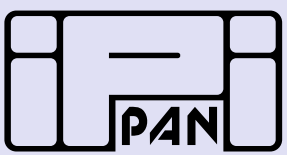
Ekstrakcja terminologii — TermoPL

Małgorzata Marciniak, Agnieszka Mykowiecka,
Piotr Rychlik



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

Warsztaty CLARIN-PL, Poznań 12-13 kwietnia 2018

Outline

Zadanie

Cel ekstrakcji terminologii:

wydobycie specyficznej terminologii z tekstów dotyczących wybranej dziedziny.

Co otrzymujemy:

- listę terminów w formie podstawowej lub w uproszczonej formie podstawowej;
- listę możemy sortować według:
  - adekwatności frazy jako terminu;
  - częstości w tekstach;
  - długości frazy
- możliwość wyboru fraz wielowyrazowyvh;
- możliwość zebrania wystąpień fraz w korpusie wraz z częstościami;
- możliwość porównań dwóch zbiorów terminów.

Zastosowania

- Tworzenie słowników dziedzinowych;
- Tworzenie zasobów do tłumaczenia tekstów;
- Automatyczna anotacja dokumentów;
- Wspomaganie wyszukiwania odpowiedzi na pytania;
- Przydatne przy streszczaniu dokumentów;
- Opracowywanie ontologii dziedziny;
- Tworzenia indeksów;
- ...

jama ustny [jama ustna] 241

śluzówka jama ustny [śluzówka jamy ustnej] 79
jama ustny czysty [jama ustna czysta] 4
suchość jama ustny [suchość jamy ustnej] 4
błona śluzowy jama ustny [błona śluzowa jamy ustnej] 3
grzybica śluzówka jama ustny [Grzybica śluzówek jamy ustnej] 2
pielęgnacja jama ustny [pielęgnacji jamy ustnej] 2
płukanie jama ustny [Płukanie jamy ustnej] 2
grzybica jama ustny [Grzybica jamy ustnej] 2
pędzlować jama ustny [pędzlowanie jamy ustnej] 2
sanacja jama ustny [sanacji jamy ustnej] 1
silny grzybica jama ustny [silna grzybica jamy ustnej] 1
jama ustny pleśniawka [jama ustnej pleśniawki] 1
zapalenie jama ustny [zapalenie jamy ustnej] 1
dno jama ustny [dno jamy ustnej] 1
jama ustny prawidłowy [jama ustna prawidłowa] 1
toaleta jama ustny [toaleta jamy ustnej] 1
jama ustny afta [jama ustnej afty] 1
higiena jama ustny [higiena jamy ustnej] 1

Choroby jamy ustnej

grzybica śluzówka jama ustny [Grzybica śluzówek jamy ustnej] 2
grzybica jama ustny [Grzybica jamy ustnej] 2
silny grzybica jama ustny [silna grzybica jamy ustnej] 1
jama ustny pleśniawka [jama ustnej pleśniawki] 1
zapalenie jama ustny [zapalenie jamy ustnej] 1
jama ustny afta [jama ustnej afty] 1

Realizacja zadania

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

Program Termo.PL

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 ujednoznacznionego korpusu danych w jednym z formatów:
  - NKJP;
  - XCES;
  - XML-owe wyjście z Korpusomatu (src/morph.xml);
  - zapis uproszczony: token # lemat # tag.
- na wyjściu: lista uporządkowanych terminów (w uproszczonych formach lub zrekonstruowanych formach podstawowych wraz z formami znalezionych fraz).

Przydatne adresy internetowe

- `http://ws.clarin-pl.eu/termopl.shtml`
- `http://zil.ipipan.waw.pl/TermoPL`

Outline

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. Częstość tej frazy w tekstach spoza dziedziny jest niższa.

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 uwzględnienić koordynację:
  - *bezsporna 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*.

Outline

Szeregowanie terminów

Bierzemy pod uwagę:

- częstość frazy;
- długość frazy;
- liczbę kontekstów.

