na $1 \ldots k$ grup
  - Reguła zatrzymania ocenia każdy podział
  - Podział danych odpowiadający optymalnej wartości reguły zatrzymania odpowiada optymalnej liczbie grup
  - Mogą być zastosowane dla dowolnej metody grupowania
  - Charakteryzują się dużym kosztem obliczeń

Reguły zatrzymania 1/2

- Adapted Gap Statistic (AGS) – porównywanie zmian w zwartości grup z wartością oczekiwaną pod warunkiem referencyjnego rozkładu danych
  1. Wygeneruj dane używając referencyjnego rozkładu prawdopodobieństwa
  2. Oblicz rozproszenie wewnątrz grup
  3. Oblicz różnicę pomiędzy modelem danych rzeczywistych a danych wygenerowanych
  4. Wybirz liczbę grup odpowiadającą najmniejszej różnicy
- $PK1(k) = \frac{crfun(k) - mean(crfun[1...maxK])}{std(crfun[1...maxK])}$
  - Optymalna liczba grup – gdy PK1 przekracza wskazany próg
  - Konieczność optymalizacji progu

Reguły zatrzymania 1/2

- $PK2(k) = \frac{crfun(k)}{crfun(k-1)}$
  - Porównanie dwóch następujących po sobie podziałów na grupy
  - Wartość optymalna dla PK2 najbliższego 1
- $PK3(k) = \frac{2*crfun(k)}{crfun(k-1)+crfun(k+1)}$
  - Podobieństwo do współczynnika Dice
  - Optymalna liczba grupy, gdy $PK3$ jest większe od odchylenia standardowego
- $CH = \frac{BGSS}{(k-1)} / \frac{WGSS}{n-k}$
  - $BGSS$ – suma kwadratów odległości pomiędzy grupami
  - $WGSS$ – suma kwadratów odległości wewnątrz grup
  - n – wielkość zbioru danych

Sposób oceny

- Problemy z definicją zadania
  - Nie wszystkie znaczenia słowa występują w danym korpusie tekstów
  - Rozkład znaczeń jest obciążony
- Dokładne trafienia nie są zawsze konieczne
- Nie wszystkie pomyłki są jednakowo groźne
- Akceptowalne odpowiedzi
  - Mniej niż 4 znaczenia – $+1$ znaczenie
  - Mniej niż 10 znaczeń – ± 1 znaczenie
  - Więcej niż 10 znaczeń – ± 2 znaczenia

Korpusy

- Korpusik US II PWr
- Korpusik-coarse – Korpusik ze znaczeniami gruboziarnistymi
- LC13 i LC13-coarse
  - Te same słowa co w Korpusiku
  - 500M słów: Korpus IPI PAN, Rzeczpospolita, korpus internetowy
- IPIC-33 – 33 niejednoznaczne słowa, dane z Korpusu IPI PAN
- Zbiory z konkurencji SensEval (Lexical Sample Task)

Eksperymenty

- Cechy: worek słów
- Grupowanie podziałowe z funkcjami celu: i1 i e1
- Redukcja wymiarowości (SVD)
  - do 30% dla słów opisanych przez zbiór cech o liczności większej niż 30000
  - do 70% dla słów opisanych przez zbiór cech o liczności większej niż 1000
- Wszystkie eksperymenty zostały powtórzone 10 raz
- Znaczenia ze Słowosieci

Wyniki dla języka polskiego

Dataset	**PK1 [%]**	**PK2 [%]**	**PK3 [%]**	**CH [%]**	**AGS [%]**
Korpusik	**69.23**	46.15	7.69	7.69	7.69
Korpusik-coarse	46.15	**76.92**	38.46	30.77	23.08
LC-13	**15.38**	**15.38**	0.00	0.00	0.00
LC-13-coarse	53.85	**76.92**	30.77	15.38	38.46
IPIC-33	6.06	**54.54**	15.16	30.30	15.16

