Analiza danych tekstowych

Łączenie informacji lingwistycznych i danych ilościowych

Agnieszka Mykowiecka

Instytut Podstaw Informatyki PAN
Polsko-Japońska Akademia Technik Komputerowych
http://zil.ipipan.waw.pl/AgnieszkaMykowiecka
agn@ipipan.waw.pl

Czy komputer/program komputerowy rozumie tekst?

Chianti to położona w sercu Warszawy trattoria w toskańskim stylu, przepełniona zapachem ziół i oliwy.

- *Chianti is located in the heart of Warsaw, Tuscan-style trattoria, full of the smell of herbs and olive oil.* (Google Translator)
- *Chianti is situated in the heart of the Tuscan-style trattoria in Warsaw full of the scent of herbs and olive oil.* (Bing Translator)

może ...

Czy komputer/program komputerowy rozumie tekst?

Chianti to położona w sercu Warszawy trattoria w toskańskim stylu, przepełniona zapachem ziół i oliwy.

- *Chianti is located in the heart of Warsaw, Tuscan-style trattoria, full of the smell of herbs and olive oil.* (Google Translator)
- *Chianti is situated in the heart of the Tuscan-style trattoria in Warsaw full of the scent of herbs and olive oil.* (Bing Translator)

może ...

Czy komputer/program komputerowy rozumie tekst?

The annual Oscar nominee luncheon brings together contenders from all categories ahead of the awards in just under three weeks' time.

- *Roczny nominowany do Oscara lunch skupia pretendentów ze wszystkich kategorii, wyprzedzając nagród w czasie niespełna trzy tygodnie.* (Google Translator, 2015)
- *Roczna nominowany do Oscara obiad łączy z pretendentów wszystkie kategorie przed nagród w czasie niespełna trzy tygodnie.* (Google Translator, 2017)
- *Roczne Oscar nominacja obiadowe łączy rywali z wszystkich kategorii przed nagród w niecałe trzy tygodnie.* (Bing Translator)

... na pewno nie

Czy komputer/program komputerowy rozumie tekst?

The annual Oscar nominee luncheon brings together contenders from all categories ahead of the awards in just under three weeks' time.

- *Roczny nominowany do Oscara lunch skupia pretendentów ze wszystkich kategorii, wyprzedzając nagród w czasie niespełna trzy tygodnie.* (Google Translator, 2015)
- *Roczna nominowany do Oscara obiad łączy z pretendentów wszystkie kategorie przed nagród w czasie niespełna trzy tygodnie.* (Google Translator, 2017)
- *Roczne Oscar nominacja obiadowe łączy rywali z wszystkich kategorii przed nagród w niecałe trzy tygodnie.* (Bing Translator)

... na pewno nie

Aplikacje NLP (Natural Language Processing)

- down-to-earth ("słabe" NLP):
  - przeszukiwanie sieci web,
  - poprawa pisowni uwzględniająca kontekst,
  - analiza trudności tekstu; ustalenie autorstwa,
  - identyfikacja nazw własnych (osoby, firmy, nazwy geograficzne),
- far-reaching ("silne" NLP)
  - zamiana tekstu na dane strukturalne/zapis formalny – „prawdziwe" rozumienie tekstu,
  - wnioskowanie i podejmowanie decyzji na podstawie danych pozyskanych z tekstu (odpowiedzi na pytania),
  - systemy dialogowe (dialog sterowany przez użytkownika).
- aplikacje "flagowe":
  - tłumaczenie Google translator
  - rozumienie mowy, wydawanie poleceń: asystenci typu Apple's Siri

Aplikacje NLP (Natural Language Processing)

- down-to-earth ("słabe" NLP):
  - przeszukiwanie sieci web,
  - poprawa pisowni uwzględniająca kontekst,
  - analiza trudności tekstu; ustalenie autorstwa,
  - identyfikacja nazw własnych (osoby, firmy, nazwy geograficzne),
- far-reaching ("silne" NLP)
  - zamiana tekstu na dane strukturalne/zapis formalny – „prawdziwe" rozumienie tekstu,
  - wnioskowanie i podejmowanie decyzji na podstawie danych pozyskanych z tekstu (odpowiedzi na pytania),
  - systemy dialogowe (dialog sterowany przez użytkownika).
- aplikacje "flagowe":
  - tłumaczenie Google translator
  - rozumienie mowy, wydawanie poleceń: asystenci typu Apple's Siri

Aplikacje NLP (Natural Language Processing)

- down-to-earth ("słabe" NLP):
  - przeszukiwanie sieci web,
  - poprawa pisowni uwzględniająca kontekst,
  - analiza trudności tekstu; ustalenie autorstwa,
  - identyfikacja nazw własnych (osoby, firmy, nazwy geograficzne),
- far-reaching ("silne" NLP)
  - zamiana tekstu na dane strukturalne/zapis formalny – „prawdziwe" rozumienie tekstu,
  - wnioskowanie i podejmowanie decyzji na podstawie danych pozyskanych z tekstu (odpowiedzi na pytania),
  - systemy dialogowe (dialog sterowany przez użytkownika).
- aplikacje "flagowe":
  - tłumaczenie Google translator
  - rozumienie mowy, wydawanie poleceń: asystenci typu Apple's Siri

Dwa nurty – formalny i statystyczny

Metody formalne:

- wyrażenia regularne,
- gramatyki formalne,
- logiki nieklasyczne, ...

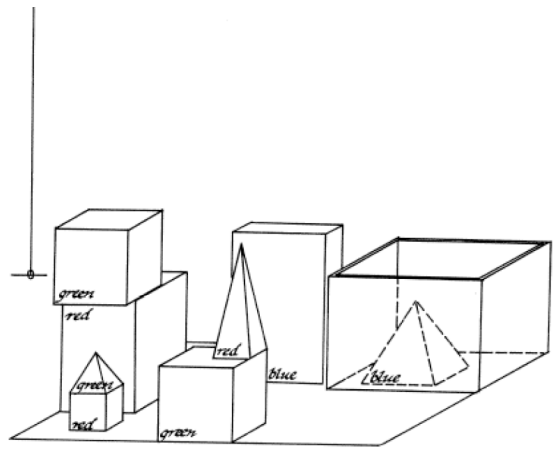
Nurt “statystyczny”:

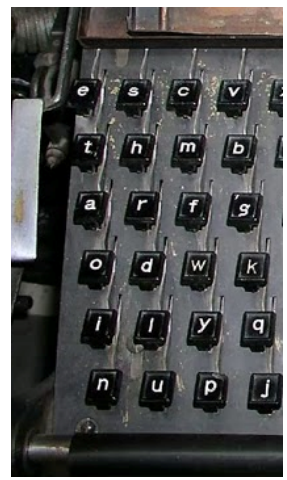
- statystyki częstościowe,
- metody maszynowego uczenia się, klasyfikatory, modele liniowe.
- sieci neuronowe ...

SHRDLU

SHRDLU, a simple dialog with a user, about a small world of objects , 1973

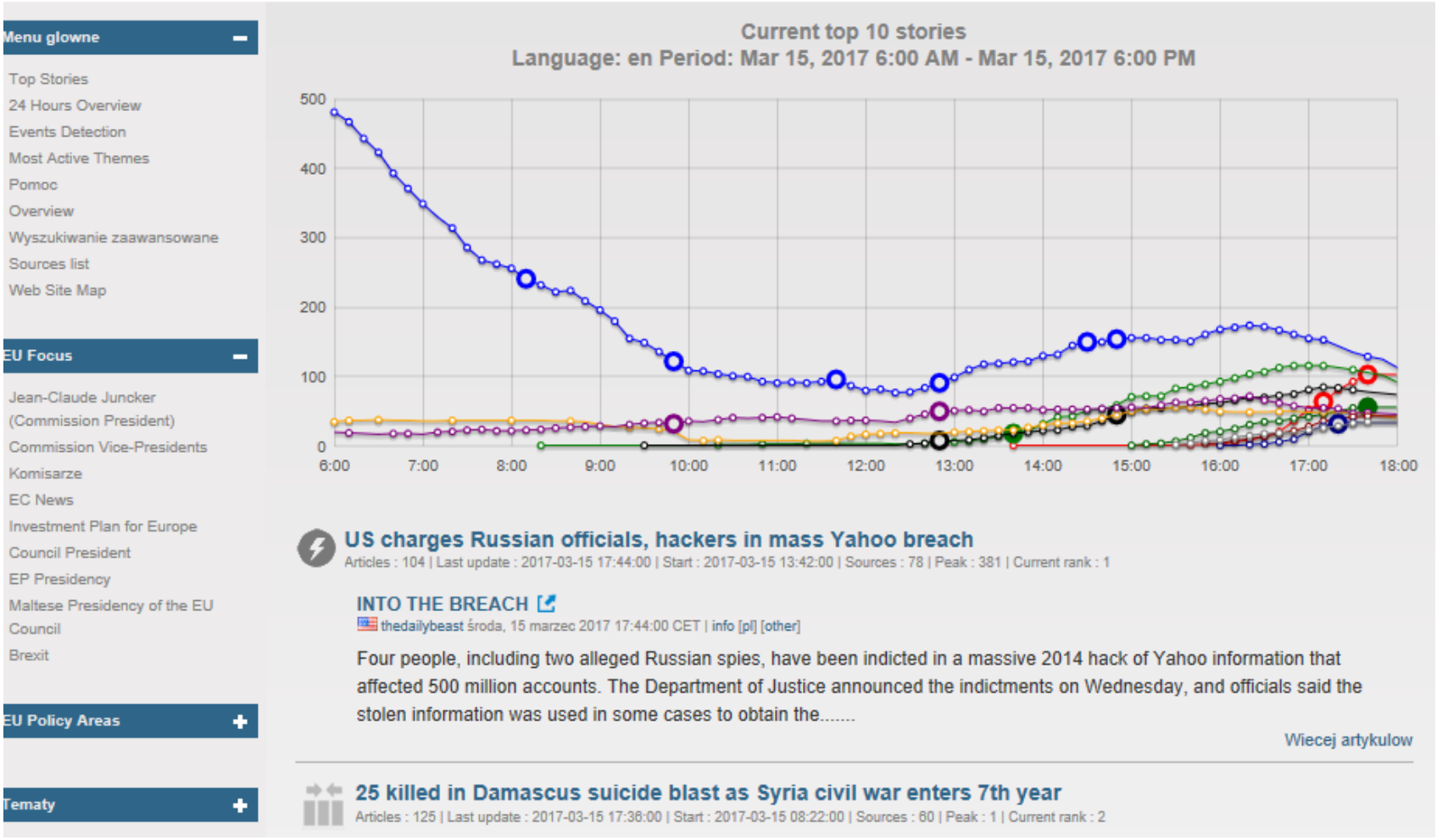
```
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.
Person: IS IT SUPPORTED?
Computer: YES, BY THE TABLE.
```