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	*przewlekły nieżyt żołądka*	*przewlekłe nieżyty żołądka*
gen	*przewlekłego nieżytu żołądka*	*przewlekłych nieżytów żołądka*
dat	*przewlekłemu nieżytowi żołądka*	*przewlekłym nieżytom żołądka*
acc	*przewlekły nieżyt żołądka*	*przewlekłe nieżyty żołądka*
inst	*przewlekłym nieżytem żołądka*	*przewlekłymi nieżytami żołądka*
loc	*przewlekłym nieżycie żołądka*	*przewlekłych nieżytach żołądka*

Wykorzystujemy uproszczoną formę podstawową:

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

Problemy z uproszczoną formą podstawową

Taką sama 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 maja bezokoliczniki jako formy podstawowe:
  - $usunięcie_{ger}$ $kamienia_{subst:gen}$ — operacja,
  - $usunięty_{ppas}$ $kamień_{subst:nom}$ — opis kamienia,

  forma uproszczona: $usunąć_{inf}$ $kamień_{subst}$.

Konteksty

planowa	*operacja*	*przepukliny*	*pachwinowej*	*lewostronnej*
	operacja	*przepukliny*	*pachwinowej*	*lewostronnej*
planowa	*operacja*	*przepukliny*	*pachwinowej*	
	operacja	*przepukliny*	*pachwinowej*	
		przepuklina	*pachwinowa*	*lewostronna*
	lewostronna	*przepuklina*	*pachwinowa*	
		przepuklina	*pachwinowa*	*prawostronna*
		przepuklina	*pachwinowa*	*obustronna*
	prawostronna	*przepuklina*	*pachwinowa*	
	uwięźnięta	*przepuklina*	*pachwinowa*	*prawostronna*

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).

Outline

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 wykorzystujemy do oceny siły powiązania pomiędzy słowami.

referencja bibliograficzna

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

Przykład

infekcja górnych dróg oddechowych

Noun_j Adj_i Noun_i Adj_i

infekcja | górnych dróg | oddechowych

infekcja górny droga oddechowy

bigram	NPMI
infekcja górny	0.66
górny droga	0.79
droga oddechowy	0.95

Porównanie dwóch metod

Poprawne gramatycznie podfrazy				Podfrazy z wykorzystaniem NPMI			
‘*infekcja*’	‘*górny*’	‘*droga*’	‘*oddechowy*’	‘*infekcja*’	‘*górny*’	‘*droga*’	‘*oddechowy*’
infekcja	*górnych*	*dróg*	*oddechowych*	*infekcja*	*górnych*	*dróg*	*oddechowych*
infekcja	*górnych*	*dróg*		—			
infekcja				*infekcja*			
	górne	*drogi*	*oddechowe*		*górne*	*drogi*	*oddechowe*
	górne	*drogi*		—			
		drogi	*oddechowe*			*drogi*	*oddechowe*
		drogi				*drogi*	

Preferowanie podziału na dwie frazy rzeczownikowe

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

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

Modyfikacja:

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

Outline

Analiza kontrastywna terminologii

Cel

Na podstawie porównania wyników ekstrakcji terminologii dla dwóch korpusów mają być wskazane frazy:

- bardziej specyficzne dla innej dziedziny (porównanie z terminologią wydobytą z innego korpusu dziedzinowego)
- terminy ogólne np.” *własny sposób*, *lewa strona*, *trudne zadanie* (porównanie z korpusem języka ogólnego).

Metody porównań

Zaimplementowane metody wykorzystują:

- Log-Likelihood (LL – logarytm wiarygodności): na ile różni się częstość konkretnego terminu w dwóch porównywanych korpusach;
- Term Frequency Inverse Term Frequency (TFITF): łączy częstość występowania w korpusie dziedzinowym z odwrotną częstością występowania w korpusie ogólnym (liczoną jako stosunek wielkości korpusu do częstości badanego terminu);
- Contrastive Selection of Multi-Word Terms(CSmw): dla terminów wielowyrazowych, uwzględnia zarówno częstość występowania pełnych terminów, ale też częstość występowania słów stanowiących element główny badanej frazy.