

Zautomatyzowane tworzenie korpusów błędów dla języka polskiego

Marcin Miłkowski

Instytut Filozofii i Socjologii PAN
Zakład Logiki i Kognitywistyki

Adres projektu: *morfologik.blogspot.com*

Korpusy błędów

Zastosowania korpusów błędów

- ♦ Wykrywanie, co jest wykroczeniem poza normę językową
- ♦ Dane do opracowywania programów szkolnych i ćwiczeń
- ♦ Materiał do wydawnictw poprawnościowych
- ♦ Materiał do narzędzi językowych

Plan referatu

- (1) Motywacja teoretyczna
 - (2) Opis metody
 - (3) Uzyskane wyniki
 - (4) Korpus błędów a inne korpusy – albo „jak to zrobić w Poliqarpie?”
-
-

Pozyskiwanie korpusów

- Ręczne:
 - ♦ Anotowane przez lingwistów
 - ♦ Anotowane przez korektorów (biura tłumaczeń, redakcje gazet)
 - Ręczne, nieformalne
 - ♦ Zbierane przez entuzjastów lub dziennikarzy (Ibis)
 - ♦ Zbierane przez prasoznawców i językoznawców (*Słownik języka niypolskiego* Pisarka)
-
-

Pozyskiwanie korpusów

- Ręczne zbieranie jest kosztowne, pracochłonne; rodzi problemy z prawami autorskimi (np. biura tłumaczeń)
- Pracochłonność grozi niską reprezentatywnością („jak nie bądź”)

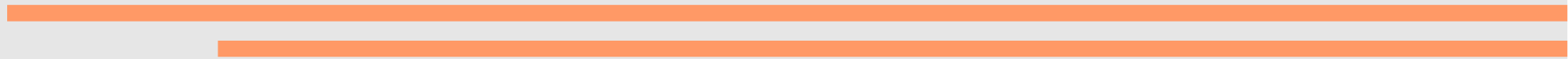
Pozyskiwanie korpusów

- Automatyczne:
 - Na podstawie istniejących dokumentów w dużych biurach tłumaczeń (tzw. LQI)
 - **Na podstawie historii poprawek dokumentu**
 - ➔ **Systemy kontroli wersji w firmach**
 - ➔ **Historia w systemach typu wiki**
 - ➔ **Automatyczny monitoring WWW**
-
-

Historia poprawek dokumentu

Hipoteza:

Częste, drobne poprawki są poprawkami błędów językowych (ortograficznych, gramatycznych, stylistycznych...).



Historia poprawek dokumentu

- Dane z Wikipedii są łatwo dostępne i stosunkowo obszerne (17 GB plik historii)
 - Wikipedyści to grupa ludzi wykształconych, więc rażące ich błędy mogą być dobrym wskaźnikiem normy (choć sam korpus nie jest reprezentatywny)
-
-

Metoda

- (1) Odfiltrowanie danych z elementów formatujących
 - (2) Segmentacja na wypowiedzenia
 - (3) Segmentacja na poszczególne wyrazy
 - (4) Porównywanie kolejnych wersji
(wystarczy diff)
 - (5) Odrzucenie zbyt obszernych poprawek; analiza frekwencyjna
-
-

Metoda

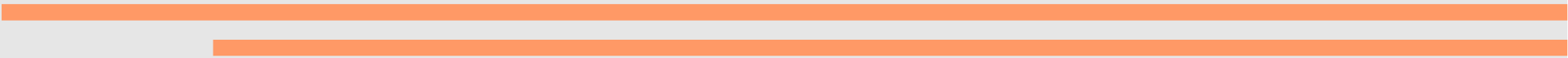
(1) Odfiltrowanie danych z elementów formatujących

Wikipedia zawiera wiele specjalnych znaczników; bez filtrowania poprawki formatowania mieszają się z poprawkami językowymi.

Metoda

(2) Segmentacja na wypowiedzenia

Ja tego nie stosowałem początkowo, ale znaczniki początku i końca wypowiedzenia ułatwiają uchwycenie struktury błędu.



Metoda

(3) Segmentacja na poszczególne wyrazy

Dzięki temu program *diff* porównuje na poziomie pojedynczych wyrazów, co jest przydatne do wykrywania skali poprawki.

Granica wyrazu: odstęp lub znak interpunkcyjny

Metoda

(4) Porównywanie kolejnych wersji

Specjalizowany program mógłby sprawdzić się lepiej, ale *diff* w odpowiednim skrypcie wygenerował zadowalające wyniki (czas generowania: ok. 2 dni)

Metoda

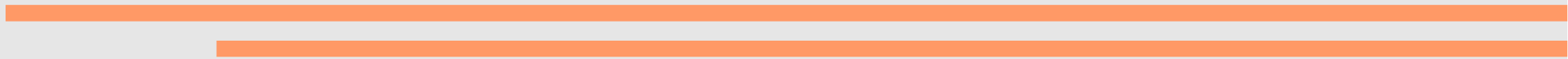
(5) Odrzucenie zbyt obszernych poprawek; analiza frekwencyjna

Dopisanie nowego hasła lub jego znaczna rozbudowa nie jest poprawką; można to zignorować i odfiltrować.

