Baza parafraz

Kamil Kędzia | Konrad Krulikowski

Zadanie

Stworzenie bazy parafraz na podstawie korpusu równoległego metodą analogiczną do wykorzystanej przy tworzeniu *The Paraphrase Database* (PPDB).

The Paraphrase Database

PPDB

220 milionów par parafraz

73 miliony parafraz frazowych

8 milionów parafraz leksykalnych

140 milionów schematów

100 milionów zdań korpusu równoległego

Zastosowania

- Wnioskowanie z tekstu
- Skracanie, streszczenie
- Upraszczanie, upiększanie, …
- Rozszerzanie / ulepszanie innych zasobów

Zastosowania: wnioskowanie z tekstu

Jeśli narzędzia do badania zależności / podobieństwa nie dają wyniku, sparafrazuj zdania korzystając z PPDB.

- Wzrost z 77,6% do 81,6% przy badaniu zależności.
- Wzrost z 81,9% do 82,7% przy badaniu podobieństwa za pomocą korelacji Pearsona.
- Spadek z 33,6% do 32,2% przy badaniu podobieństwa za pomocą błędu średniokwadratowego.

Zastosowania: skracanie, streszczanie

Riots were sparked by twelve of the cartoons that are offensive to the islamic prophet.

15 słów

72 + 15 znaków

Zastosowania: skracanie, streszczanie

12 comics insulting Mohammad caused riots.

6 słów

36 + 6 znaków

Zastosowania: skracanie, streszczanie

12 comics insulting Mohammad caused riots.

60% spadek liczby słów

50% + 1,16% spadek liczby znaków

Riots were sparked by twelve of the cartoons that are offensive to the islamic prophet.

Zastosowania: upraszczanie, ...

Upraszczenie:	*accelerate*
	speed up
Upiększanie:	*set up*
	establish
Łagodzenie:	*abso-fucking-lutely*
	indeed

Zastosowania: upraszczanie, ...

Degenderyzacja: c*hairman*

chairperson

Rozwijanie skrótów: *ROI*

Return On Investment

Inferencja typów: *Hudson*

Hudson Bay

Zastosowania: PPDB i PropBank

*few economists **expect** the data to show*

await

wait

hope

anticipate

…

grant

let

Zastosowania: precision vs. recall

	All	Lexical	One-To-Many	Phrasal	Syntactic
S	Paraphrases (424MB, 6.8M rules)	Paraphrases (1.7MB, 31k rules) Identity (16MB, 437k rules)	One-To-Many (3.8MB, 47k rules) Many-To-One (3.8MB, 47k rules)	Paraphrases (42MB, 637k rules) Identity (170MB, 4.1M rules)	Constituent (38MB, 585k rules) Non-Constituent (343MB, 5.6M rules)
M	Paraphrases (757MB, 11.9M rules)	Paraphrases (1.7MB, 69k rules) Identity (16MB, 468k rules)	One-To-Many (7.6MB, 94k rules) Many-To-One (7.6MB, 94k rules)	Paraphrases (42MB, 1.2M rules) Identity (170MB, 4.3M rules)	Constituent (69MB, 1.0M rules) Non-Constituent (601MB, 9.6M rules)
L	Paraphrases (1.5GB, 23.5M rules)	Paraphrases (12MB, 198k rules) Identity (19MB, 503k rules)	One-To-Many (16MB, 188k rules) Many-To-One (16MB, 188k rules)	Paraphrases (209MB, 3.0M rules) Identity (191MB, 4.5M rules)	Constituent (148MB, 2.2M rules) Non-Constituent (1.2GB, 18.2M rules)
XL	Paraphrases (2.8GB, 43.2M rules)	Paraphrases (33MB, 548k rules) Identity (20MB, 532k rules)	One-To-Many (31MB, 376k rules) Many-To-One (31MB, 376k rules)	Paraphrases (486MB, 6.9M rules) Identity (198MB, 4.7M rules)	Constituent (300MB, 4.4M rules) Non-Constituent (2.1GB, 31.4M rules)
XXL	Paraphrases (5.7GB, 86.4M rules)	Paraphrases (125MB, 2.1M rules) Identity (21MB, 559k rules)	One-To-Many (61MB, 752k rules) Many-To-One (61MB, 752k rules)	Paraphrases (1.5GB, 20.2M rules) Identity (204MB, 4.8M rules)	Constituent (644MB, 9.3M rules) Non-Constituent (3.6GB, 54.8M rules)
XXXL	Paraphrases (12.2GB, 169M rules)	Paraphrases (451MB, 7.6M rules) Identity (22MB, 570k rules)	One-To-Many (117MB, 1.5M rules) Many-To-One (117MB, 1.5M rules)	Paraphrases (4.9GB, 68.4M rules) Identity (207MB, 4.9M rules)	Constituent (1.1GB, 16.1M rules) Non-Constituent (5.1GB, 77.4M rules)

