PolEval 2019

Task 2

Lemmatization of proper names and multi-word phrases

Zbigniew Bronk
(badacz niezależny)

Zadanie

- Zadanie wymaga przekształcenie fraz wejściowych do *postaci podstawowej* opisanej w *Wytycznych lematyzacji fraz rzeczownikowych i przymiotnikowych*
- *biskupom wrocławskim*
  - *biskup wrocławski*
- *Wyspach Brytyjskich*
  - ** Wyspa Brytyjska*
  - ***Wyspy Brytyjskie***
- *Prezydent Boliwii*
  - ** prezydent Boliwia*
  - ** Prezydent Boliwia*
  - *? **prezydent Boliwii***
  - *? Prezydent Boliwii*

Etapy mojego rozwiązania

1. Analiza morfologiczna (rozszerzona)
2. Eliminacja wybranych interpretacji
3. Ujednoznacznienie interpretacji analizy morfologicznej i analiza składniowa
4. Generowanie wariantów lematów
5. Wybór najczęściej spotykanego wariantu lematu na podstawie danych korpusowych

Analiza morfologiczna

- *Analizator morfologiczny Morfeusz* (wariant SGJP) z modułem dla Javy:
  - segmentacja – podział wejściowej frazy na segmenty,
  - analiza morfologiczna, której wynikiem są *interpretacje morfologiczne* poszczególnych segmentów.
- Moje rozszerzenie analizy:
  - uzupełnienie na podstawie danych *Multisłownika* interpretacji w wypadku form nierozpoznanych (tzw. „ignów") i niezgodnych wielkości liter,
  - w wypadku niezinterpretowanych słów zawierających znak łącznika (-) wykonywana jest resegmentacja i reinterpretacja składników (np. *Radwańska-Paryska*)
  - zgadywanie „przymiotnikowych" i „apostrofowych" interpretacji „ignów",
  - rozszerzenie interpretacji o dane frekwencyjne z anotowanego korpusu NKJP1.2 i dużego korpusu nieanotowanego,
  - w wypadku czasowników dodatkowa informacja o przechodniości i łączliwości z „się",
  - uporządkowanie listy interpretacji w kolejności od najczęściej do najrzadziej spotykanych.

Interpretacje Morfeusza

0	1	Ala	Ala	subst:sg:nom:f	imię	
			Al	subst:sg:gen.acc:m1	imię	
			Alo	subst:sg:gen.acc:m1	imię	
1	2	ma	mój:a	adj:sg:nom.voc:f:pos		
			mieć	fin:sg:ter:imperf		
2	3	kota	kota	subst:sg:nom:f	nazwa pospolita	
			kot:s1	subst:sg:gen.acc:m2	nazwa pospolita	
			kot:s2	subst:sg:gen.acc:m1	nazwa pospolita	pot., środ.

A po rozszerzeniu…

Gramatyka polska

Rozszerzenie analizatora morfologicznego Morfeusz SGJP

Fraza: Ala ma kota

Wpisz frazę, który ma być analizowana.

Analizuj

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
0	1	Ala	Ala	**Ala**	**subst:sg:nom:f**	imię		7	Wikisłownik, SJP.PL, PWN (SJP+SO+SJPDor.), Sł. nazwisk	SGJP:1.0	95%
				Al	subst:sg:gen:m1			0	SJP.PL, PWN (SJP+SO+SJPDor.), Sł. nazwisk		2%
					subst:sg:acc:m1						1%
				Alo	subst:sg:gen:m1						
					subst:sg:acc:m1						0%
1	2	ma	ma	**mieć**	**fin:sg:ter:imperf**			2161	Wikisłownik, WSJP, SJP.PL, PWN (SJP+SO+SJPDor.), SJPDor.		100%
				mój:a	adj:sg:nom:f:pos			0	Wikisłownik, SJP.PL, PWN (SJP+SO+SJPDor.), SJPDor.		0%
					adj:sg:voc:f:pos						
2	3	kota	kota	**kot:s1**	**subst:sg:gen:m2**	nazwa pospolita		23	Wikisłownik, WSJP, SJP.PL, PWN (SJP+SO+SJPDor.), SJPDor.		41%
					subst:sg:acc:m2						38%
				kot:s2	subst:sg:gen:m1		pot. środ.	1			10%
				kota	subst:sg:nom:f			0	Wikisłownik, SJP.PL, PWN (SJP+SO+SJPDor.), SJPDor.		6%
				kot:s2	subst:sg:acc:m1		pot. środ.	1	Wikisłownik, WSJP, SJP.PL, PWN (SJP+SO+SJPDor.), SJPDor.		5%

Objaśnienia do kolumn pojawiają się po wskazaniu kursorem numeru kolumny w nagłówku.