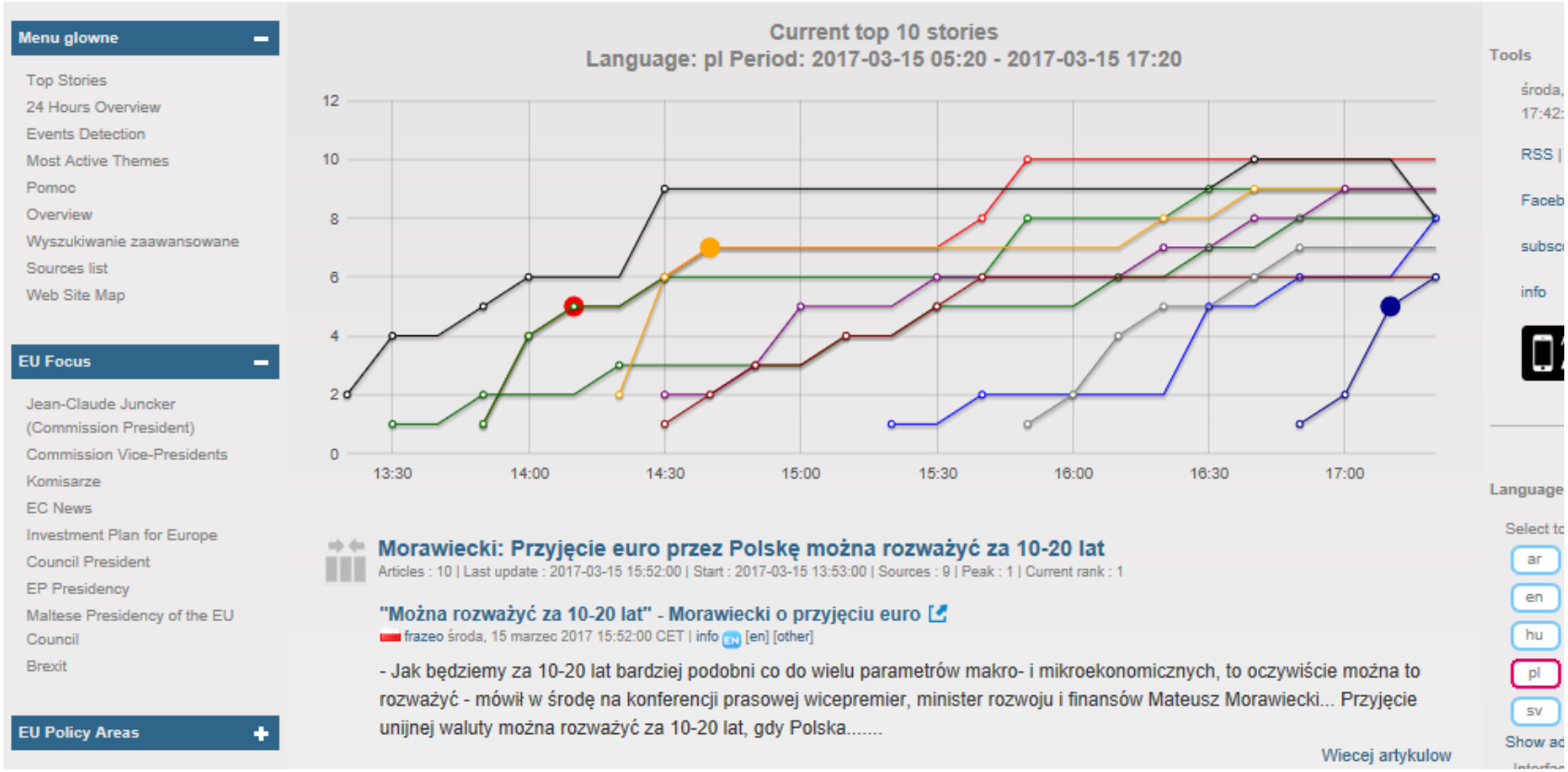
Media Monitor

http://emm.newsbrief.eu/NewsBrief/clusteredition/pl/latest.html



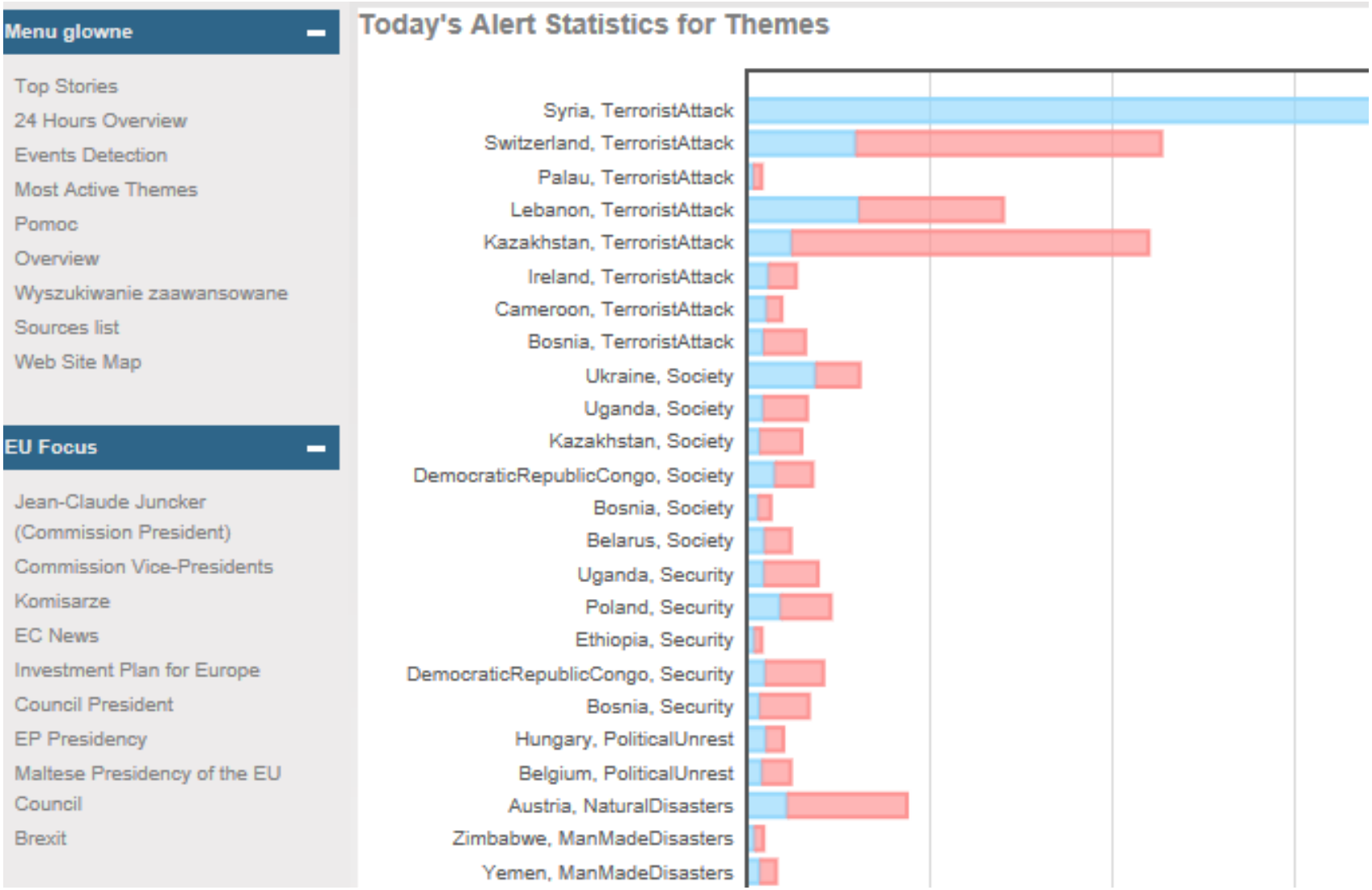
Media Monitor

http://emm.newsbrief.eu/NewsBrief/clusteredition/pl/latest.html



Media Monitor

http://emm.newsbrief.eu/NewsBrief/clusteredition/pl/latest.html



Co osiągnęliśmy?

- Żaden system komputerowy nie modeluje w sposób satysfakcjonujący komunikacji w języku naturalnym

  $\Downarrow$

- dla prawie wszystkich zadań są tylko rozwiązania przybliżone

Automatyczna analiza języka, ”wyznacznik celu”

Dave Bowman:

Open the pod bay doors, HAL.

HAL:

I’m sorry Dave, I’m afraid I can’t do that.