Za pomocą analizy frekwencyjnej i łatwych testów można wyłować częste poprawki.

Wyniki – częste literówki

- Uzyskane wyniki posłużyły do ulepszenia list autokorekty w OpenOffice.org



Wersja pierwotna autokorekty

cjociaż - chociaż

czerweic - czerwiec

czesc - część

czesto - często

czrewiec - czerwiec

czwatrek - czwartek

czynnność - czynność

czynnoać - czynność

czynnosc - czynność

czynność - czynność

część - część

dobzre - dobrze

dokumnety - dokumenty

dr. - dr

dzwiek - dźwięk

dźwięk - dźwięk

grudizeń - grudzień

grudzeiń - grudzień

gruszień - grudzień

Wersja poprawiona

a nad to a nadto
a pro po à propos
absorbacja absorpcja
absorbacji absorpcji
aleji alei
alternatywana alternatywna
aż nad to aż nadto
bedzie będzie
bierzaco bieżąco
bieze bierze
bieże bierze
biznes plan biznesplan
brut brud
byc być
byl był
chor chór
wogóle w ogóle

Surowe dane – „literówki”

województwie	województwa	26162
-	16003	
zamieszkiwało	zamieszkiwały	2738
zamieszkiwały	zamieszkiwało	2646
podstawowe	przeгляд	2519
się	się	2343
		1854
,	,	1760
też:	też	1753
(†	(zm.	895
Gmina=Saint	Gmina=Saint	851
także	też	834
	w	797
-		762
to		587
to	-	584
E.	Edward	512

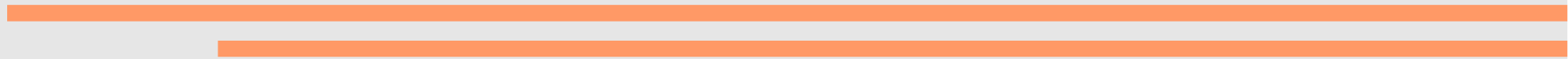
Surowe dane – „literówki”, c.d.

województwie 509
(* (ur. 487
C. Cornelis 466
roku, roku 448
singiel singel 423
— 418
Kolejka Clausura 376
K. Karl 375
US USA 365
zewnątrzne: zewnątrzne 335
(ur. 330
zewntrzne zewnątrzne 318
także: też: 313
tą tę 312
roku r. 312
p.w. pw. 303
(urodzony (ur. 296
Wojny wojny 289

Jak to zrobić w Poliqarpie?

Dwa poziomy zgodności korpusu:

- poziom stosowanego tagsetu
- poziom oprogramowania (możliwość stosowania podobnego języka zapytań)



Jak to zrobić w Poliqarpie?

Dwa poziomy zgodności korpusu:

- poziom stosowanego tagsetu
 - ♦ Tagset IPI (lub bardzo zbliżony)
- poziom oprogramowania (możliwość stosowania podobnego języka zapytań)
 - ♦ Poliqarp (z rozszerzeniami statystycznymi)

Kodowanie poprawek

Poprawki można kodować:

- W korpusie paralelnym: jako pary oryginał–poprawka (niemożliwe w obecnej wersji Poliqarpa)
 - W korpusie zwykłym: oryginały jako dane właściwe, a poprawki jako metadane
 - W korpusie zwykłym: poprawki z oznaczeniem poprawek w stylu *diff* lub przez dodanie znaczników
-
-

Oznaczanie poprawek

- Styl *diff*:
 - ♦ dodany wyraz = „+wyraz”
 - ♦ usunięty wyraz = „-wyraz”

Takie rozwiązanie utrudnia wyszukiwanie form wyrazowych, konieczność używania wyrażeń regularnych (= wolniejsze działanie).

Oznaczanie poprawek

- W znacznikach
 - dodany wyraz ze znacznikiem „ins”
 - usunięty wyraz ze znacznikiem „del”
 - nowa „kategoria” gramatyczna: „correction”

Takie rozwiązanie ułatwia wyszukiwanie form wyrazowych, nie trzeba używać wyrażeń regularnych (= szybsze działanie).

Ograniczenia

- Próbną wersja ma bałagan w znacznikach (niedoskonałości projektu Morfologik)
 - Bez funkcji statystycznych (przynajmniej sortowania wg frekwencji) jest to znacznie mniej przydatne
 - Funkcje kolokacji byłyby interesujące
 - Format paralelny może być klarowniejszy
-
-

I to wszystko.

Dziękuję za uwagę!

Projekt Morfologik:
morfologik.blogspot.com