Segmentacja i podstawowa analiza morfologiczna: Analizator morfologiczny Morfeusz
Dane frekwencyjne: Narodowy Korpus Języka Polskiego (1.2M)

Opracowanie: Zbigniew Bronk. Kontakt: Z[kropka]Bronk[at]ai[kropka]pl

Wyniki analizy morfologicznej

Ogółem fraz	1997	
Nierozpoznane przez Morfeusza (ign) lub niezgodne co do wielkości liter	620	
Występujące w Multisłowniku	163	*Walt, Disney, Bogota, Kordoba, Hestia, Kartuzy(?!), Bukowina (!)*
Odgadnięte "przymiotniki"	15	*Cyplu Szluknowskim, Przełęczy Jugowskiej, Rówieńkowego Potoku, Łabajowej, Piotra Xięskiego*
Odgadnięte "formy z apostrofem"	2	*couchsurfing'u, Oetker'a*
Pozostało nierozpoznanych	440	*Apple Store, KONG VI+, Fryderyce von Reden Metują, Vaclavicach, Michalovicach,Jizerą, Jesenikach*

- Po

Eliminacja wybranych interpretacji

- interpretacje z lematem pisanym wielką literą w wypadku słów (orth) pisanych małą (np. dla *polski* interpretacja ~~*Polski*(subst:sg:gen)~~)
- interpretacje, w których wielkość wewnętrznych liter nie jest zgodna (np. z lematem ~~*ipod*~~, dla słowa *iPod*),
- rzeczownikowe interpretacje homonimiczne z interpretacjami przymiotnikowymi (np. ~~stary(subst)~~),
- rzeczownikowe interpretacje z lematem w liczbie pojedynczej, gdy występują też homonimiczne interpretacje plurale tantum (np. ~~*rodzic*~~*, rodzice;* ~~*tarapata*~~*, tarapaty*),
- homonimiczne formy deprecjatywne rzeczowników męskoosobowych,
- interpretacje wołaczowe,
- interpretacje bardzo rzadkie (o bardzo niskiej frekwencji).

Analiza składniowa

- Mój prototypowy parser quasi-zależnościowy (Java)
  - Na wejściu parser otrzymuje ciąg list wariantywnych interpretacji morfologicznych, a wynikiem jego działania jest ciąg pojedynczych interpretacji wyrazów frazy wejściowej uzupełnionych o *relacje zależnościowe* (atrybuty *head* i *deprel)* pomiędzy segmentami.

Id	Form	Lemma	Tag	Head	DepRel
1	Zespołu	zespół	subst:sg:gen:m3	0	root
2	Parków	park	subst:m3:pl:gen	1	nmod
3	Krajobrazowych	krajobrazowy	adj:pos:m3:pl:gen	2	amod
4	Województwa	województwo	subst:sg:gen:n:ncol	1	nmod
5	Małopolskiego	małopolski	adj:adjp:pos:n:sg:gen	4	amod

Id	Form	Lemma	Tag	Head	DepRel
1	Policjantów	policjant	subst:pl:gen:m1	0	root
2	w	w	prep:loc:nwok	1	zb:pmod
3	służbowych	służbowy	adj:pos:m3:pl:loc	4	amod
4	uniformach	uniform	subst:pl:loc:m3	2	zb:prep

Relacje zależnościowe

- Podstawa: standard *Universal Dependencies*

advmod	przymiotnik -> przysłówek	8
amod	rzeczownik -> przymiotnik	635
appos	rzeczownik -> rzeczownik, zw. zgody	118
cc	spójnik	4
conj	element dołączony szeregowo	4
nmod	rzeczownik -> rzeczownik, zw. rządu	167
obj	dopełnienie	2
punct	znak interpunkcyjny	182
zb:adja	przymiotnik -> f. przyprzymiotnikowa (np. kujawsko-pomorski)	3
zb:num	liczebnik -> rzeczownik	7
zb:pmod	rzeczownik -> fraza przyimkowa	54
zb:prep	przyimek -> rzeczownik	51

jednowyrazowych – 1066
niepowiązanych – 260

Błędy analizy

Id	Form	Lemma	Tag	Head	DepRel
1	Galla	Galla	subst:f:sg:nom	0	root
2	Anonima	anonim	subst:sg:gen:m1	1	nmod

Id	Form	Lemma	Tag	Head	DepRel
1	Matki	matka	subst:sg:gen:f	0	root
2	Natury	natura	subst:sg:gen:f	1	nmod

Id	Form	Lemma	Tag	Head	DepRel
1	Doliny	Dolina	subst:f:sg:gen	0	root
2	Białki	Białka	subst:f:sg:gen	1	appos