Stanley Kubrick, Arthur C. Clarke
2001: A Space Odyssey (1968)

Komunikacja w języku naturalnym - etapy

- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ $c1$ – *aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Komunikacja w języku naturalnym - etapy

- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ $c1$ – *aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Komunikacja w języku naturalnym - etapy

- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ $c1$ – *aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Komunikacja w języku naturalnym - etapy

- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ *c1 – aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Komunikacja w języku naturalnym - etapy

- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ *c1 – aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Komunikacja w języku naturalnym - etapy

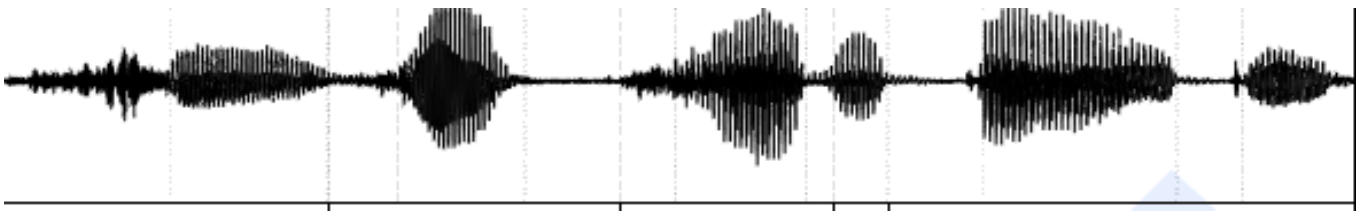
- mowa: ...
- tekst
  Open the pod bay doors, HAL.
- reprezentacja znaczenia
  $r1 = REQUEST(e, open) \wedge agent(e, HAL) \wedge object(e, poddoors)$
- interpretacja w kontekście (wiedza, cel, sytuacja ...)
  $(r1 \Rightarrow t1 \wedge t1 \Rightarrow c2) \wedge c2$ contradicts with $c1$ *c1 – aktualny cel*
- reakcja, wybór rodzaju odpowiedzi: decline $r1$
- sformułowanie słowne odpowiedzi:
  I'm sorry Dave, I'm afraid I can't do that.

Przyczyny trudności

1. problemy z rozstrzyganiem **niejednoznaczności**:
   - wyrażenia językowe mają często (częściej niż sobie to uświadamiany) więcej niż jedno znaczenie,
   - prawidłowe rozstrzygniecie niejednoznaczności wymaga często danych z wielu poziomów analizy;
2. brak wystarczająco ogólnych **metod reprezentacji semantyki**;
3. jest **wiele języków** naturalnych różniących się między sobą nie tylko na poziomie syntaktycznym.

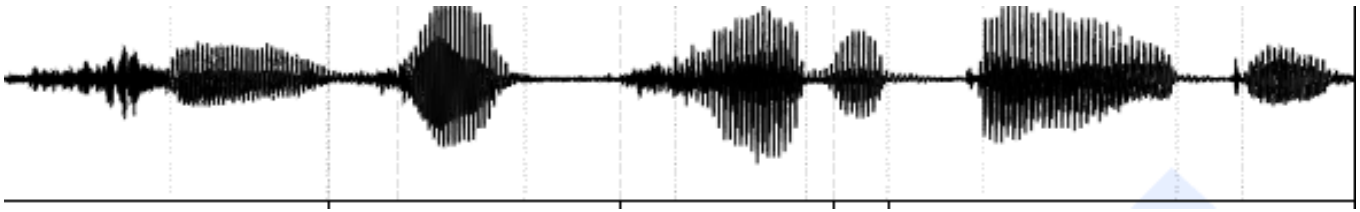
ASR – rozpoznawanie mowy

- $\rightarrow$ fala akustyczna $\rightarrow$ opis numeryczny wybranych cech



ASR – rozpoznawanie mowy

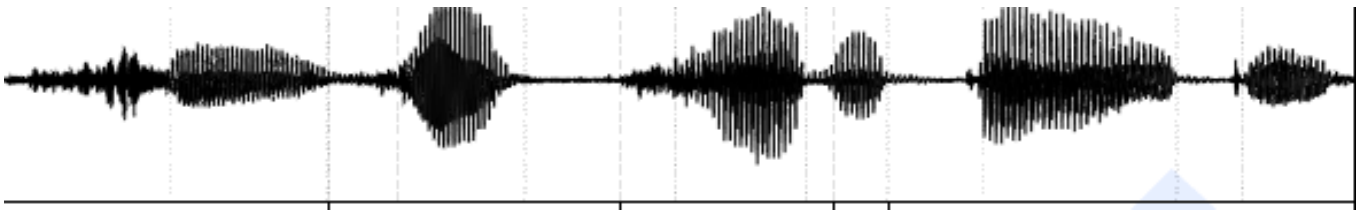
- $\rightarrow$ fonemy



sh ly j ax s h ae dx ax b ey b ly

ASR – rozpoznawanie mowy

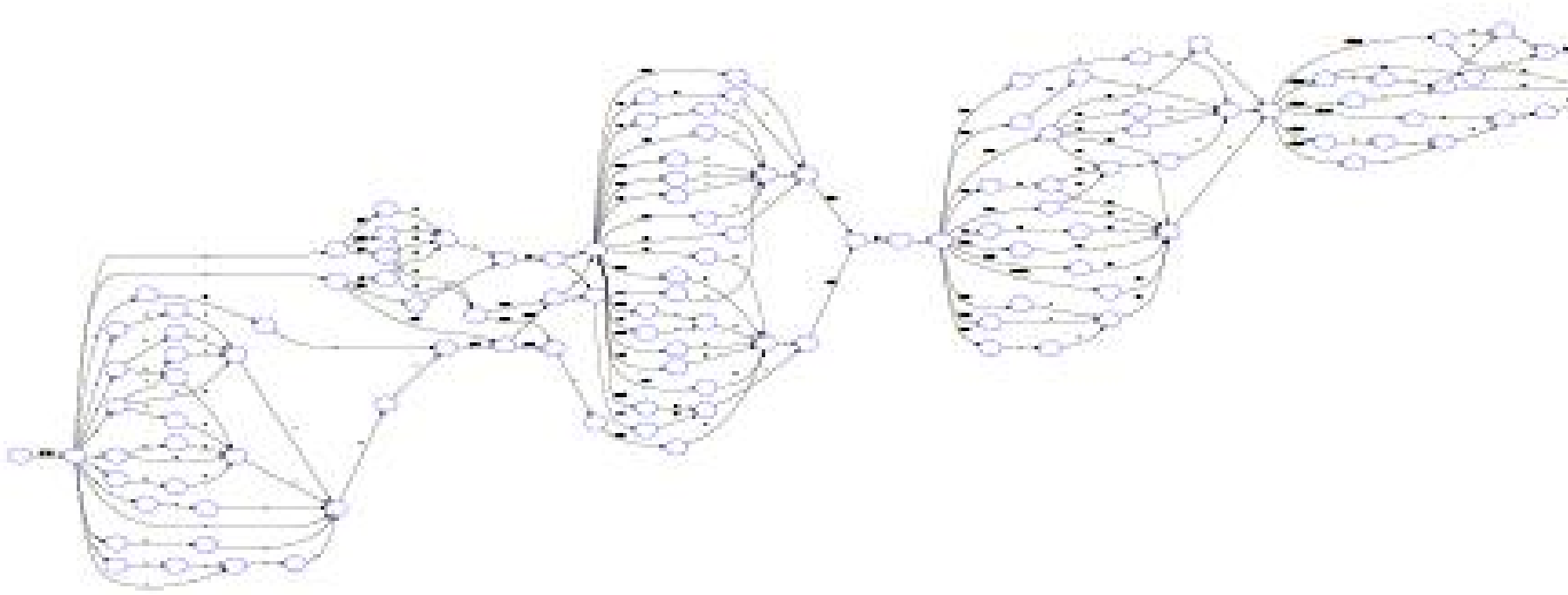
- fonemy $\rightarrow$ słowa



```
sh  ly  j  ax   s  h  ae dx ax b   ey   b ly

She     just     had      a   baby
```

ASR – rozpoznawanie mowy



Główne przyczyny niejednoznaczności:

- różnice w wymowie, barwie głosu, niestaranność wymowy ...
- hałasy tła,
- nakładające się głosy wielu mówców,
- niejednoznaczności fonologiczne, m o rz e $\rightarrow$ może, morze

Analiza tekstu – podział na zdania

Teorie syntaktyczne opisują pojedyncze zdania, trzeba zatem je wydzielić.

- zdanie: *sekwencja segmentów zakończona kropką (? ! ...)*?
- niejednoznaczność kropki

W 1964 r. Akademia Krakowska obchodziła swoje 600-lecie.

Wysiłki zmierzające do ustalenia wojennej przeszłości bośniackich imigrantów rozpoczęły się w 2004 r. Informacje przekazywane służbom imigracyjnym USA ...

Dz.U. 1998 nr 137 poz. 887. Ustawa z dnia 13 października 1998 r. o systemie ubezpieczeń społecznych

Analiza tekstu – segmentacja/tokenizacja

Identyfikacja elementów zdania (słowa, liczby, daty, adresy www ...).

Najprostsze kryterium: sekwencja znaków między spacjami, znakami interpunkcyjnymi lub granicami wierszy.

- czasem to za mało: www.ipipan.waw.pl 3,5 100 000
- czasem za dużo:

  dwuipółwieczny, dwuipółletni, dwuipółgodzinny, dwuipółmilowy, ...

  biało-czerwony, niebiesko-pomarańczowo-żółty
  poszedłbym; bym poszedł
- niejednoznaczności segmentacji:

  Coś ty zrobił? / Co ty zrobiłeś / Coś zostało zrobione.

