

Ekstrakcja terminologii z tekstów w języku polskim — program TermoPL

Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik

Seminarium IPI PAN, 11 stycznia 2016

Zadanie

Cel ekstrakcji terminologii:

wydobycie specyficznej terminologii z tekstów dotyczących wybranej dziedziny.

Zastosowania:

- tworzenie słowników dziedzinowych;
- tworzenie zasobów do tłumaczenia tekstów;
- wstępny krok przy opracowywaniu ontologii;
- anotacja dokumentów i wspomaganie wyszukiwania odpowiedzi na pytania;
- przydatne przy streszczaniu dokumentów;
- ...

Przykład

Plik Edycja Widok Historia Zakładki Narzędzia Pomoc

CLARIN2015 Submission 6

<https://easychair.org/conferences/submission.cgi?submission=2427016;a=9E>

Często odwiedzane Bezplatna usługa poc... Dostosuj łącz Windows Media Windows

For all questions related to processing your submission you should contact the conference organizers. [Click here to see information about this conference.](#)
 All **reviews sent to you** can be found at the bottom of this page.

Paper 6

Title:	TermoPL tool for automatic extraction of Polish terminology
Submission	
Author keywords:	terminology extraction Polish C-value
EasyChair keyphrases:	term candidate (70), automatic terminology extraction (63), base form (60), noun phrase (60), nap gen (50), terminology extraction (50), narodowy korpus j ezyka (40), orthographic form (40), simplified form (40)
Abstract:	The paper presents the TermoPL tool created to extract terminology from domain corpora in Polish. The program extracts term candidates with help of a simple grammar recognizing noun phrases and uses the C-value method to rank them. The program accepts as input morphologically annotated and disambiguated domain texts and creates a list of terms, the top part of which comprise domain terminology.
Time:	Jul 13, 10:44 GMT

Realizacja zadania

- Zgromadzenie tekstów dziedzinowych.
- Wstępna analiza lingwistyczna — tagowanie (przypisanie formy podstawowej, części mowy oraz charakterystyki morfologicznej).
- Identyfikacja fraz — kandydatów na terminy.
- Szeregowanie fraz.
- Selekcja fraz.

Co rozumiemy pod pojęciem terminu?

Definicja słownikowa

Wyraz albo połączenie wyrazowe o specjalnym, konwencjonalnie ustalonym znaczeniu naukowym lub technicznym. (Doroszewski)

Definicja robocza

Fraza rzeczownikowa, która w tekstach dziedzinowych występuje dostatecznie często by przypuszczać, że opisuje pojęcie istotne dla dziedziny. Występowanie tej frazy w tekstach spoza dziedziny jest rzadsze.

Struktura gramatyczna terminów w języku polskim

- rzeczownik, akronim lub skrót rzeczownika:
 - *podatek, angiografia,*
 - *PKB, USG*
 - *ust.(awa),*
- rzeczownik z przymiotnikiem (który wystąpił po lub rzadziej przed rzeczownikiem):
 - *stosunki gospodarcze,*
 - *granulocyty obojętnochłonne;*
- sekwencja rzeczownika z rzeczownikiem w dopełniaczu:
 - *udar_{n,nom} mózgu_{n,gen};*
 - *kodeks_{n,nom} pracy_{n,gen};*
- kombinacja powyższych dwóch struktur:
 - *europejski_{adj} rynek_{n,nom} usług_{n,gen} finansowych_{adj},*
 - *wodonercze niewielkiego stopnia dolnego układu podwójnego nerki prawej;*

Struktura gramatyczna terminów w języku polskim

- fraza rzeczownikowa modyfikowana frazą przyimkową:
 - *wierzytelność podatnika wobec skarbu państwa,*
 - *podatek dochodowy od osoby fizycznej;*
 - *poziom hormonów we krwi;*
- można uwzględnić koordynację:
 - *bezsportna i wymagalna wierzytelność podatnika wobec skarbu państwa,*
 - *zapalenie mózgu i rdzenia,*
 - *oddział alergologii, endokrynologii i pediatrii ogólnej.*

Wykluczenie niektórych słów/fraz

Terminy nie powinny składać się ze:

- słów wskazujących na określenie czasu, jak np: *miesiąc, dzień*;
- nazwy dni i miesięcy, np: *styczeń, poniedziałek*;
- przymiotników wymagających kontekstu do interpretacji np: *inny, niektóry, jakiś, pewien*.