Tabela: Precyzja reguł zatrzymania dla języka polskiego

Wyniki dla języka angielskiego

Dataset	**PK1 [%]**	**PK2 [%]**	**PK3 [%]**	**CH [%]**	**AGS [%]**
SensEval-1	11.11	**22.22**	14.81	11.11	16.67
SensEval-2	4.10	**31.50**	1.37	2.73	9.59
SensEval-3	1.75	**17.54**	1.75	0.00	1.75
SensEval-4	3.00	**62.00**	34.00	12.00	29.00

Tabela: Precyzja reguł zatrzymania dla języka angielskiego

Dyskusja

- Znaczenia drobnoziarniste: nawet zawodowi leksykografowie nie są zgodni
- Dodatkowe eksperymenty z innymi zbiorami znaczeń
  - Słownik Języka Polskiego PWN
  - Wyższa precyzja: np. wzrost z 54.54% do 63.63% dla PK2 na IPIC-33
- Wyniki dla języka angielskiego są niskie, ale spójne z wynikami znanymi z literatury i wynikami dla języka polskiego
- CH i AGS wykorzystują odległość euklidesową

Opis problemu

- Wiele metod grupowania
- Różny rozkład znaczeń dla różnych słów
- Wstępne badania pokazały, że wykorzystując jedną metodę grupowania można osiągnąć najlepsze wyniki tylko dla 64% słów
- Dwa główne sposoby oceny jakości metod grupowania:
  - wykorzystujące kryteria wewnętrzne
  - wykorzystujące kryteria zewnętrzne

Badane metody grupowania

- k-średnich
- k-medoidów
- grupowanie hierarchiczne aglomeracyjne
- grupowanie hierarchiczne podziałowe
- grupowanie oparte na podziale grafu metodą min-cut
- 7 funkcji celu dla metod hierarchicznych

Badane metody oceny jakości grupowania

- Indeks Dunna – $\mathcal{DI} = \frac{d_{\min}}{d_{\max}}$, gdzie $d_{\min}$ oznacza minimalną odległość pomiędzy 2 obiektami nienależącymi do tej samej grupy, a $d_{\max}$ maksymalną odległość pomiędzy obiektami w tej samej grupie.
- Miara oczekiwanej gęstości (ang. *Expected Density Measure, EDM*) – porównywanie gęstości grafów zbudowanych w obrębach grup do grafu zbudowanego na całych danych.
- Indeks C – $CI = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$, gdzie S jest sumą odległości pomiędzy obiektami należącymi do tej samej grupy, $S_{\min}$ ($S_{\max}$) jest sumą I najmniejszych (największych) odległości pomiędzy dwoma dowolnymi obiektami.
- Funkcje celu

$$I_1 = \text{maximize} \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{\vec{v},\vec{u} \in S_i} sim(\vec{v}, \vec{u}) \right) \quad (1)$$

$$I_2 = \text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{\vec{v},\vec{u} \in S_i} sim(\vec{v}, \vec{u})} \quad (2)$$

Zbiory danych

- Korpusik US II PWr
- GPW – 28 słów z dziedziny ekonomii w 14106 tekstach
- NE – podzbiór NKJP (46 słów w 3732 tekstach)
- Zbiory z konkurencji SensEval

Porównanie z BCubed

Corpus	**CI**	**DI**	**DI_2**	**EDM**	**I_1**	**I_2**
GPW	0.797	0.808	0.685	0.688	0.634	0.578
NE	0.742	0.762	0.558	0.698	0.684	0.552
Korpusik	0.553	0.557	0.480	0.566	0.533	0.482
S-1	0.625	0.631	0.385	0.616	0.521	0.385
S-2	0.544	0.559	0.344	0.528	0.397	0.344
S-3	0.582	0.591	0.355	0.575	0.510	0.356
S-4	0.730	0.740	0.502	0.703	0.691	0.499
Average	**0.653**	**0.664**	**0.473**	**0.625**	**0.567**	**0.457**