Analiza morfologiczna

Morfeusz-SGJP

- identyfikacja form podstawowych (haseł słownikowych), kategorii gramatycznej i wartości cech morfologicznych
- rezultaty są różne, w zależności od przyjętych reguł segmentacji i analizatora morfologicznego, ale zawsze są niejednoznaczne

0	1	Dla	**dla**	prep:gen
1	2	brata	**brat**	subst:sg:gen.acc:m1
			bratać	fin:sg:ter:imperf
2	3	Jacka	**[NIEZNANE]**	
3	4	kawa	**kawa**	subst:sg:nom:f
4	5	rozpuszczalna	**rozpuszczalny**	adj:sg:nom:f:pos
5	6	to	**ten**	adj:sg:nom.acc:n1.n2:pos
			to	conj
			to	pred
			to	qub
			to	subst:sg:nom.acc:n2
6	7	nie	**nie**	qub
			on	ppron3:sg:acc:n1.n2:ter:_:praep
				ppron3:pl:acc:m2.m3.f.n1.n2.p2.p3:ter:_:praep
7	8	jest	**być**	fin:sg:ter:imperf
8	9	kawa	**kawa**	subst:sg:nom:f

Analiza morfologiczna

Morfeusz-SIaT

dla	dla	prep:gen
brata	brat	subst:sg:gen.acc:m1
	bratać	fin:sg:ter:imperf
Jacka	Jacek	subst:sg:gen.acc:m1
	Jack	subst:sg:gen.acc:m1
kawa	Kawa	subst:sg:nom:m1
	kawa	subst:sg:nom:f
rozpuszczalna	rozpuszczalny	adj:sg:nom.voc:f:pos
to	ten	adj:sg:nom.acc.voc:n1.n2:pos
	to	conj
	to	pred
	to	qub
	to	subst:sg:nom.acc:n2
nie	nie	conj
	nie	qub
	on	ppron3:sg:acc:n1.n2:ter:_:praep|
		ppron3:pl:acc:m2.m3.f.n1.n2.p2.p3:ter:_:praep
jest	być	fin:sg:ter:imperf
kawa	Kawa	subst:sg:nom:m1
	kawa	subst:sg:nom:f

Analiza syntaktyczna

- identyfikacja fraz i ich powiązań – budowa drzewa struktury
- niektóre niejednoznaczności są eliminowane, ale powstają kolejne:

  Dla brata$_{gen}$ Jacka$_{gen}$ kawa rozpuszczalna to nie jest kawa.

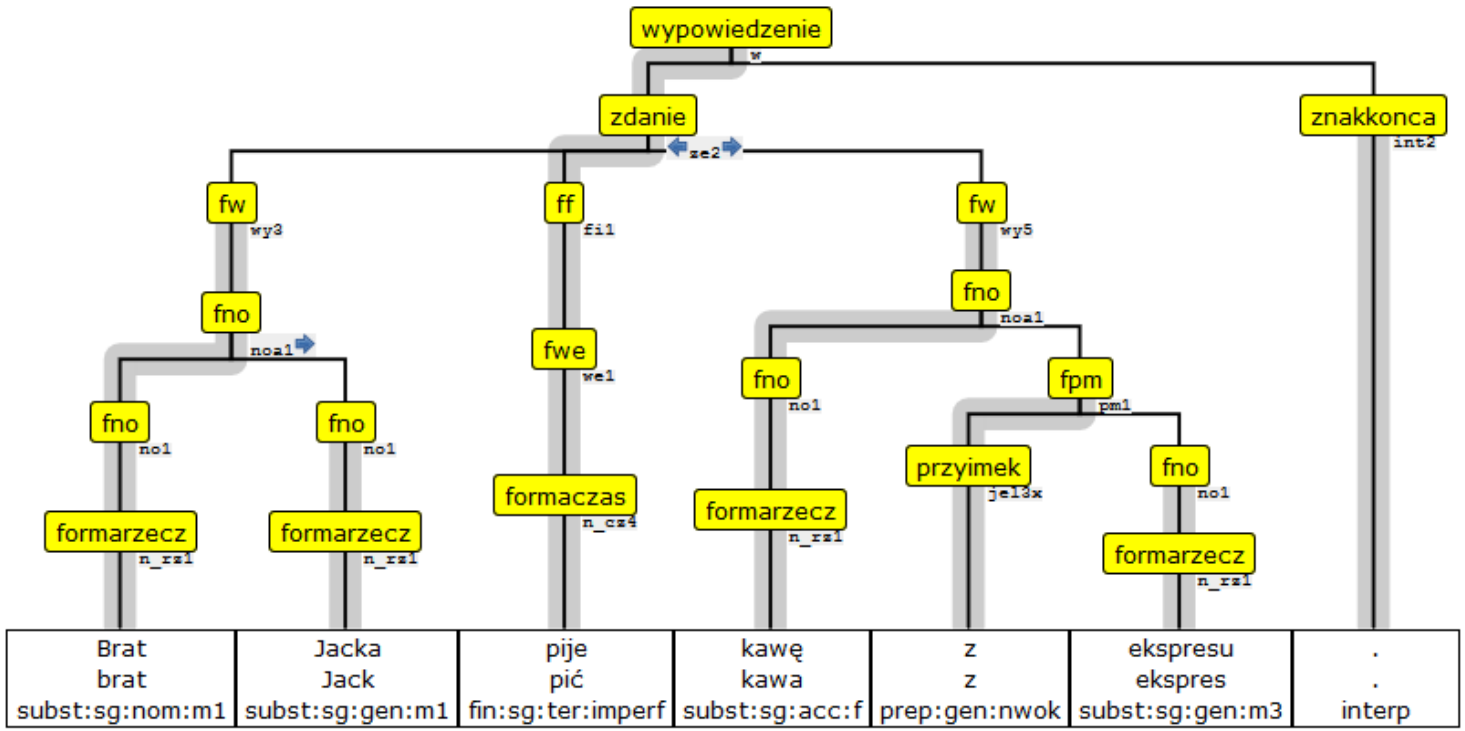
  [dla brata Jacka] [kawa rozpuszczalna] ...

  [dla brata] [Jacka kawa rozpuszczalna] ...

Analiza syntaktyczna

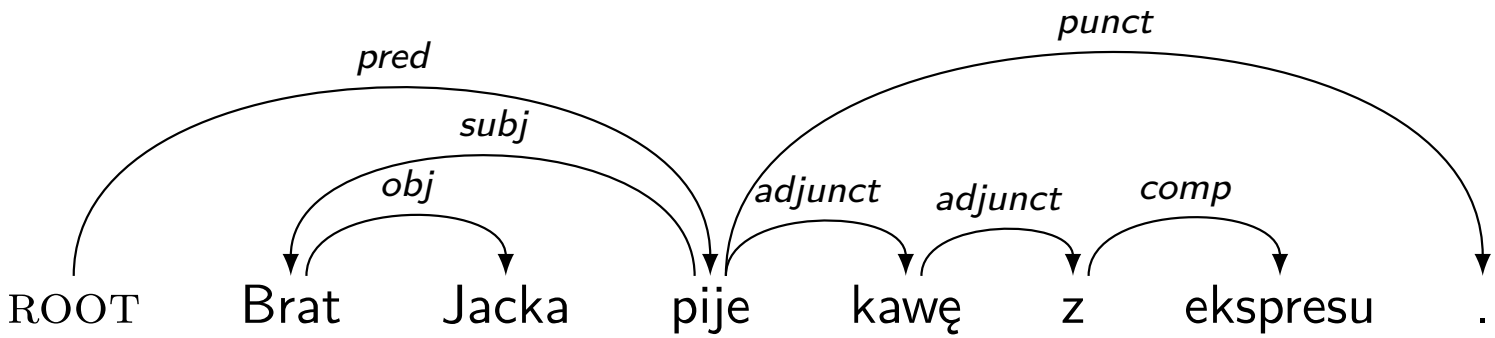
- drzewo rozbioru zgodne z gramatyką frazową (parser Świgra, Woliński)

Drzew:	40
Czas analizy (s):	0.10925884999983282
Krawędzi:	153
Użytecznych krawędzi:	71
Kroków wywodu:	121246



Analiza syntaktyczna

- analiza zależnościowa (parser zależnościowy, Wróblewska)



Analiza semantyczna

- przypisanie interpretacji (na ogół rozbiorowi syntaktycznemu)

  Jan chodzi [do kina] [z przyjaciółmi].

predykat:	chodzić
aktor:	Jan
miejsce docelowe:	kino
??:	przyjaciele

Cats are mammals	$\forall x Cat(x) \implies Mammal(x)$
I have a cat	$\exists x Cat(x) \wedge Own(I, x)$

Analiza semantyczna

- niektóre rozbiory syntaktyczne odpowiadają niemożliwym/mało prawdopodobnym interpretacjom semantycznym

  Jan chodzi [do kina] [z przyjaciółmi].

  *Jan chodzi [do [kina [z przyjaciółmi]]].

  Jan chodzi [do kina] [z nudów].

  Jan chodzi [do [kina z [dobrym nagłośnieniem]]].

  *Jan chodzi [do kina] [z [dobrym nagłośnieniem]].

- czasem różne rozbiory syntaktyczne odpowiadają temu samemu znaczeniu, czasem różnym

  Po dwunastej jest spotkanie [w biurze] [w [pokoju 101 [na pietrze]]]

  Po dwunastej jest spotkanie [w biurze] [w pokoju 101] [na pietrze]

Rozstrzyganie niejednoznaczności – uwzględnienie kontekstu

Większe fragmenty tekstu czy konwersacji, wiedza o świecie pomagają wybrać interpretację

Pacjent opuścił salę operacyjną w dobrym stanie

Kto/co było w dobrym stanie?

pacjent opuścił [salę operacyjną] [w dobrym stanie]

pacjent opuścił [salę operacyjną [w dobrym stanie]]

Przykład 1 – ekstrakcja terminologii

Author keywords:	Relation Extraction Word Embeddings PU Learning Knowledge Extraction
EasyChair keyphrases:	word embedding (190), vector space (110), semantic relation (100), relationship extraction (90), lexical unit (80), positive sample (80), text corpus (70), desired semantic relation (63), seed set (60), semantic information (40), euclidean similarity (40), seed recall (40)
Abstract:	This paper discusses extraction of semantic relationships between lexical units from text data using deep learning technique of representation called word embeddings. We examine the vector space created using this technique and~assess its ability to preserve semantic information. We propose a method for extracting semantic relationship between words using exemplary seed set as training data for machine learning algorithm.