Należy wykluczyć przyimki złożone:

- *[w kierunku]* zapalenia nerek → *kierunek zapalenia nerek*;
- *[pod postacią]* podatku VAT → *postać podatku VAT*;
- *[pod kątem]* diagnostyki obrazowej → *kąt diagnostyki obrazowej*;
- *[pod kątem]* prostym → *kąt prosty*.

Gramatyka

NPP : \$*NAP* *NAP_GEN**;

NAP[*agreement*] : *AP** *N* *AP**;

NAP_GEN[*case* = *gen*] : *NAP*;

AP : *ADJ* | *ADJA* *DASH* *ADJ* | *PPAS*;

N[*pos* = *subst*, *ger*];

ADJ[*pos* = *adj*];

ADJA[*pos* = *adja*];

PPAS[*pos* = *ppas*];

DASH[*form* = " - "];

Szeregowanie terminów

Dla każdej frazy kandydackiej p liczona jest wartość C-value:

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

p — rozważana fraza,

LP — zbiór fraz zawierających p ,

$r(LP)$ — liczba różnych fraz w LP ,

$I(p) = \log_2(length(p))$, jeśli p ma długość 1 to przyjmujemy stałą n_p : $I(p) = 0.1$;

referencja bibliograficzna

Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. Journal on Digital Libraries 3 (2000) 115–130

Identyfikacja fraz

	pojedyncza	mnoga
nom	<i>przewlekły nieżyt żołądka</i>	<i>przewlekłe nieżyty żołądka</i>
gen	<i>przewlekłego nieżytu żołądka</i>	<i>przewlekłych nieżytów żołądka</i>
dat	<i>przewlekłemu nieżyтови żołądka</i>	<i>przewlekłym nieżytom żołądka</i>
acc	<i>przewlekły nieżyt żołądka</i>	<i>przewlekłe nieżyty żołądka</i>
inst	<i>przewlekłym nieżytem żołądka</i>	<i>przewlekłymi nieżytami żołądka</i>
loc	<i>przewlekłym nieżycie żołądka</i>	<i>przewlekłych nieżytach żołądka</i>

Wykorzystujemy uproszczoną formę podstawową:

- *przewlekły nieżyt żołądka* → *przewlekły nieżyt żołądek*;
- *ostra niewydolność nerek* → *ostry niewydolność nerka*.

Problemy z uproszczoną formą podstawową

Taką samą uproszczoną formę podstawową mają:

- frazy w liczbie mnogiej i pojedynczej np. *zapalenie ucha* i *zapalenie uszu*, uproszczona: *zapalenie ucho*;
- przymiotniki w różnych stopniach (mały, mniejszy) np. *miednica mała* (częściej *mała miednica* — opisuje rozmiar) podczas gdy *miednica mniejsza* (określenie anatomiczne), uproszczona: *miednica mały*;
- pozytywne i zanegowane imiesłowy przymiotnikowe . *powiększony*/*niepowiększony* mają formę podstawową *powiększyć_{inf}*;
- gerundia i imiesłowy mają bezokoliczniki jako formy podstawowe:
 - *usunięcie_{ger}* *kamienia_{subst:gen}* — operacja,
 - *usunięty_{ppas}* *kamień_{subst:nom}* — opis kamienia,forma uproszczona: *usunąć_{inf}* *kamień_{subst}*.

Konteksty

<i>planowa</i>	<i>operacja</i>	<i>przepukliny</i>	<i>pachwinowej</i>	<i>lewostronnej</i>
	<i>operacja</i>	<i>przepukliny</i>	<i>pachwinowej</i>	<i>lewostronnej</i>
<i>planowa</i>	<i>operacja</i>	<i>przepukliny</i>	<i>pachwinowej</i>	
	<i>operacja</i>	<i>przepukliny</i>	<i>pachwinowej</i>	
		<i>przepuklina</i>	<i>pachwinowa</i>	<i>lewostronna</i>
	<i>lewostronna</i>	<i>przepuklina</i>	<i>pachwinowa</i>	
		<i>przepuklina</i>	<i>pachwinowa</i>	<i>prawostronna</i>
		<i>przepuklina</i>	<i>pachwinowa</i>	<i>obustronna</i>
	<i>prawostronna</i>	<i>przepuklina</i>	<i>pachwinowa</i>	
	<i>uwięźnięta</i>	<i>przepuklina</i>	<i>pachwinowa</i>	<i>prawostronna</i>

Liczenie kontekstów

Metody liczenia kontekstów (ograniczamy do jednego słowa):

- 1 liczba różnych kontekstów liczona po obu stronach razem;
- 2 suma różnych kontekstów po obu stronach;
- 3 maksimum z kontekstów liczonych z lewej i prawj strony osobno.