PPDB — metoda

1. Wybierz fragment ze zdania
2. Przetłumacz zdanie
3. **Znajdź tłumaczenie fragmentu**
4. Znajdź inne zdanie z takim fragmentem
5. Przetłumacz na język źródłowy
6. **Znajdź parafrazę**
7. **Oceń**

PPDB — metoda

1. Wybierz fragment ze zdania
2. Przetłumacz zdanie
3. **Znajdź tłumaczenie fragmentu**
4. Znajdź inne zdanie z takim fragmentem

5. Przetłumacz na język źródłowy

6. **Znajdź parafrazę**
7. **Oceń**

PPDB — przykład

1. **Uważa** on, **że** rządy europejskie powinny…
2. *European governments should … **believes** Mr Beffa.*
3. ***Believe***
4. *…set different goals in line with what they **believe to be** attainable.*
5. …wyznaczyć inne cele zgodnie z tym, co **w ich oczach jest** możliwe do osiągnięcia.
6. **w ich oczach jest**
7. **…**

Istniejące rozwiązanie

PPDB zawiera zasób w języku polskim…!

… niestety bardzo słabej jakości.

Wśród 63 wpisy dla wyrazu „zły”:

- 52 to formy wyrazu „zły” lub „źle”,
- 10 to formy wyrazu „niewłaściwy”,
- 1 to wyrażenie „zły obrót”

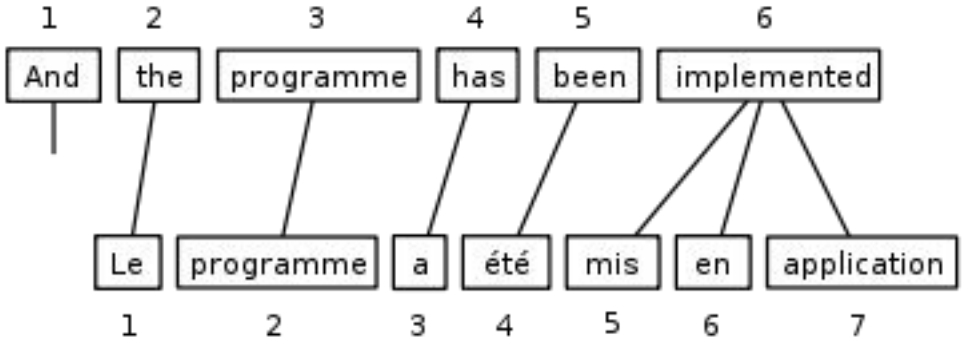
Wyzwanie: wyrównanie wyrazów

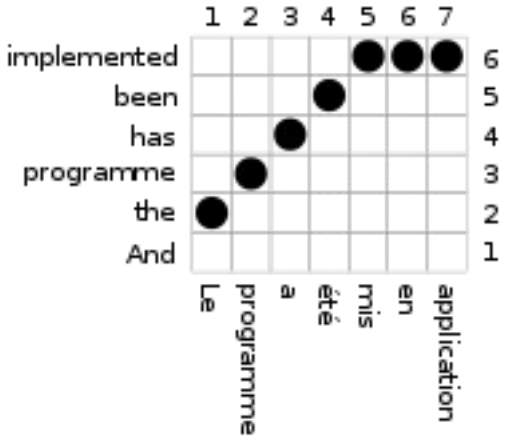
Dopasowanie zależności między tłumaczonymi wyrazami w dwutekście:

- graf dwudzielny między wyrazami tekstów
- modele bezkontekstowe i kontekstowe
- zwykle uczenie bez nadzoru

Przykład — wyrównanie wyrazów

Wyrównanie dwutekstu angielsko-francuskiego:





Wyzwanie: homonimia i polisemia

Homografia: identyczne formy wyrazowe o różnych znaczeniach.

fleksja: „dam” od „dać” i „dam” od „dama”

słowotwórstwo: „ranny” w zn. „zraniony”, „ranny” w zn. „poranny”

leksyka: „bal” („przyjęcie”), „bal” („kłoda”)

składnia: „zdrada przyjaciela” (agent zdrady)

Wyzwanie: homonimia i polisemia

Polisemia: jedno słowo w różnych znaczeniach.

