Ekstrakcja terminologii z tekstów w języku polskim — program TermoPL

Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik

Seminarium IPI PAN, 22 lutego 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;
- ...

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.

Wbudowana gramatyka terminów

- 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*;

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*.

Szeregowanie terminów

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

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

p — rozważana fraza,
LP — zbiór fraz zawierających p,
$r(LP)$ — liczba różnych fraz w LP,
$l(p) = log_2(length(p))$, jeśli p ma długość 1 to przyjmujemy stałą np: $l(p) = 0.1$;

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

Działamy na uproszczonych formach

	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*.

Gramatyczne podfrazy — problem

Gramatycznie poprawne zagnieżdżone frazy:

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

Przykłady frazy o silnym powiązaniu słów:

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

NPMI – Normalised Pointwise Mutual Information

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

Where:

- ‘x y’ jest bigramem składajacym się z lematów tokenów x i y,
- p(x,y) jest prawdopodobińśtwem 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.

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*	

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 ujednoznacznionego 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).

Termo.PL

Prezentacja