Konteksty dla frazy: *przepuklina pachwinowa*:

- 1 'operacja'–'lewostronny', 'operacja'–[pusty], [pusty]–'lewostronny', 'lewostronny'–[pusty], [pusty]–'prawostronny', [pusty]–'obustronny', 'prawostronny'–[pusty], 'uwięźnięty'–'prawostronny';
- 2 'operacja', 'lewostronny', 'prawostronny', 'obustronny', 'uwięźnięty';
- 3 'operacja', 'lewostronny', 'prawostronny', 'uwięźnięty' (lewych o jeden więcej).

Ocena wyników

- Termin medyczny: *ostry niezbyt żołądka* (ok. 88 % górnej części listy terminów w testach opisanych w artykule *Terminology Extraction from Medical Texts in Polish*);
- Termin ogólny: *pora nocy, dół* (fragment frazy *dół biodrowy*);
- Terminy niepoprawne wynikające z:
 - niedostatków gramatyki: *dziewczynka skierowana* z frazy *dziewczynka skierowana do chirurga*;
 - błędów anotacji: *Lacidofil zalecenia* z dwóch całkowicie odrębnych fraz bez znaków przestankowych (*zalecenia* otagowane jako dopełniacz a nie mianownik);
 - urwanych fraz: *infekcja dróg, USG jamy*.

Problem uciętych fraz

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:

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

NPMI – Normalised Pointwise Mutual Information

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

Where:

- 'x y' jest bigramem składającym się z lematów tokenów x i y,
- $p(x,y)$ jest prawdopodobieństwem bigramu 'x y' w korpusie,
- $p(x)$, $p(y)$ jest prawdopodobieństwem unigramów 'x' i 'y' w korpusie .

referencja bibliograficzna

Gerlof Bouma, 2009, *Normalized (pointwise) mutual information in collocation extraction.*, w: *Proceedings of the Biennial GSCL Conference 2009*, strony 31—40.

Algorytm ustalający frazy zagnieżdżone

candidate_term (phr)

if phr jest poprawną frazą rzeczownikową to
 dodaj phr do listy terminów

if length(phr) > 1

 znajdź wszystkie pozycje i , w których
 phr można podzielić zgodnie z gramatyką

for wszystkich pozycji i

 wylicz NPMI(i -tego bigramu we phr)

 posortuj NPMI od najmniejszej do największej wartości

$j :=$ miejsce z najniższą NPMI

 podziel phr na phr1 i phr2 w j -tym miejscu

candidate_term(phr1)

candidate_term(phr2)

Przykład

infekcja górnych dróg oddechowych

Noun; Adj; Noun; Adj;

infekcja | górnych dróg | oddechowych

infekcja górny droga oddechowy

bigram	NPMI
infekcja górny	0.65658
górny droga	0.78773
droga oddechowy	0.95089

Porównanie dwóch metod

Poprawne gramatycznie podfrazy

<i>'infekcja'</i>	<i>'górnny'</i>	<i>'droga'</i>	<i>'oddechowy'</i>
<i>infekcja</i>	<i>górnnych dróg</i>	<i>oddechowych</i>	
<i>infekcja</i>	<i>górnnych dróg</i>		
<i>infekcja</i>			
	<i>górne</i>	<i>drogi</i>	<i>oddechowe</i>
	<i>górne</i>	<i>drogi</i>	
		<i>drogi</i>	<i>oddechowe</i>
		<i>drogi</i>	

Podfrazy z wykorzystaniem NPMI

<i>'infekcja'</i>	<i>'górnny'</i>	<i>'droga'</i>	<i>'oddechowy'</i>
<i>infekcja</i>	<i>górnnych dróg</i>	<i>oddechowych</i>	
	<i>górne</i>	<i>drogi</i>	<i>oddechowe</i>
		<i>drogi</i>	<i>oddechowe</i>
		<i>drogi</i>	