Korelacja z BCubed

Corpus	**CI**	**DI**	**DI_2**	**EDM**	**I_1**	**I_2**
GPW	0.632	0.512	0.235	-0.026	-0.255	-0.251
NE	0.758	0.386	-0.346	0.378	-0.437	-0.261
SCWSD	0.672	0.326	-0.346	0.549	0.070	0.085
S-1	0.625	0.599	0.068	0.692	-0.699	-0.667
S-2	0.631	0.456	-0.356	0.516	-0.592	-0.741
S-3	0.722	0.382	-0.529	0.627	-0.17	-0.269
S-4	0.762	0.417	-0.23	0.616	-0.372	-0.311
Average	**0.686**	**0.440**	**-0.215**	**0.479**	**-0.350**	**-0.345**

Ocena ręczna

- Badania dla całego Korpusu IPI PAN
- Wybrano 5 słów do ręcznej oceny: agent, pociąg, rakieta, zamek, zbiór
- Znaczenia określone na podstawie Słowosieci
- Indek C jako metoda wyboru algorytmu grupowania
- Wnioski:
  - Nie znaleziono wszystkich znaczeń
  - Większość znaczeń została wykryta poprawnie
  - Dla *pociąg* tylko jedno znaczenie
  - Dla *zamek* wykryto wszystkie gruboziarniste znaczenia

Dyskusja

- Metody oceny wewnętrznej grupowania są pomocne w nienadzorowanym wyborze algorytmu grupowania dla słowa
- Metody te opierają się na danych korpusowych
- Największa korelacja z BCubed została osiągnięta dla Indeksu C
- Zaproponowany protokół badawczy nie jest ograniczony do problemu ujednoznaczniania znaczeń leksykalnych słów

Wydobywanie przykładów użycia znaczeń

- Dobry przykład użycia powinien być[1]
  - typowy
  - informatywny
  - jednoznaczny
- W LexCSD[2] podejście oparte na centroidach grup wspomaganych dodatkowym filtrowaniem heurystycznym, np.
  - wymóg zawierania pewnej liczby rzeczowników
  - heurystyczne usuwanie fragmentów i podpisów tabel
  - usuwanie przykładów użycia opisujących nazwy własne

1. A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlỳ. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX*, pages 425–32, 2008
2. B. Broda, M. Maziarz, and Piasecki. Tools for plWordNet development. presentation and perspectives. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC'12)*, 2012

