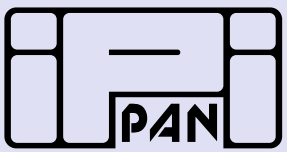
Wykorzystanie semantyki dystrybucyjnej do oceny metaforyczności polskich fraz rzeczownikowych

Agnieszka Mykowiecka, Aleksander Wawer
Małgorzata Marciniak, Piotr Rychlik



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 12 lutego 2018

Outline

Projekt

Projekt NCN

tytuł: *Wykorzystanie metod kompozycyjnej semantyki dystrybucyjnej do identyfikacji i rozróżniania znaczeń w języku polskim*;

akronim: CoDeS;

kierownik: Agnieszka Mykowiecka;

strona: `http://zil.ipipan.waw.pl/CoDeS`

Nieco historii

"the meaning of words lies in their use"

L. Wittgenstein *Philosophical investigations* (1953)

"You shall know a word by the company it keeps!"

J. R. Firth, *Studies in Linguistic Analysis* (1957)

"The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents"

Z. S. Harris, *Distributional structure* w Word (1954)

Realizacja idei

Impulsy rozwoju:

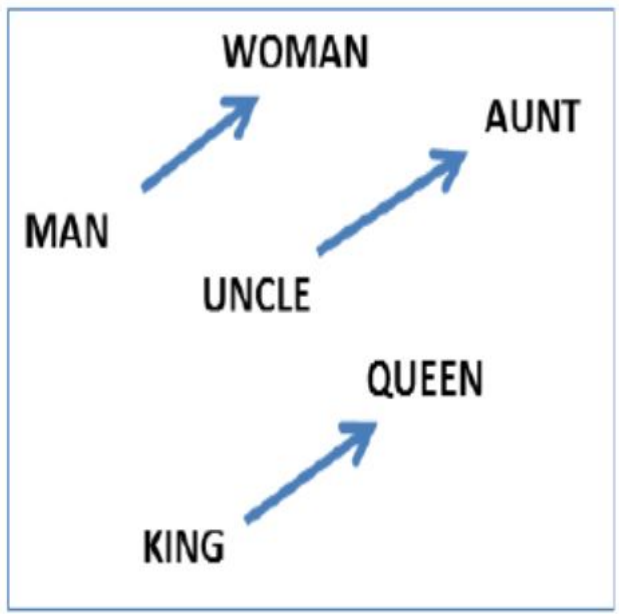
- dostępność korpusów;
- rozwój mocy obliczeniowej komputerów.

Word embeddings (zanurzenie?/inne tłumaczenie?):

- Słowa/formy słów reprezentowane przez wysokowymiarowe wektory.
- Podobieństwo wektorów (miara kosinusowa) zwykle odzwierciedla podobieństwo semantyczne (dla polskich form również morfosyntaktyczne).

Realizacja idei, cd

T. Mikolov, W. Yih, G. Zweig. *Linguistic Regularities in Continuous Space Word Representations*, HLT-NAACL, (2013)



analogia słów: a:b c:d

jeśli x_a, x_b, x_c są odpowiednimi wektorami to $x_d = x_b - x_a + x_c$

Outline

Modele word2vec

W oparciu o korpusy, których słowa stanowią obserwacje, przewidujemy:

- słowo na podstawie okna kontekstu (CBOW — Continuous Bag of Words);
- kontekst na podstawie słowa (skip-gram).

Parametry do ustalenia:

- algorytm uczący: hierarchical softmax, negative sampling;
- rozmiar wektora, zwykle od 50 — 400;
- częstość słów w kontekście;
- wymiar okna kontekstów.

Modele dla polskiego

- Wygenerowane przy pomocy pakietu *gensim* (Python).
- Na podstawie korpusów NKJP, WikiPL, otagowanych przy pomocy programu Concraft-pl (J. Waszczuk, 2012).
- Wygenerowane na podstawie form oraz lematów.
- Wektory o długości 100 i 300.
- Dla wszystkich POS, i ograniczone do słów z 19 POS.
- Ograniczone do słów, które wystąpiły co najmniej 30 lub 50 razy w korpusie.