Ekstrakcja terminologii

1	,	zapalenie (116), ma zapalenie (5), wirusowe zapalenie (5), nerwobóle i zapalenie (4)
2	.	zapalenie płuc . (104), zapalenie opon mózgowych . (16), zapalenie . (13), . zapaleni
3	płuco	zapalenie płuc (273), zapalenie oskrzeli i płuc (2), płucu wywiązało się zapalenie (1)
4	na	na zapalenie (158), na wirusowe zapalenie (21), na reumatoidalne zapalenie (6), na
5	i	i zapalenie (34), zapalenie płuc i (32), zapalenie oskrzeli i (14), zapalenie opon mózg
6	mieć	mam zapalenie (18), ma zapalenie (17), miał em zapalenie (8), miała m zapalenie (5)
7	być	em zapalenie (11), m zapalenie (6), jest zapalenie (5), było zapalenie (3), był chory m
8	-	- zapalenie (17), zapalenie opon mózgowo - (5), zapalenie płuc - (4), - wirusowe zap
9	wątroba	zapalenie wątroby (68), zapalenie wirusowe wątroby (2), zapalenie trzustki i wątrob
10	(	(zapalenie (28), zapalenie wątroby ((8), zapalenie stawów ((4), (przewlekłe zapale
11	się	się zapalenie (18), zapalenie się (14), się na zapalenie (3), się przyplątało zapalenie
12	oskrzele	zapalenie oskrzeli (47), zapalenie płuc i oskrzeli (8), oskrzeli ; zapalenie (1), zapaler
13	wirusowy	wirusowe zapalenie (47), zapalenie wirusowe (2), zapalenie (1), wirusową mózgu (z
14	w	zapalenie płuc w (8), w zapalenie (5), w szpitalu na zapalenie (3), zapalenie w (3), za
15	opona	zapalenie opon (46), zapalenie mózgu i opon (3), opon (2), zapalenie (1)

Ekstrakcja terminologii

Struktura gramatyczna terminów w jezyku polskim

- rzeczownik, akronim lub skrót rzeczownika:
  dochód, pasteryzacja
  EKG, ONZ
  ust.(awa), (ęp)
- rzeczownik z przymiotnikiem:
  stosunki gospodarcze
  kodeks karny
- sekwencja rzeczownika z rzeczownikiem w dopełniaczu:
  zapalenie_{nom} płuc_{gen}
  $\text{kodeks}_{n;nom}$ $\text{pracy}_{n;gen}$
- kombinacja powyższych struktur:
  europejski rynek usług finansowych
  wodonercze niewielkiego stopnia dolnego układu podwójnego nerki prawej
- frazy przyimkowe, koordynacja ...

Ekstrakcja terminologii

Identyfikacja kandydatów

- wejście: dane oznaczone morfologicznie (POS, wartości cech morfologicznych)
- kandydatami na terminy są frazy zidentyfikowane za pomocą reguł syntaktycznych
- przykładowa gramatyka (formalizm programu TermoPI):

$$
\begin{aligned}
NPP &: \$NAP \;\; NAP\_GEN^*; \\
NAP[agreement] &: AP^* \;\; N \;\; AP^*; \\
NAP\_GEN[case = gen] &: NAP; \\
AP &: ADJ \mid ADJA \;\; DASH \;\; ADJ \mid PPAS;
\end{aligned}
$$

N[*pos=subst,ger*]; $DASH[form="-"]$; $ADJ[pos=adj]$;
$ADJA[pos=adja]$; $PPAS[pos=ppas]$;

Ekstrakcja terminologii

Lista frekwencyjna fraz dla danych szpitalnych

Show 1000 top-ranked terms Multi-word terms only Search:

#	Rank	Term	C-value	Length	Freq_s	Freq_in	Context #
1	16	badanie	1372,25	1	13753	10383	340
2	34	oddział	836,14	1	8443	4977	61
3	50	stan	583,46	1	5902	4986	74
4	58	leczenie	523,55	1	5259	2181	93
5	64	zalecenie	475,33	1	4771	2071	117
6	69	krew	460,8	1	4687	3237	41
7	71	mocz	446,54	1	4494	1488	52
8	76	dzień	432,58	1	4340	924	65
9	81	wynik	401,1	1	4049	3725	98
10	84	szpital	384,94	1	3861	396	34
11	88	ciało	376,99	1	3853	3159	38
12	99	pacjent	339,94	1	3492	3333	36
13	95	MG	344,9	1	3453	12	3
14	100	zmiana	337,93	1	3385	1488	260
15	113	leczenie	317,32	1	3209	2505	70
16	111	rozpoznanie	319,84	1	3200	54	34
17	114	płeć	316,93	1	3178	61	7
18	1	kod pacjenta	3116	2	3116	0	0
19	2084	kod	0	1	3116	3116	1
20	117	epikryza	309,77	1	3101	23	7
21	2	wynik BADAŃ	3055,92	2	3092	1335	37
22	125	norma	289,84	1	2922	424	18

Ekstrakcja terminologii

Rangowanie fraz

- slightly modified C-value coefficient (Frantzi et al, 2000) comprising both one-word and multi-word phrases in one terminology list:

$$C\text{-}value(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)) & if \;\; r(LP) > 0, \\ l(p) * freq(p), & if \;\; r(LP) = 0 \end{cases} \quad (1)$$

- l increases weight for longer phrases. It is equal to the logarithm of phrase length for multi-word expressions and a constant (e.g. 0.1) for one word terms.
- LP is a set of different phrases containing p, and $r(LP)$ is the number of their types,
- $r(LP)$ is counted as a number of different pairs of the nearest left and right words combined together.

Ekstrakcja terminologii

"nadmiarowe" elementy fraz

- słowa wskazujące na określenie czasu, jak np: miesiąc, czwartek, styczeń, poniedziałek
- przymiotników wymagających kontekstu do interpretacji np: inny, niektóry, jakiś, pewien, dalszy
- przyimki złożone:
  [w kierunku] zapalenia nerek $\rightarrow$ kierunek zapalenia nerek
  [pod postacia] podatku VAT $\rightarrow$ postać podatku VAT
  [pod katem] diagnostyki obrazowej $\rightarrow$ kat diagnostyki obrazowej;
  [pod katem] prostym $\rightarrow$ kat prosty

Ekstrakcja terminologii

Wyniki (dla podanej gramatyki i wybranego zbioru danych medycznych)

- poprawne terminy medyczne, np. ciśnienie tętnicze krwi
- terminy ogólne: pora nocy, kolejna próba
- terminy niepoprawne; błędy mogą wynikać z:
  - niedostatków gramatyki:
    dziewczynka skierowana z frazy dziewczynka skierowana do chirurga;
    5-letni chłopiec
  - błędów anotacji:
    Lacidofil zalecenia (zalecenia otagowane jako dopełniacz a nie mianownik);
  - złego podziału fraz na podfrazy:
    infekcja dróg, USG jamy.

Ekstrakcja terminologii

Niepoprawne podterminy

- Przykłady frazy o silnym powiązaniu słów:

  w medycynie: pęcherzyk żółciowy, jama brzuszna, staw kolanowy
  w ekonomii: papiery wartościowe, fundusz inwestycyjny
  w angielskim: contact lens

- Gramatycznie poprawne zagnieżdżone frazy, które nie są terminami dziedzinowymi:

  [zapalenie pęcherzyka] żółciowego
  [USG jamy] brzusznej
  [operacja lewego stawu] kolanowego
  [giełda papierów] wartościowych
  [uczestnik funduszu] inwestycyjnego
  [soft contact] lens

Ekstrakcja terminologii

Próba “karania” niepoprawnych podfraz – NPMI

Wydzielanie fraz zaczynamy od “najsłabszego” miejsca. Siłę związku mierzymy jako:

Normalised Pointwise Mutual Information

$$NPMI(x,y) = \left(ln \ \frac{p(x,y)}{p(x)p(y)} \right) \Big/ - ln \ \ p(x,y) \qquad (2)$$

‘x y’ jest bigramem składajacym sie z lematów tokenów x i y,
p(x,y) jest prawdopodobinstwem bigramu ‘x y’ w korpusie,
p(x), p(y) jest prawdopodobienstwem unigramów ‘x’ i ‘y’ w korpusie.

Ekstrakcja terminologii

Efekt uwzględnienia NPMI

nominalna **roczna** **stopa** **procentowa**

nominal annual *interest* *rate*

.436 **.456** **.802**

nominalna roczna stopa
roczna stopa
stopa $\Longrightarrow$ ***stopa***
roczna stopa procentowa ***roczna stopa procentowa***
stopa procentowa ***stopa procentowa***

Ekstrakcja terminologii

Wyniki programu TermoPL dla danych szpitalnych

#	Rank	Term	C-value	Length	Freq_s	Freq_in	Context #
1	1	kod pacjenta	3116	2	3116	0	0
2	2	karta informacyjna LECZENIA SZPITALNEGO	2560	4	1281	1	1
3	3	wzór białych krwinek	2379,03	3	1501	0	0
4	4	stan ogólny dobry	1955,84	3	1234	0	0
5	5	stan ogólny	1940,77	2	2055	1485	13
6	6	pęcherz moczowy	1854,84	2	1892	1189	32
7	7	wynik badań dodatkowych	1845,69	3	1197	65	2
8	8	wynik BADAŃ	1845,4	2	1848	91	35
9	9	karta informacyjna	1836,4	2	2153	1583	5
10	10	jama brzuszna	1820,61	2	1841	367	18
11	11	badanie ogólne	1783,09	2	1822	428	11
12	12	pęcherzyk żółciowy prawidłowy	1570,7	3	992	1	1
13	13	płytka krwi	1549,4	2	1554	23	5
14	14	nerki prawidłowej wielkości	1494,62	3	943	0	0
15	15	badanie dodatkowe	1482,5	2	1690	1245	6
16	17	układ kielichowo-miedniczkowy nieposzerzony	1351,97	3	854	2	2
17	18	stan dobry	1346	2	1347	1	1
18	19	moczowód niewidoczny	1332	2	1333	2	2
19	20	pęcherz moczowy niewypełniony	1283,82	3	810	0	0

Ekstrakcja terminologii

Frazy ze słowem 'zapalenie'

zapalenie płuc
zapalenie oskrzeli
zapalenie wyrostka robaczkowego
przewlekłe zapalenie wątroby typu
odoskrzelowe zapalenie płuc
zapalenie gardła
ostre zapalenie wyrostka robaczkowego
zapalenie ucha środkowego
podejrzenie ostrego zapalenia wyrostka robaczkowego
zapalenie opon mózgowo-rdzeniowych
kierunek zapalenia wyrostka robaczkowego
przewlekłe zapalenie
powód zapalenia płuc
Obturacyjne zapalenie oskrzeli
atopowe zapalenie skóry
zapalenie ucha
zapalenie wątroby

Przykład 2

Jak zapisać wiedzę, że:

słoń jest bardziej podobny do tygrysa niż do słońca?