Automatycznie wydobyte przykłady użycia znaczeń

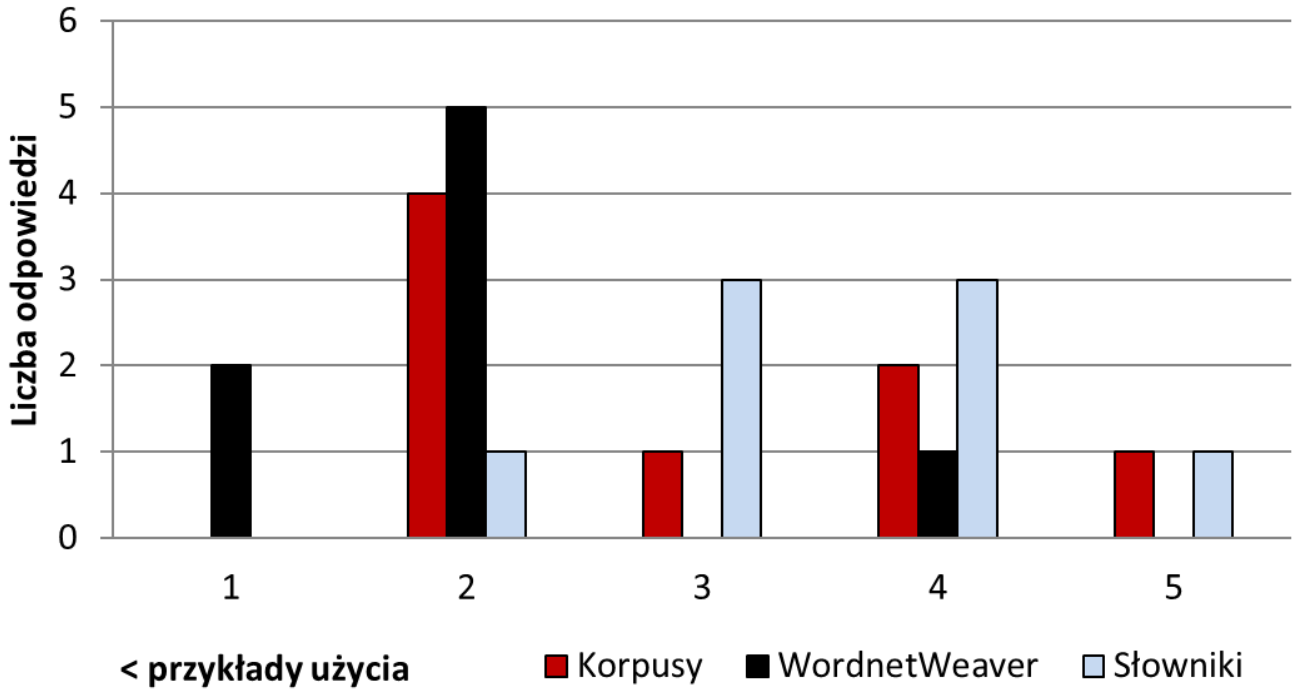
Kleić

- W domach jednorodzinnych drzwi wewnętrzne najczęściej wykonane są z drewna litego lub **klejonego**.
- Oczy mu się **kleiły**; zaczął szukać po omacku butelki.
- Rozmowa się zbytnio nie **kleiła**. Nic dziwnego.
- Z meczu na mecz gramy coraz lepiej. Gra się nam **klei**, zgrywamy się, stwarzamy dużo sytuacji bramkowych.