`http://dsmodels.nlp.ipipan.waw.pl/w2v.html`

A. Mykowiecka, M. Marciniak, P. Rychlik, *Testing word embeddings for Polish*, Cognitive Studies / Études Cognitives, 17, 2017

Formy vs lematy

piernik (lematy)		piernik (formy)		pierników (formy)	
pierniczek	0.650	wiatraka	0.536	ciast	0.756
keks	0.591	kiszonych	0.494	ciastek	0.731
piernikowy	0.586	czekoladopodobny	0.450	piernika	0.724
marcepan	0.570	kołaczy	0.432	tortów	0.709
strucla	0.558	kiszenia	0.426	wypieków	0.697
sękacz	0.556	sernik	0.426	wafli	0.689
marcepanowy	0.540	waniliowy	0.422	kiełbas	0.687
ciasto	0.537	kwaszonych	0.421	placków	0.686
ciastko	0.532	garnka	0.407	ciasteczek	0.675
jabłecznik	0.532	konserwowych	0.400	czekoladowych	0.663
torcik	0.519	orzechowy	0.399	bakalii	0.659
karpatka	0.518	upieczone	0.396	smakołyków	0.655
chałwa	0.514	budyń	0.396	pierogów	0.654
wafel	0.514	pomidor	0.391	ciasta	0.652
bakaliowy	0.511	pasztetem	0.386	makaronu	0.649

Podobieństwo wyrazów wieloznacznych

Model lematowy

myszka:

mysz, myszy, ptaszek, kotek, ZEMAN, króliczek, łapka, miś, myszke, klawiatura, paluszek, małpka, klawisz, Misia, trackball, guziczek, robaczek, ptaszka, stworek, kaczuszka, joystick, Miś, Miki, pyszczek, jeż, piórko, zwierzątko, Pasek, zajączek, Misiek

Podobieństwo wyrazów wieloznacznych

Model lematowy

myszka:

mysz, myszy, ptaszek, kotek, ZEMAN, króliczek, łapka, miś, myszke, klawiatura, paluszek, małpka, klawisz, Misia, trackball, guziczek, robaczek, ptaszka, stworek, kaczuszka, joystick, Miś, Miki, pyszczek, jeż, piórko, zwierzątko, Pasek, zajączek, Misiek

Model dla form

myszką:

myszą, myszki, kursorem, klawiszem, przyciskiem, kliknąć, kliknięciem, klawisz, klawiatury, kursora, wciśnij, ikonkę, klawisza, ikonę, klawiszami, rysikiem, przeciągnij, kliknięcie, strzałkę, wciśniętym, pulpitu, naciśnij, klawiaturę, przycisków, myszy, klawiaturą, kursor, przyciskami, wyświetlaną , klikając

myszka:

mucha, papuga, koza, jeż, kozak, sroka, Kaczor, kret, Kozak, rybka, wilk, . . . inne zwierzęta

Analogie

Zbiór testowy:

- 200 analogii semantycznych podzielonych na 49 typów, takich jak relacje: rodzinne, zawodowe, geograficzne, zdrobnienia itd.,
- 20 analogii gramatycznych: przypadki, liczby, stopnie

Przykłady (model lematowy):

woda – powódź, ogień – ?

Analogie

Zbiór testowy:

- 200 analogii semantycznych podzielonych na 49 typów, takich jak relacje: rodzinne, zawodowe, geograficzne, zdrobnienia itd.,
- 20 analogii gramatycznych: przypadki, liczby, stopnie

Przykłady (model lematowy):

woda – powódź, ogień – ?
pożar, nawałnica, kataklizm, wichura, trzęsienie, pożoga, gradobicie, huragan, zniszczenie, klęska

Analogie

Zbiór testowy:

- 200 analogii semantycznych podzielonych na 49 typów, takich jak relacje: rodzinne, zawodowe, geograficzne, zdrobnienia itd.,
- 20 analogii gramatycznych: przypadki, liczby, stopnie

Przykłady (model lematowy):

woda – powódź, ogień – ?
pożar, nawałnica, kataklizm, wichura, trzęsienie, pożoga, gradobicie, huragan, zniszczenie, klęska

chleb – zboże, ser – ?