Jak stwierdzić, że zdanie:

Samochód Kasi jest czerwony. jest bardziej podobne znaczeniowo do

Kasi auto polakierowane jest na czerwono. niż do

Sweter Kasi jest czerwony.

lub, że na następujące pytania można udzielić tej samej odpowiedzi?

Jak włączyć transmisję danych w telefonie?
Jak mogę włączyć transmisję danych?
Co (trzeba) zrobić by działała transmisja danych w telefonie?
I jeszcze mam pytanie o transmisję danych. Jak ją uruchomić?

Słowa

Wiele relacji między słowami

hiperonimia/hiponimia	słoń	$\rightarrow$	ssak
	biec	$\rightarrow$	przemieszczać się
co-hiponimia	słoń	$\rightarrow$	tygrys
	biec	$\rightarrow$	gnać
antonimia	krótki	$\rightarrow$	długi
synonimia	auto	$\rightarrow$	samochód
holonimia/meronimia	auto	$\rightarrow$	silnik

Słowa

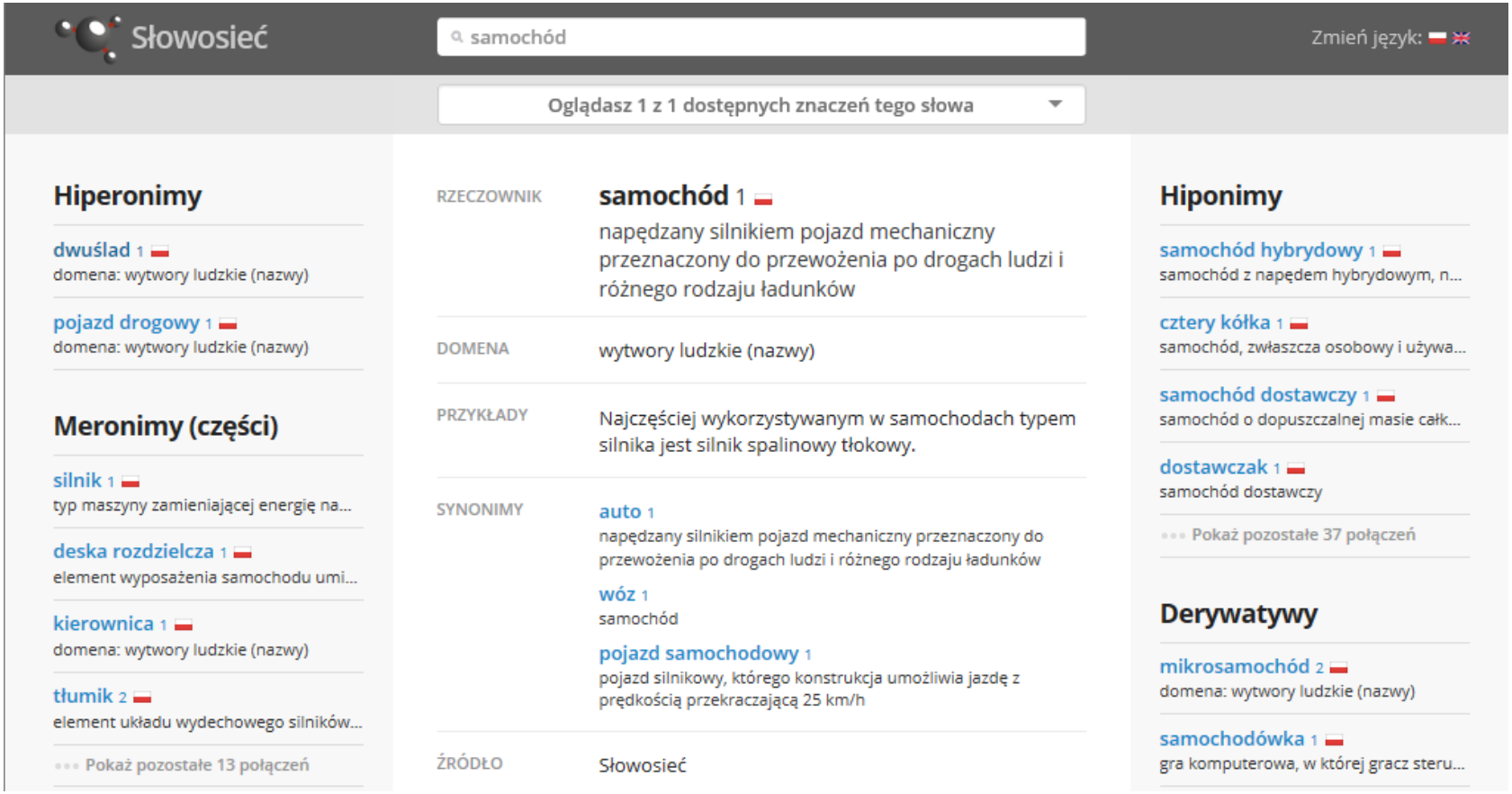
Wiele znaczeń jednego słowa

- baza znaczeń słów plWordnet 3.0, http://plwordnet.pwr.wroc.pl/wordnet/

	Verbs	Nouns	Adjectives	Sum
lemmas	17391	126482	26961	170834
lexical units	31834	166938	45514	244286
synsets	21663	123709	38868	184240
monosemous lemmas	10264	103566	16338	130168
polisemous lemmas	7127	22916	10623	40666

Słowa

plWordnet, przykład hasła



Semantyka dystrybucyjna

Hipoteza dystrybucyjna

Leksemy o podobnych cechach dystrybucyjnych mają podobne znaczenie

Firth (1957): "You shall know a word by the company it keeps!"

Modele dystrybucyjne (DSM) to implementacja hipotezy dystrybucyjnej służąca do dokonywania przybliżonej analizy semantycznej.

Podobieństwo słów, metody dystrybucyjne

tesguino example (Jurafsky & Martin)

Butelka **tesguino** stoi na stole.

Wszyscy lubią pić **tesguino**.

Tesguino sprawia, że jesteś pijany.

Tesguino przyrządzane jest z kukurydzy.

Co wiadomo o **tesguino**?

tesguino jest napojem alkoholowym (wyrabianym z kukurydzy)

Podobieństwo słów, metody dystrybucyjne

tesguino example (Jurafsky & Martin)

Butelka **tesguino** stoi na stole.

Wszyscy lubią pić **tesguino**.

Tesguino sprawia, że jesteś pijany.

Tesguino przyrządzane jest z kukurydzy.

Co wiadomo o **tesguino**?

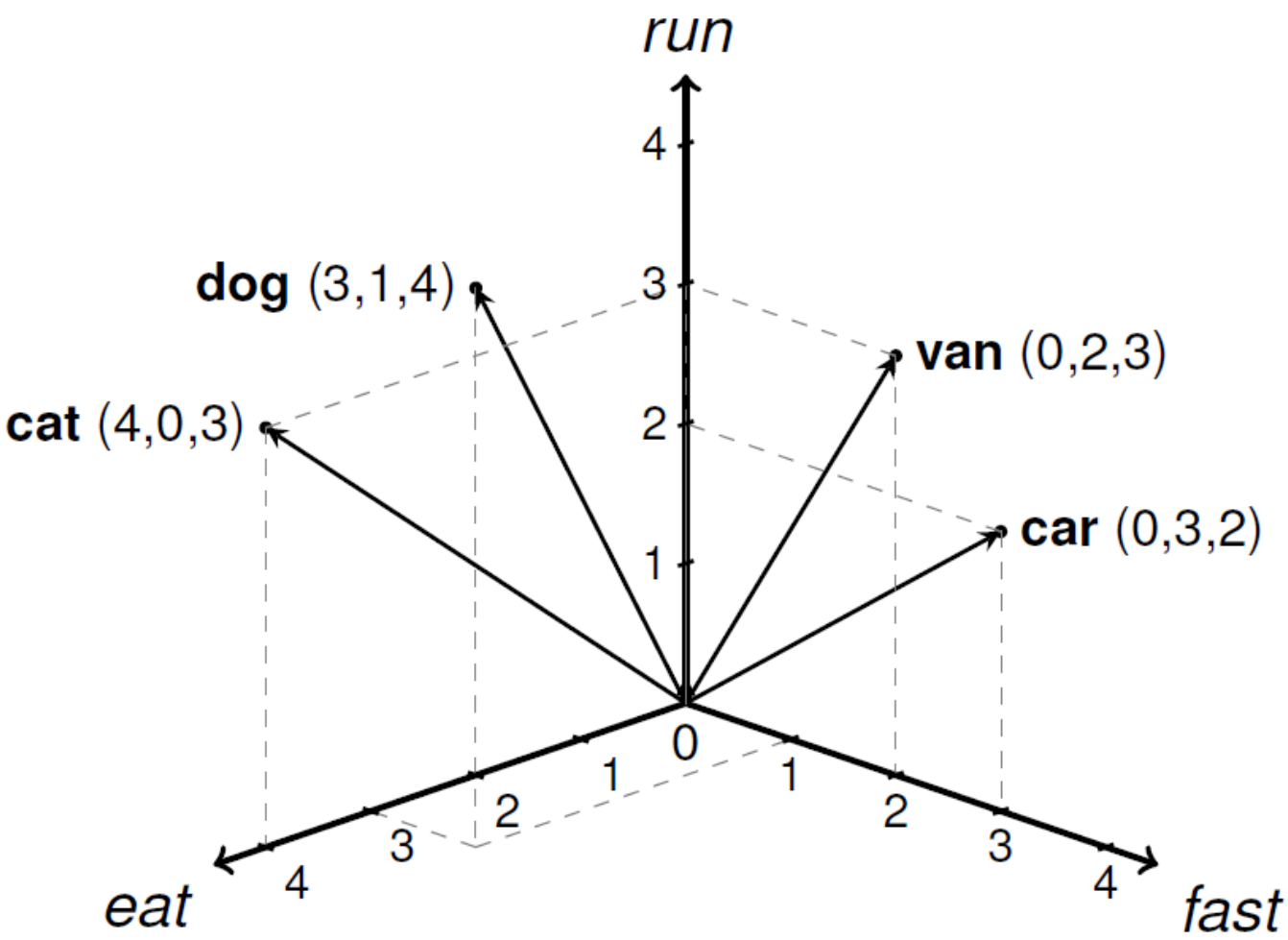
tesguino jest napojem alkoholowym (wyrabianym z kukurydzy)