WordnetLoom 1.48 (Centralna - Robocza) NEKST
Plik Edycja Perspektywa Inne Okno Pomoc
Relacje jednostek Relacje synsetów Edycja synsetów Wizualizacja w...
Jednostki Kandydaci
Część mowy:
czasownik
Numer paczki:
116 (491)
Przelicz
Wyszukaj:
Szukaj
Kandydaci:
kołatać się
najść
nasunąć się
przeminąć
rozwiać się
sczeznąć
skojarzyć się
spełznąć
świtać
taić się
ulecieć
ulotnić
ulotnić się
wywietrzeć
zagościć
zaświtać
złagodnieć
Liczba elementów: 17
najść #116
najść
kandydat
kandydat
kandydat
kandydat
wypełniać 3
obudzić się 1*
zwiać 2* (ruch)
bzn
pierzchnąć 1*
hipo
hiper
uciec 1 (ruch) ...
bzn
hiper
hipo
umknąć 1*
Podgląd
najść
Synset
Jednostki w synsecie:
uciec 1 (ruch)
ujść 1* (ruch)
wziąć nogi za pas 1* (ruch)
zbiec 2* (ruch)
Komentarz:
brak danych
Właściciel: Aleksandra Pawlikowska
Relacje jednostki
uciec 1 (ruch)
Od
aspektowość czysta DK-NDK
uciekać 1 (ruch)
Do
aspektowość czysta NDK-DK
uciekać 1 (ruch)
Przykłady Schowek
by z nią porozmawiać i zwierzyć się ze swych szalonych myśli. Gdy tylko zaczął mówić, rzekła: „Wiem, co chcesz powiedzieć. Już dawno temu też naszły mnie takie myśli, ale nie chciałam się nimi z tobą dzielić, bo wiem, że nie wierzysz w miłość. Miłość chyba istnieje, tylko jest czym
mniej nie codziennie człowiek dostarcza sobie kulturalnej stymulacji więc dziś będzie o czym innym. mam nadzieje że nie stracę przez to moich regularnych czytelników o ile tacy istnieją. Naszła mnie dziś myśl że w sieci funkcjonuje pod kilkoma nickami - co więcej nie mają one wiele wspólnego z tym kim jestem. W jednym portalu społecznościowym mam nick po
młodego. Ku temu rzekł pan Bojanowski: Wątpię ja, aby kto nalazł sie taki. Dołożył pan Dersniak: Mężczyzna trudno, a białagłowa żadną miarą sie nie najdzie. Na to pan Myszkowski powiedział: Wdy w. m., panie Dersniaku, pana Bojanowskiego wydać niechcesz, abyś niemiał białymgłowam podnoskow dawać, a one
tego schematu? — Jest to po prostu pierwsza rzecz, która przyszła mi na myśl — powiedział Spence. — Być może na przyjęciu był ktoś, kogo nagle naszła przemożna chęć morderstwa. Być może zrobił już kiedyś coś takiego, a może tylko miał na to ochotę. Chcę przez to powiedzieć, że być może w przeszłości
Dobrze - powiedział szybko. - Zgadzam się, że Carewe'a trzeba uciszyć, ale jej nie. Zabierzmy... zabierzmy ją ze sobą. - Co cię naszło, Manny? - Przecież ona jest w ciąży - wykrztusił Pleeth. - I co z tego? - spytał Barenboim, lekko marszcząc czoło. - Nie jesteś
. - Właśnie. Dlatego zlekceważyłem go po przebudzeniu. Wiedziałem, że przez krótki czas posługiwałem się Mocą, ale pomyślałem, że po prostu naszedł mnie ten sen i że wszystko potem mi się przyśniło. No więc w tym śnie, w ten dziwny sposób właściwy snom, nagle zjawił się obok mnie Młotek
jak smagnięcie bicza i coś przeorało powietrze tuż przy jego prawym uchu. Pierwszy raz do mnie strzelają - pomyślał. - Dziewięć lat w policji, dopóki mnie nie naszli i nie zniszczyli wszystkiego (czterech krawężników, czterech w cywilu i jeden z wywiadu), i nigdy do mnie nikt nie strzelał, dopiero dziś. Znowu eksplozja
O takiej długości, podgrzej je w ciepłej wodzie... Nimber jęknął tylko raz, gdy Lawler rozpłaszczył mu ramiona tak, aby obojczyk wygiął się i kość naszła z po-wrotem na miejsce. Potem zamknął oczy i wydawał się me-dytować, podczas gdy Lawler robił, co mógł, aby zmniejszyć opuchlinę i unieruchomić mu ramię, zabezpieczając
kinie, dzieliły pojedyncze oparcia. Te po moich bokach były zajęte przez ręce kolegów. Obaj trzymali dłonie dziwnie wykrzywione ku górze - gotowe na przyjęcie mojej, gdyby naszła mnie taka
Cache włączony, zapytania: (47%), czas życia: 2 min., zapytań: 14, wielkość: 0 tyś. | Ogólne zużycie pamięci: 71 MB

Ankiety

Rysunek: „W pracy nad Słowosiecią korzystam raczej z...” 1 – przykładów użycia, 5 – innych źródeł danych.



B. Broda, M. Maziarz, and Piasecki. Tools for plWordNet development. presentation and perspectives. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC'12)*, 2012

Podsumowanie

Tytuł

Uczenie nienadzorowane w wydobywaniu znaczeń leksykalnych słów

- Charakterystyka dziedziny
- Opis metody LexCSD
- Opis problemu szacowania liczby znaczeń
- Wybór algorytmu grupowania dla słowa
- Wybór przykładów użycia znaczeń leksykalnych

Dziękuję za uwagę!

Przewodnik, Słownik języka polskiego PWN

1. «osoba, która idąc przodem, wskazuje komuś drogę»
2. «osoba, która zawodowo oprowadza turystów po jakimś terenie»
3. «osoba stojąca na czele jakiejś grupy, wpływająca na kogoś»
4. «książka podająca wiadomości z historii, geografii danego regionu»
5. «książka zawierająca praktyczne wiadomości z jakiejś dziedziny»
6. «ciało, substancja zdolne do przewodzenia ciepła lub prądu elektrycznego; też: ciało dobrze przewodzące prąd elektryczny w warunkach normalnych»
7. «osobnik z grupy tego samego gatunku ptaków lub zwierząt idący, lecący na czele stada, klucza»

• przewodniczy, przewodnicki, przewodnikowy • przewodniczka