Analogie

Zbiór testowy:

- 200 analogii semantycznych podzielonych na 49 typów, takich jak relacje: rodzinne, zawodowe, geograficzne, zdrobnienia itd.,
- 20 analogii gramatycznych: przypadki, liczby, stopnie

Przykłady (model lematowy):

woda – powódź, ogień – ?
pożar, nawałnica, kataklizm, wichura, trzęsienie, pożoga, gradobicie, huragan, zniszczenie, klęska

chleb – zboże, ser – ?
rzepak pszenica, kukurydza, wołowina, nasiono, sadzeniak, ziemniak, pomidor, burak, warzywo

Wnioski

- Modele lematowe dają lepsze wyniki dla analogii, podczas gdy modele oparte na formach dają lepsze wyniki dla synonimii.
- Modele CBOW dają lepsze wyniki dla analogi, podczas gdy modele skip-gram dają lepsze wyniki dla synonimii (dla angielskiego większość badań wskazuje na model skip-gram z negative sampling).
- Zwiększanie rozmiaru wektorów i korpusów, na podstawie których są tworzone wektory, nie koniecznie prowadzi do polepszenia wyników.
- Lepszej jakości korpusy polepszają wyniki, np. pomaga usuwanie rzadkich słów.

Outline

Dane

- Zebranie 404 fraz o metaforycznym znaczeniu (zawierają 202 przymiotniki).
- Klasyfikacja przymiotników na 11 grup (np. zmysły, wymiar, kolor, . . .) posiadających 47 podgrup (odpowiednio: słuch, głębokość, biały).
- Klasyfikacja rzeczowników na: abstrakcyjne i konkretne.
- Przygotowanie listy fraz przymiotnik-rzeczownik z NKJP zawierających wybrane przymiotniki.
- Uzupełnienie danych spośród najczęściej występujących.
- Weryfikacja danych przez 2 anotatorki.

Przykłady fraz

fraza	przymiotnik	rzeczownik	typ
głęboki wstyd	wymiar -> głębokość	abstrakcyjny	M
głęboka analiza	wymiar -> głębokość	abstrakcyjny	M
głęboka bruzda	wymiar -> głębokość	konkretny	L
głęboka cisza	wymiar -> głębokość	abstrakcyjny	M
głęboka dziupla	wymiar -> głębokość	konkretny	L
głęboka jaskinia	wymiar -> głębokość	konkretny	L
głęboka myśl	wymiar -> głębokość	abstrakcyjny	M
głęboka noc	wymiar -> głębokość	konkretny	M
głęboka wiedza	wymiar -> głębokość	abstrakcyjny	M
głęboka woda	wymiar -> głębokość	konkretny	B
głęboka zaspa	wymiar -> głębokość	konkretny	L
głęboka zmarszczka	wymiar -> głębokość	konkretny	L
głęboki podział	wymiar -> głębokość	abstrakcyjny	M
głęboki rów	wymiar -> głębokość	konkretny	L
głęboki skłon	wymiar -> głębokość	konkretny	L
głęboki śnieg	wymiar -> głębokość	konkretny	L

Statystyka danych

typ frazy	liczba		
	L	M	B
wszystkie frazy	990	449	175
zmysły	20	51	8
wymiar	130	33	15
kolor	183	29	24
właściwość fizyczna	90	29	20
temperatura	49	25	8
materiał	77	16	8
czystość/brud	22	25	6
emocje	26	22	9
światło/ciemność	43	25	12
zwierzęta	24	26	13
społeczeństwo	22	22	7

Outline

Reprezentacja fraz AN

M. Baroni i R. Zamparelli, *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space*, Proceedings of EMNLP 2010

$$\mathbf{A}_{(a)}\ \mathbf{n} = \mathbf{p}$$

gdzie:

- **p** wektor reprezentujący frazę AN
- **n** wektor reprezentujący rzeczownik z rozważanej frazy
- $\mathbf{A}_{(a)}$ macierz reprezentująca przekształcenie (zależna od przymiotnika z rozważanej frazy)