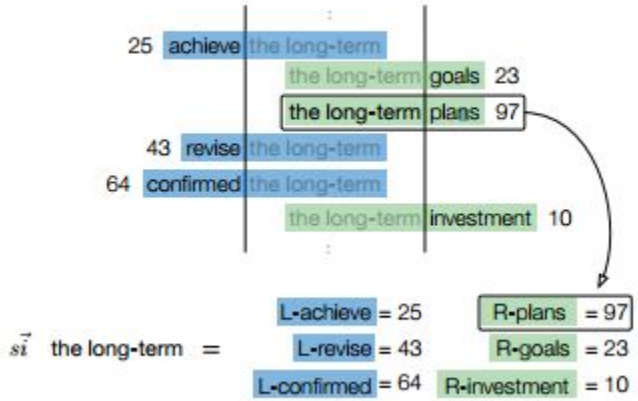
„powód” w zn. „przyczyna, motyw”,
„powód” w zn. „osoba pozywająca do sądu”

„język” w zn. „organ anatomiczny”,
„język” w zn. „system komunikacji”

Wyzwanie: specyfika polszczyzny

- Fleksyjność zamiast pozycyjności
  etykietowanie, formy słownikowe, heurystyki
- Szyk ruchomy i ustalony
  wyrównanie, heurystyki
- Inna składnia
  etykietowanie, wyrównanie, analizowanie

Na ratunek: n-gramy



Na ratunek: *n*-gramy

Post hoc ergo propter hoc

Correlation does not imply causation

Na ratunek: *n*-gramy

Correlation does not imply causation

Na ratunek: *n*-gramy

Wyzwania:

1. homonimia
2. polisemia
3. fleksja
4. składnia

Zasób — *n*-gramy Google

8,3 miliona 4-gramów

11,9 miliona 3-gramów

6,2 miliona 2-gramów

Zasób — *n*-gramy ze zbalansowanego NKJP

232,4 miliona 5-gramów

217,5 miliona 4-gramów

170,1 miliona 3-gramów

75,3 miliona 2-gramów

Nasze rozwiązanie

Wykorzystane zasoby

- ~75 tys. zdań korpusu równoległego
- ~1,5 mln leksemów w języku źródłowym
- ~1,5 mln leksemów w języku osiowym

Dla porównania z PPDB:

- ~100 mln zdań korpusu równoległego

Porównanie wyników — „opiera się na”

W PPDB

- jest oparta na
- jest oparty na
- bazuje na
- oparta jest na
- polega na

U nas (z 56)

- na bazie
- wykorzystuje
- oparty na
- oparte o
- zbudowana na
- wynika z
- na podstawie
- według
- wynikające z
- korzysta z
- stawia na
- polega na
- polegał na
- oparty na modelu

Porównanie wyników — „na podstawie”

W PPDB

- jest oparta na
- jest oparty na
- bazuje na
- oparta jest na
- polega na

U nas (z 219)

- ze źródeł
- z
- na
- dzięki
- opierano na
- zbudowana na
- opierają na
- opartą na
- uwzględniającą
- pozyskiwany z
- bazują na
- wynika z
- opracowane na
- wynikające z

Porównanie wyników — „w zderzeniu”

W PPDB

U nas (z 3)

- w miejscu, gdzie uderzyła
- w momencie uderzenia
- w wyniku uderzenia
- w chwili uderzenia

Porównanie wyników — „między innymi”

W PPDB

U nas (z 163)

- również
- włączając
- także
- wśród innych
- w tym

Porównanie wyników — „są członkami”

W PPDB

U nas (z 2)

- to członkowie
- to państwa członkowskie

Porównanie wyników — „stanowi dobry przykład”