Wektory współwystępowania

	dog	*drive*	*eat*	*fast*	*play*	...	*the*	*wheel*
car	0	3	0	2	0	⋮	2	1
cat	1	0	3	0	1	⋮	2	0
dog	0	0	3	0	2	⋮	2	0
van	0	3	0	1	0	⋮	3	1

co-occurrence matrix

Wektory współwystępowania



Wektory współwystępowania

Słowa reprezentowane przez bliskie wektory są semantycznie podobne

	car	*cat*	*dog*	*van*
car	1			
cat	0.33	1		
dog	0.60	0.94	1	
van	0.92	0.50	0.76	1

Podobieństwo wektorów można ustalać za pomocą różnie zdefiniowanych miar podobieństwa (najczęściej jest to cosinus kąta miedzy nimi).

Modele dystrybucyjne

Modele dystrybucyjne słów oparte są (bezpośrednio lub pośrednio) na częstościach występowania słów w tekście.

Sposób tworzenia wektorów reprezentacyjnych słowa:

- częstości względne współwystępowania wybranych słów i wybranych kontekstów, znormalizowane z wykorzystaniem PMI, redukcja wymiarów za pomocą LSA
- współczynniki uzyskane z reprezentacji tekstu za pomocą sieci neuronowej - word2vec (Mikolow, 2013)
- inne metody, np. GLoVE (Pennington, J., Socher, R. & Manning, C. 2014).

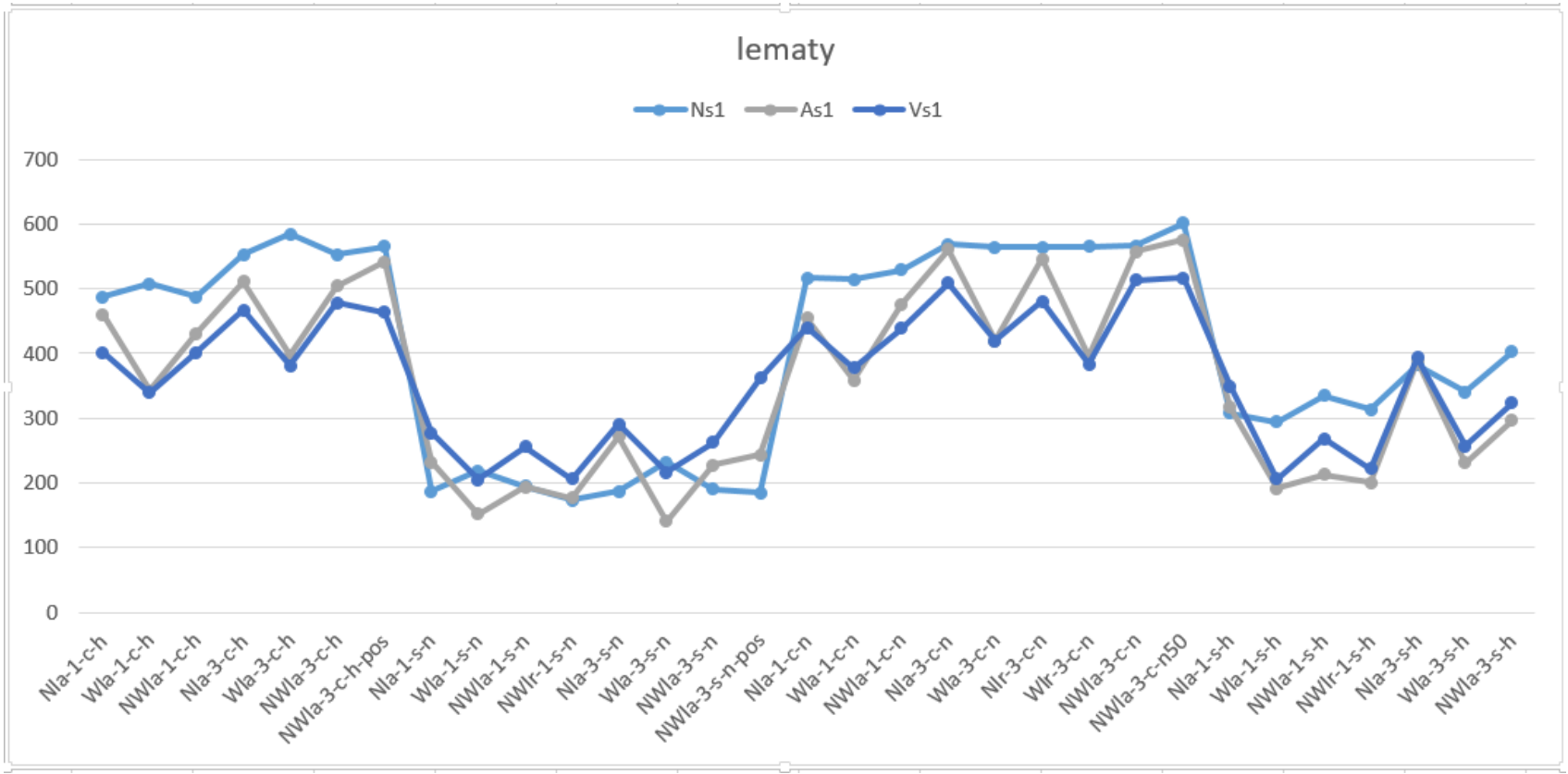
Szukanie słów podobnych

dom plWordnet

podobieństwo	jednostka leksykalna
0.193	domek ⌄
0.186	budynek ⌄
0.173	mieszkanie ⌄
0.152	kamienica ⌄
0.141	domostwo ⌄
0.141	rodzina
0.141	rezydencja
0.139	chata ⌄
0.129	chałupa ⌄
0.126	posiadłość
0.125	pokój
0.123	pałac ⌄
0.120	dwór ⌄
0.118	budowla
0.115	hotel
0.115	stodoła
0.110	gmach ⌄
0.109	garaż
0.109	posesja
0.108	sklep

NKJP+Wikipedia		Wikipedia	
domek	0.73 ∨	pałacyk	0.71
chałupa	0.64 ∨	domek	0.70 ∨
barak	0.62	dworek	0.68 ∨
chata	0.61 ∨	pałac	0.67 ∨
mieszkanie	0.60 ∨	pawilon	0.66
willa	0.60 ∨	szpital	0.65
dworek	0.58 ∨	lokal	0.65
altanka	0.57	pensjonat	0.64
gospoda	0.56	budynek	0.64 ∨
sypialnia	0.56	apartament	0.62
kamienica	0.55 ∨	sierociniec	0.62
świetlica	0.55	dwór	0.60 ∨
domostwo	0.55 ∨	grobowiec	0.60
Domek	0.55	pasaż	0.59
piwnica	0.55	zajazd	0.58
plebania	0.55	ogród	0.58
schronisko	0.55	zameczek	0.58
pałacyk	0.54	kościół	0.57
szałas	0.54	klasztor	0.56
chatka	0.54 ∨	lamus	0.56
pensjonat	0.54	sklep	0.56
suterena	0.53	przytułek	0.55
baraczek	0.53	konwikt	0.55
pałac	0.53 ∨	żłobek	0.55
altana	0.53	ośrodek	0.55
stodoła	0.53	refektarz	0.55
pokoik	0.53 ∨	kościółek	0.55
budynek	0.52 ∨	zakład	0.55

Szukanie słów podobnych



Przypisanie znaczenia słowu w kontekście

WSD - word sense disambiguation

- model dystrybucyjny słów nie rozwiązuje problemu wieloznaczności leksykalnej,
- wektor dystrybucyjny słowa zawiera wszystkie znaczenia,
- w przypadku nierównego rozkładu znaczeń - największą wagę mają znaczenie najczęstsze.

Przypisanie znaczenia słowu w kontekście

WSD - word sense disambiguation

problem 1:	trudność, trudna sytuacja, z którą człowiek musi się zmierzyć
problem 2:	zagadnienie, zadanie do rozwiązania, kwestia do przemyślenia, podjęcia decyzji

Raport podkreśla, że w wielu krajach poważnym problemem/1 jest wysokie bezrobocie.