Rozpoznawanie fraz M i L — podejście podstawowe

D. Gutierrez, E. Shutova, T. Marghetis, B. Bergen, *Literal and metaphorical senses in compositional distributional semantic models*, Proceedings of ACL 2016

Podobne jak przy wieloznaczności rozróżniamy:

$$\mathbf{A}_{LIT(a)}\ \mathbf{n}_i = \mathbf{p}_i$$

$$\mathbf{A}_{MET(a)}\ \mathbf{n}_i = \mathbf{p}_i$$

Cel:

nauczenie się macierzy $\hat{\mathbf{A}}_{LIT(a)}$ na podstawie fraz literalnych i $\hat{\mathbf{A}}_{MET(a)}$ na podstawie fraz metaforycznych

Rozpoznawanie fraz M i L — podejście podstawowe, cd

Dla nowej frazy p_i o wektorze $\mathbf{p}_i$ sprawdzamy
Jeśli:

$$\cos(\mathbf{p}_i, \hat{\mathbf{A}}_{MET(a)}\ \mathbf{n}_i) > \cos(\mathbf{p}_i, \hat{\mathbf{A}}_{LIT(a)}\ \mathbf{n}_i)$$

to jest metaforyczna wpp literalna.
Dane:

- 3991 literalnych,
- 4601 metaforycznych,
- 23 przymiotników.

Wyniki z artykułu dla 10-krotnej walidacji krzyżowej:

F-score 0.817, recall 0.793, precision 0.842, accuracy 0.809.

Wyniki metody dla polskich danych, przymiotniki

przym.	dziedzina	nb.L	nb.M	F L	F M	F L reg	F M reg
biały	kolor	38	8	**0.900**	0.064	0.897	**0.178**
ciepły	temperatura	20	9	0.851	0.505	**0.879**	**0.556**
czarny	kolor	45	9	**0.943**	**0.755**	0.531	0.386
głęboki	wymiar	8	9	0.655	0.446	**0.712**	**0.659**
gorący	temperatura	10	9	0.715	0.193	**0.764**	**0.466**
gruby	wymiar	19	4	**0.893**	**0.588**	0.752	0.328
pełny	pełne/puste	7	14	0.690	**0.852**	**0.774**	0.842
wysoki	wymiar	22	9	0.837	0.550	**0.871**	**0.657**