Preferowanie podziału na dwie frazy rzeczownikowe

*prawidłowa*_{adj} *mikroflora*_{noun} *górných*_{adj} *dróg*_{noun} *oddechowych*_{adj}
—> *prawidłowa mikroflora* oraz *górne drogi oddechowe*

*częste*_{adj} *infekcje*_{noun} *górných*_{adj} *dróg*_{noun} *oddechowych*_{adj} —>
częste modyfikuje całą frazę *infekcje górnych dróg oddechowych*

Modyfikacja:

- szukamy najśłabszej pozycji pozwalającej podzielić frazę na dwie podfrazy rzeczownikowe;
- jeśli różnica pomnięcy nastłabszym miejscem podziału a tym dzielącym na dwie frazy rzeczownikowe jest mniejsza od ustalonego progu to preferujemy podział na dwie frazy rzeczownikowe.

Ewaluacja 2000 fraz

	Poprawne		Przyczyna niepoprawności		
	medyczne	ogólne	gramatyka	anotacja	obcięte
s-phrases	1,778	84	48	25	65
— wyłącznie	174	7	8	3	49
s&npmi-phrases	1,823	85	48	27	17
— wyłącznie	219	8	8	5	1
Wspólne	1,604	77	40	22	16

Fraza *operacja przeszczepienia* przesunęła się o 300 pozycji do góry w metodzie z wykorzystaniem NPMI

Selekcja fraz na podstawie korpusu ogólnego

	jedno.	wielo.	razem
Wszystkie	2,113	464	2,577
C-value większe w NKJP	1,319	193	1,512
$C-v_{med} > 3.0 \ \& \ C-v_{NKJP} > C-v_{med}$	96	11	107
2K medyczne $\& \ C-v_{NKJP} > C-v_{med}$	16	0	16

Najdłuższa wspólna fraza:

objawy infekcji górnych dróg oddechowych

Wspólne frazy

Wielowyrazowe frazy z C-value powyżej 3.0 w medycznych danych i wyższym C-value w NKJP

fraz	Korpus medyczny		NKJP	
	C-value	pozycja	C-value	position
<i>duży stopień</i>	9.25	2,817	16.00	479
<i>jedna strona</i>	4.00	5,266	36.00	131
<i>członek rodzina</i>	4.00	5,509	10.00	963
<i>intensywna terapia</i>	3.00	6,674	4.00	3,260
<i>pani doktor</i>	3.00	6,750	6.50	1,674
<i>jedna noc</i>	3.00	6,750	5.00	1,674
<i>pierwszy etap</i>	3.00	7,051	8.00	1,281
<i>lewa noga</i>	3.00	7,092	5.00	2,472
<i>podjąć decyzję</i>	3.00	7,215	8.00	1,295
<i>własna prośba</i>	3.00	7,238	5.00	2,505
<i>dom dziecka</i>	3.00	7,252	6.00	1,885

Publikacje

- Marciniak, M. i Mykowiecka, A. *Construction of a Medical Corpus Based on Information Extraction Results*. Control & Cybernetics, 40(2), 337—360, (2011)
- Marciniak, M. and Mykowiecka, A. *Terminology Extraction from Medical Texts in Polish*. Journal of Biomedical Semantics, 5. (2014)
- Marciniak, M. and Mykowiecka, A. *Nested Term Recognition Driven by Word Connection Strength*. Terminology, 21(2), 180–204, (2015)
- Marciniak M. *Domain corpora as a source of information* Monograph Series, volume 4, Institute of Computer Sciences PAS

Program

Opracowany w ramach projektu Clarin.PI

- Java Runtime Environment w wersji 7 lub nowszej;
- Wymaga Morfeusza 2 do wygenerowania formy podstawowej z uproszczonej formy;
- Wymaga otagowanego i ujednoznaczonego korpusu danych w jednym z formatów:
 - NKJP;
 - XCES;
 - zapis uproszczony: token # lemat # tag.
- na wyjściu: lista uporządkowanych terminów (w uproszczonych formach lub zrekonstruowanych formach podstawowych wraz z formami znalezionych fraz).

Prezentacja