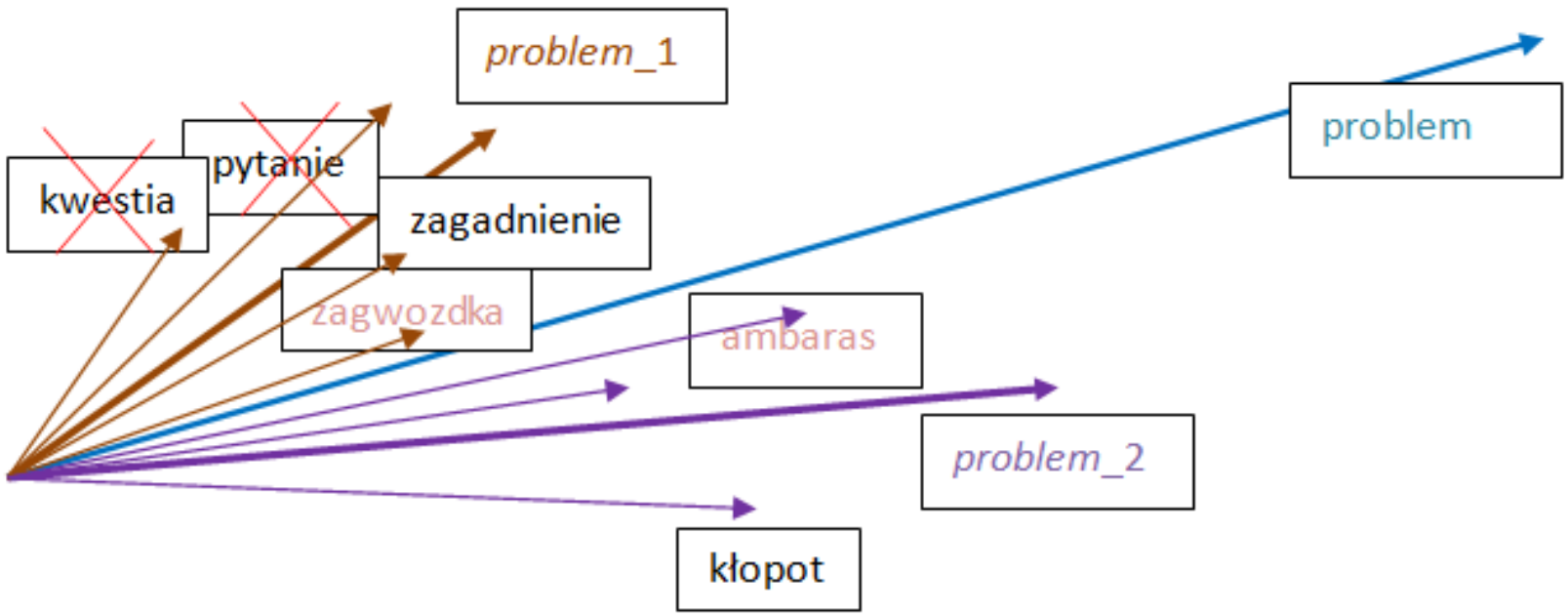
Pozostaje wreszcie problem/2 walorów dydaktycznych książki.

Przypisanie znaczenia słowu w kontekście

WSD - word sense disambiguation

problem_1: kwestia, pytanie, zagwozdka, zagadnienie

problem_2: bieda, kłopot, ambaras



Przypisanie znaczenia słowu w kontekście

Dane

- model dystrybucyjny dla wyznaczonych zestawów słów/znaczeń(Wikipedia).
- anotacja semantyczna Składnicy – zbioru 8241 zweryfikowanych "ręcznie" drzew rozbioru (ok. 50K słów)
- przymiotniki, rzeczowniki i czasowniki w tym zdaniach oznaczone są znaczeniami z plWordnet wersja 2.0,
- 17410 oznaczeń w 2785 (34%) zdaniach; ok. 2072 tokenów w tych zdaniach (12%) nie ma definicji w plWordnet.
- 303 wystąpień 124 słów "obsługiwanych" przez tę metodę, 5571 wystąpień słów jednoznacznych,
- np. słowo 'problem' wystąpiło 38 razy, 23 razy w znaczeniu 1, a 15 w znaczeniu 2, 20 wystąpień oznaczonych poprawnie

Przypisanie znaczenia słowu w kontekście

Metoda

1. metoda nienadzorowana: korzystając z poniższego wzoru (Taddy, 2015) wyznaczamy prawdopodobieństwa zdania ograniczonego do kontekstu wyrazów jednoznacznych i jednego z alternatywnych wektorów sensu i wybieramy to z najwyższym prawdopodobieństwem

$$logp_{\mathcal{V}}(w) = \sum_{j=1}^{T} \sum_{k=1}^{T} \mathbb{1}_{[1 \leqslant |k-j| \leqslant b]} logp_{\mathcal{V}}(w_k | w_j) \tag{3}$$

2. metoda nadzorowana:
   - dane: po 2 wektory prawo- i lewostronnego kontekstu; słowo reprezentowane jest przez wektor reprezentujący wszystkie znaczenia
   - trenowanie modelu składającego się LSTM i gęstej warstwy z sigmoidalną funkcją aktywacji, Keras 1.0.1 `https://keras.io/`)
   - w końcowym kroku wektor wyjściowy sieci porównywany jest z wekorami sensów

Przypisanie znaczenia słowu w kontekście

Metoda

1. metoda nienadzorowana: korzystając z poniższego wzoru (Taddy, 2015) wyznaczamy prawdopodobieństwa zdania ograniczonego do kontekstu wyrazów jednoznacznych i jednego z alternatywnych wektorów sensu i wybieramy to z najwyższym prawdopodobieństwem

$$logp_{\mathcal{V}}(w) = \sum_{j=1}^{T} \sum_{k=1}^{T} \mathbb{1}_{[1 \leqslant |k-j| \leqslant b]} logp_{\mathcal{V}}(w_k | w_j) \tag{3}$$

2. metoda nadzorowana:
   - dane: po 2 wektory prawo- i lewostronnego kontekstu; słowo reprezentowane jest przez wektor reprezentujący wszystkie znaczenia
   - trenowanie modelu składającego się LSTM i gęstej warstwy z sigmoidalną funkcją aktywacji, Keras 1.0.1 `https://keras.io/`)
   - w końcowym kroku wektor wyjściowy sieci porównywany jest z wekorami sensów

Przypisanie znaczenia słowu w kontekście

Rezultaty

Method	Settings	Precision
random baseline	N/A	0.47
MFS baseline	N/A	0.73
pagerank*	N/A	0.52
unsupervised	5 word context	0.507
	10 word	0.529
supervised	750 epochs	0.673
	2000 epochs	0.690
	4000 epochs	0.667

Kędzia et al., wynik dla wszystkich słów

Metody radzenia sobie z problemami

- niejednoznaczności: przybliżenia, wybór najbardziej prawdopodobnego rozwiązania; łączenie podejścia statystycznego i formalnego;
- zapis semantyki i metod wnioskowania: formuły logiczne. ontologie, semantyka dystrybucyjna, semantyka kompozycyjno-dystrybucyjna;
- wielość języków: budowa zasobów i aplikacji wielojęzycznych, baz łączących słowa/nazwy pojęć w wielu językach.

Łączenie nurtu formalnego ze statystycznym.

Metody radzenia sobie z problemami

- niejednoznaczności: przybliżenia, wybór najbardziej prawdopodobnego rozwiązania; łączenie podejścia statystycznego i formalnego;
- zapis semantyki i metod wnioskowania: formuły logiczne. ontologie, semantyka dystrybucyjna, semantyka kompozycyjno-dystrybucyjna;
- wielość języków: budowa zasobów i aplikacji wielojęzycznych, baz łączących słowa/nazwy pojęć w wielu językach.

Łączenie nurtu formalnego ze statystycznym.

Literatura

1. Jurafsky Daniel, Martin James H. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition: Upper Saddle River, 2009
2. Hajnicz, E. 2014. Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proc. of the 7th Int. WordNet Conference (GWC 2014)*, pp 23–31, Tartu, Estonia.
3. Kędzia, P., Piasecki, M. and Orlińska, M. 2015. Word sense disambiguation based on large scale Polish CLARIN heterogeneous lexical resources. *Cognitive Studies— Études cognitives*, 15:269–292.
4. Marciniak, M. and Mykowiecka, A. Terminology Extraction from Medical Texts in Polish. Journal of Biomedical Semantics, 5. (2014)
5. Marciniak, M. and Mykowiecka, A. Nested Term Recognition Driven by Word Connection Strength. Terminology, 21(2), 180–204, (2015)
6. Marciniak, M. and Mykowiecka, A., Rychlik, P. TermoPL - a Flexible Tool for Terminology Extraction, LREC 2016, Portorož, Slovenia, European Language Resources Association (ELRA).
7. Mykowiecka, A., Marciniak, M., Rychlik, P. Recognition of non-domain phrases in automatically extracted lists of terms, Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016), Osaka 2016
8. Wawer, A., Mykowiecka A., Supervised and Unsupervised Word Sense Disambiguation on Word Embedding Vectors of Unambigous Synonyms, Proc. of the SENSE Workshop EACL, Walencja, 2017
9. Woliński, M., Głowińska, K. and Świdziński, M. 2011. A preliminary version of Składnica—a treebank of Polish. In Z. Vetulani, editor, *Proc. of the 5th Language & Technology Conference*, pp 299–303, Poznań, Poland.

Linki

- TermoPL - web service `http://ws.clarin-pl.eu/termopl.shtml`
- TermoPl - program do pobrania `http://zil.ipipan.waw.pl/TermoPL`
- Korpusomat - program do oznaczenia tekstu tagami morfologicznymi `http://korpusomat.nlp.ipipan.waw.pl/`
- Concraft - tagger do pobrania `http://zil.ipipan.waw.pl/Concraft`