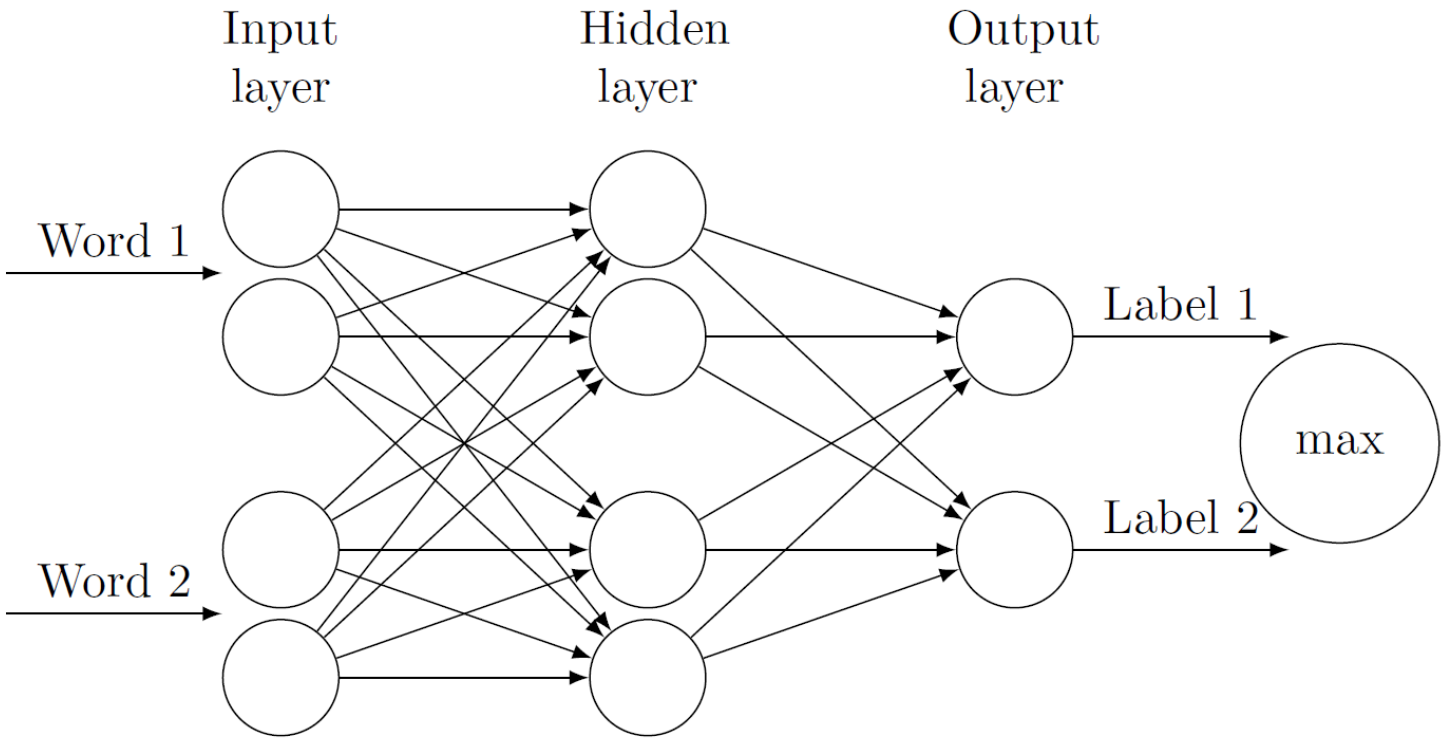
Wyniki dla polskich danych, dziedziny

dziedzina	nb. L	nb. M	F L	F M	F L reg	F M reg
zwierzęta	24	26	0.671	0.175	**0.710**	**0.414**
czystość	20	27	**0.786**	**0.829**	0.761	0.804
kolor	176	28	0.811	0.416	**0.944**	**0.664**
wymiar	128	33	0.866	0.409	**0.905**	**0.470**
emocje	26	21	0.647	0.374	**0.716**	**0.420**
pełny/pusty	22	16	0.780	0.787	**0.906**	**0.887**
światło/ciemność	39	25	**0.819**	**0.484**	0.773	0.230
właściwosci fizyczne	82	26	0.840	**0.422**	**0.864**	0.393
zmysły	18	42	**0.526**	**0.474**	0.524	0.441
społeczeństwo	20	22	0.683	0.563	**0.802**	**0.771**
materiał	72	15	**0.908**	**0.293**	0.907	0.247
temperatura	43	25	**0.804**	**0.391**	0.799	0.305

Rozwiązanie oparte na wektorach słów

A. Wawer, A. Mykowiecka. *Detecting metaphorical phrases in the Polish language.* Proceedings of RANLP, wrzesień 2017

Architektura sieci:



Wyniki

		2 dense layers			3 dense layers		
	nb	P	R	F1	P	R	F1
metaforyczne	453	0.86	0.90	0.90	0.85	0.81	0.83
literalne	988	0.92	0.94	0.93	0.92	0.93	0.92
średnia ważona	1441	0.90	0.90	0.90	0.90	0.90	0.90

Dodanie informacji o dziedzinach przymiotników nie pomaga:

	nb	P	R	F1
metaforyczne	453	0.86	0.79	0.83
literalne	988	0.91	0.94	0.92
średnia ważona	1441	0.89	0.90	0.89

Wyniki metody dla angielskiego, rozpoznawanie M/L

	dane Gutierrez et al. 2016			
	nb	P	R	F1
metaphors	4596	0.966	0.973	0.969
non metaphors	3991	0.969	0.961	0.965
średnia ważona	8587	0.967	0.965	0.966

Podobne rozwiązanie (też wrzesień 2017):

Y. Bizzoni, S. Chatzikyriakidis, M. Ghanimifard, *"Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs* Proceedings of the Workshop on Stylistic Variation, 2017

dla danych Gutierrez et al. raportują wyniki:

- trenowane na 500 frazach accuracy: 91,5%
- trenowane na 8000 frazach accuracy: 98,5%

Wyniki oryginalne: P 0.842, R 0.793, F1 0.817.