Generowanie lematów i ich wariantów

Zachowuje się wielkość pierwszej litery składników. W wypadku składników na pierwszej pozycji zdania, które mają pospolitą interpretację, bierze się pod uwagę oba warianty.	*Karpaczu Górnym -> Karpacz **G**órny* *(…) Jeziora Lednickiego -> **J**ezioro Lednickie* *(.) Kozica górska -> **K**ozica górska* lub ***k**ozica górska*
W wypadku elementu głównego zastępuje się go formą mianownikową, przy czym, jeśli forma tego elementu jest w liczbie mnogiej, a nie jest on typu *plurale tantum* ani nie jest on częścią nazwy własnej, ani nie jest elementem połączenia szeregowego rzeczowników, to obok wariantu mnogiego, tworzony jest wariant w liczbie pojedynczej (preferowany). W przypadku przymiotnikowego (posp.) elementu głównego, otrzymuje on formę rodzaju męskiego.	*tarapatach finansowych -> tarapat**y** finansowe* *Wyspy Brytyjskie -> Wysp**y** Brytyjskie* *pracowników i emerytów sektora publicznego -> pracowni**cy** i emery**ci** sektora publicznego* *ruinach zamkowych -> ruin**a** zamkowa* lub *ruin**y** zamkowe* *wczesnopiastowskiej -> wczesnopiastow**ski***
W wypadku składników dowiązanych relacją typu *amod,* otrzymują one formę z przypadkiem i liczbą zgodną z nadrzędnikiem.	*miejscowych górników -> miejscow**y** górnik*

Reguły lematyzacji c.d.

W wypadku składników dowiązanych relacją typu *nmod,* zachowują one swój przypadek i liczbę; w wypadku, gdy nadrzędnik zmienia liczbę z mnogiej na pojedynczą, to obok formy podrzędnika z liczbą mnogą, bierze się też pod uwagę formę z liczba pojedynczą.	*miłośniczką tłumów -> miłośniczka tłumów* *łóżek pacjentów -> łóżko pacjent**ów*** lub *łóżko pacjent**a***
W wypadku składników dowiązanych relacją typu *zb:num,* jeśli nadrzędny liczebnik ma kategorię rządu z wartością *rec*, otrzymują one przypadek dopełniacza, a jeśli kategoria ta ma wartość *congr* (związek zgody), to przypadek jest uzgadniany z liczebnikiem (mianownik).	*pięcioma gwiazdkami -> pięć gwiazd**ek*** *trzech poniemieckich schronisk -> trzy poniemieckie schronisk**a***
W wypadku składników dowiązanych relacją typu *zb:prep,* przypadek jest zachowany, a jeśli nadrzędnik frazy przyimkowej zmienił liczbę, to obok formy podrzędnika z liczbą mnogą, bierze się też pod uwagę wariant z liczba pojedynczą.	*stolikami barowymi pod wodą -> stolik barowy pod wodą* *policjanci w służbowych uniformach -> policjant w służbowych uniform**ach*** lub *policjant w służbowym uniform**ie***
Pozostałe składniki, w tym składniki niedowiązane (np. nierozpoznane), nie podlegają przekształceniom lematyzacyjnym	*Górną Beczwa(ign) -> Górna Beczwa* *Piotrkowską 282 -> Piotrkowska 282* *Apple(ign) Store(ign) -> Apple Store* *Czyżynką(ign) -> Czyżynk**ą*** *Rožnovie(ign) pod Radhoštěm(ign) -> Rožnov**ie** pod Radhoštěm*

Wybór spośród wariantywnych lematów

- W wypadku, gdy proces lematyzacji zwrócił kilka wariantów możliwych lematów, wybranym lematem staje się ten, którego formy są najczęstsze w DKnA.
- Obok frekwencji formy podstawowej wariantu lematu, sprawdzana była też frekwencja formy dopełniacza.
- W przypadku fraz bardziej rozbudowanych, które nie mają wystąpień w DKnA, dodatkowo badana była frekwencja *lematów skróconych*.

(.) „Prezydent Boliwii”	
***P**rezydent Boliwii*	8
***p**rezydent Boliwii*	**30**

„ruinach zamkowych”	
ruina zamkowa	2
ruiny zamkowej	0
ruiny zamkowe	3
ruin zamkowych	26

(.) „Policjantów w służbowych uniformach”	
policjanci w służbowych uniformach	0
policjant w służbowych uniformach	0
policjant w służbowym uniformie	0
policjanta w służbowych uniformach	0
policjanta w służbowym uniformie	0
policjant w uniformach	0
policjant w uniformie	0
policjanta w uniformach	0
policjanta w uniformie	**2**

Wnioski

- Należy udoskonalić analizę składniową, zrezygnować z ujednoznaczniania na etapie analizy, a lematyzację przeprowadzać na zbiorze wariantywnych interpretacji morfoskładniowych.
- Lematyzowane frazy analizować w ich kontekście.
- Dopracować „zgadywacze form nieznanych”
- W szerszym zakresie rozstrzygać niejednoznaczności powiazań składniowych z wykorzystaniem KnA
- Zaindeksować wykorzystywany korpus tekstów (KnA), żeby sprawniej uzyskiwać dane o frekwencji poszczególnych wariantów zlematyzowanych fraz.

Pytania?

Dziękuję za uwagę!