Dane Tsvetkov et al., podział na M/L

Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, C. Dyer
Metaphor Detection with Cross-Lingual Model Transfer
Proceedings of ACL, 2014

- Klasyfikator (lasy losowe)
- Zrównoważony zbiór 1753 fraz
- Wykorzystywane informacje:
  - supersensy z WordNetu dla N i V, własne dla Adj;
  - ocena abstrakcyjności i możliwości wizualizacji, np. *zemsta* jest trudniejsza do wizualizacji niż *tortury*;
  - wektory dystrybucyjne

Acc: 0.86

Dane Tsvetkov et al., podział na M/L

M. Rei, L. Bulat, D. Kiela, E. Shutova, *Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection* Proceedings of EMNLP, 2017 (też wrzesień 2017)

- modelowanie interakcji między reprezentacjami składowych frazy
- Wynik z artykułu: Acc: 82,9, P: 90,3, R: 73,8, F1: 81,1

dane Tsvetkov et al., nasze wyniki
(15epok. 10-krotne sprawdzenie krzyżowe)

	nb	P	R	F1
metaforyczne	882	0.88	0.84	0.86
literalne	871	0.85	0.88	0.86
średnia	1753	0.86	0.86	0.86

Rozpoznawanie fraz L, M i B

Rozpoznawanie fraz L, M i B, wektory 100 elementowe

		2 dense layers						3 dense layers		
		only vectors			vectors+adjective domains					
	nb	P	R	F1	P	R	F1	P	R	F1
metaforyczne	453	0.78	0.79	0.78	0.85	0.75	0.80	0.80	0.75	0.78
literalne	988	0.85	0.86	0.86	0.84	0.89	0.87	0.85	0.89	0.87
oba znaczenia	170	0.36	0.33	0.34	0.40	0.38	0.39	0.44	0.41	0.43
średnia ważona	1611	0.78	0.78	0.78	0.80	0.80	0.80	0.80	0.80	0.80

Rozpoznawanie fraz L, M i B, wektory 300 elementowe

		2 dense layers		
	nb	P	R	F1
metaforyczne	453	0.82	0.76	0.79
literalne	988	0.85	0.87	0.86
oba znaczenia	170	0.39	0.39	0.39
średnia ważona	1611	0.79	0.79	0.79

Outline

Dane

Z NKJP wybrano 1817 zdań zawierających frazy typu "B", dla których ustalono typ użycia

literalny: 545;

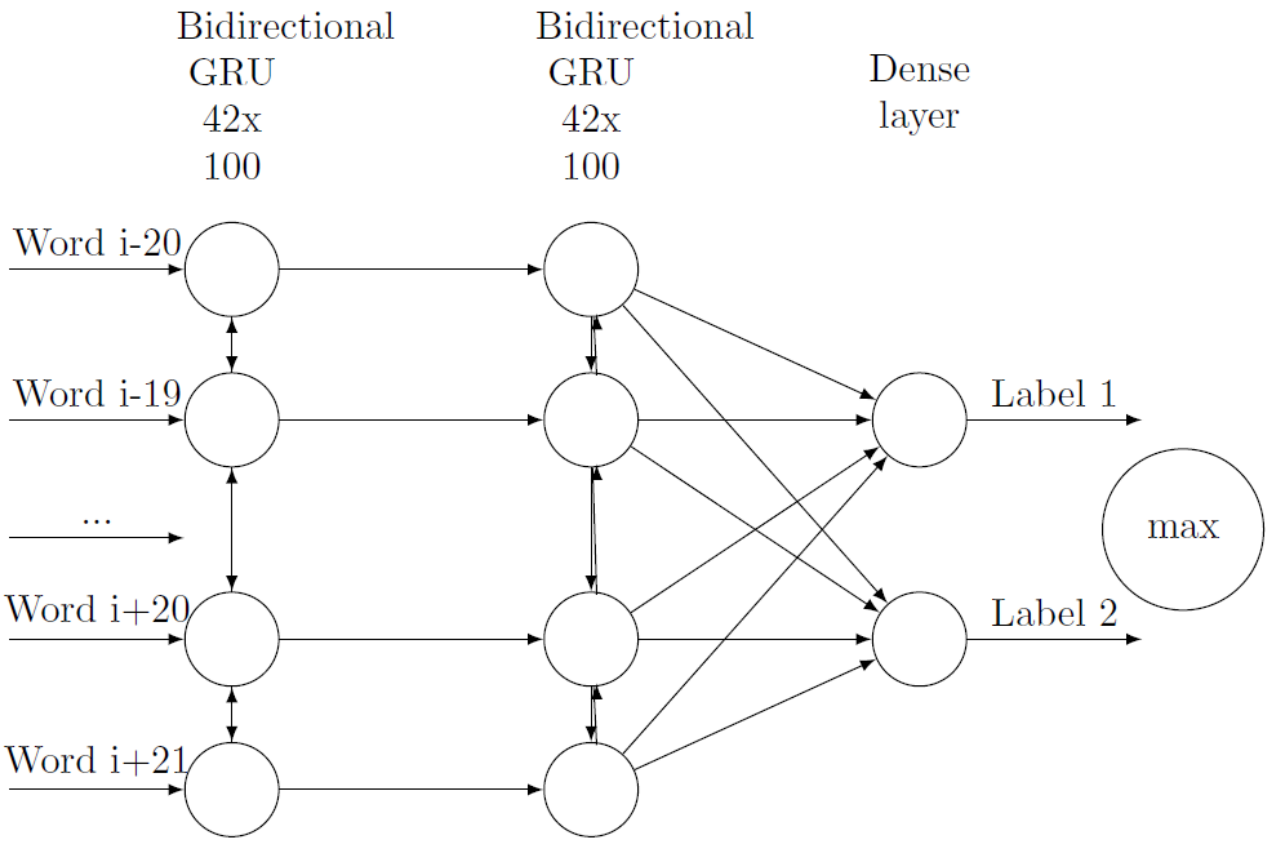
metaforyczny: 1272.

Przykłady:

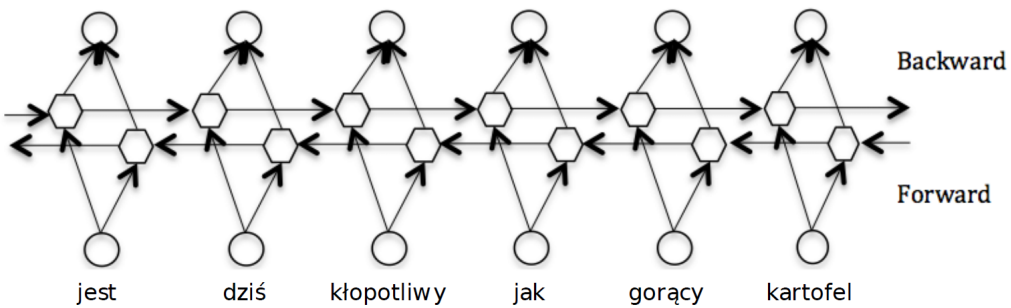
literalne: *Na szczęście jacht znajdował się już w bezpiecznej przystani.*

metaforyczne: *[..] pałac był bez gospodarza, zdany na kaprysy aury, aktywność szabrowników i różnego rodzaju marginesu społecznego, który znalazł sobie w murach bezpieczną przystań.*

Architektura sieci



architektura warstwy GRU:



Wyniki

kontekst 20 słów przed i 20 po frazie (20 epok)

		2 warstwy GRU			3 warstwy GRU		
	nb	P	R	F1	P	R	F1
metaforyczne	1272	0.85	0.87	0.86	0.83	0.89	0.86
literalne	545	0.68	0.63	0.66	0.69	0.58	0.63
średnia ważona	1817	0.80	0.80	0.80	0.79	0.80	0.79

po zmniejszeniu kontekstu do 10 słów, 2 warstwy GRU

	nb	P	R	F1	
metaforyczne	1272	0.86	0.84	0.85	
literalne	545	0.65	0.69	0.67	
średnia ważona	1817	0.8	0.8	0.8	