W PPDB

U nas (z 1)

- jest dobrym przykładem
- dobrym przykładem jest

Porównanie wyników — „cierpi na brak”

W PPDB

U nas (z 1)

- cierpi na niedobór
- ma regularnie problemy z powodu braku

Porównanie wyników — „uważa się za”

W PPDB

U nas (z 28)

- postrzegano jako
- jest uznawane za
- przytoczone jako
- poczytywane za
- traktowana jak
- są postrzegane jako
- postrzegany jest jako
- uważano za

Ocena: 3 miliony 3-gramów

wyjście	wyjście	ocena
kurs	szkolenie	0.79
kurs	zajęcia	0.83
kurs	uczeń	0.97
kurs	szkoła	0.98

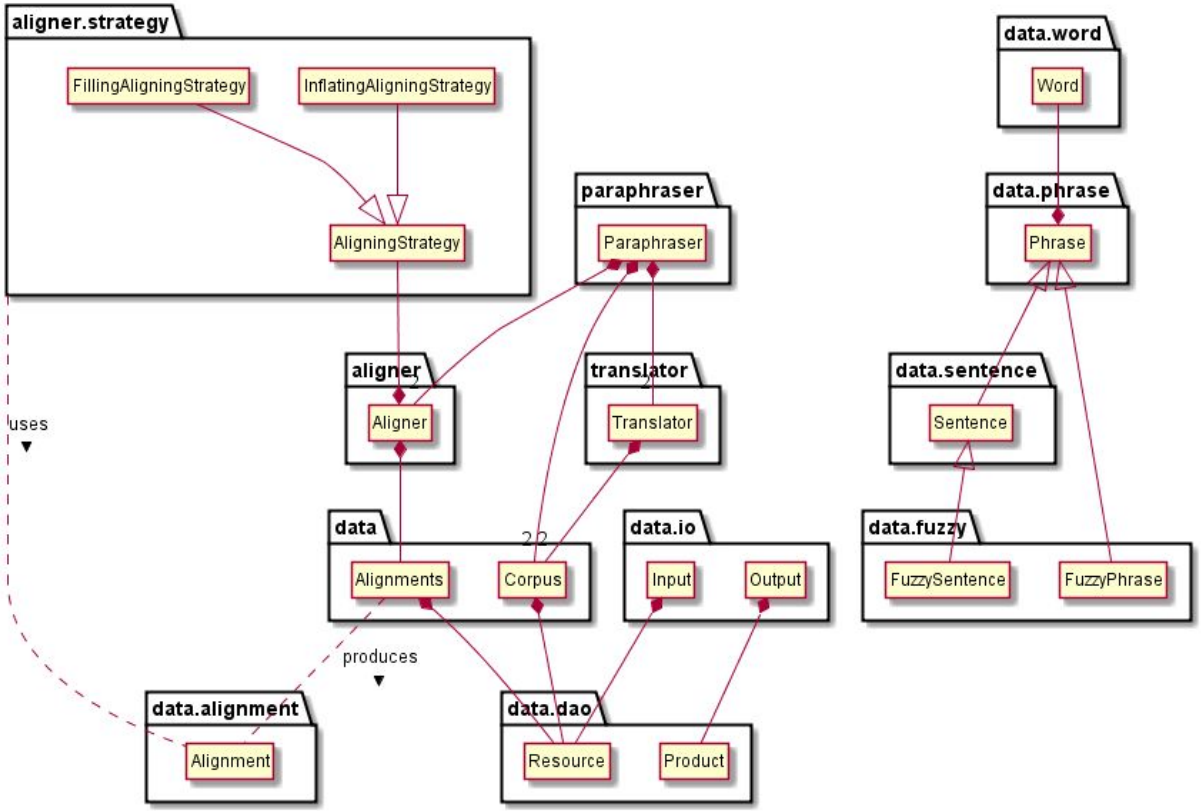
Ocena: homonimia i polisemia

problem	wejście	połączenie	wyjście	ocena
fleksja	dać	dam	dama	1.0
słowotwórstwo	zranił	ranny	porannych	1.0
leksyka	przyjęcie	bal	kłoda	1.0
polisemia	przyczyna	powód	oskarżyciel	1.0

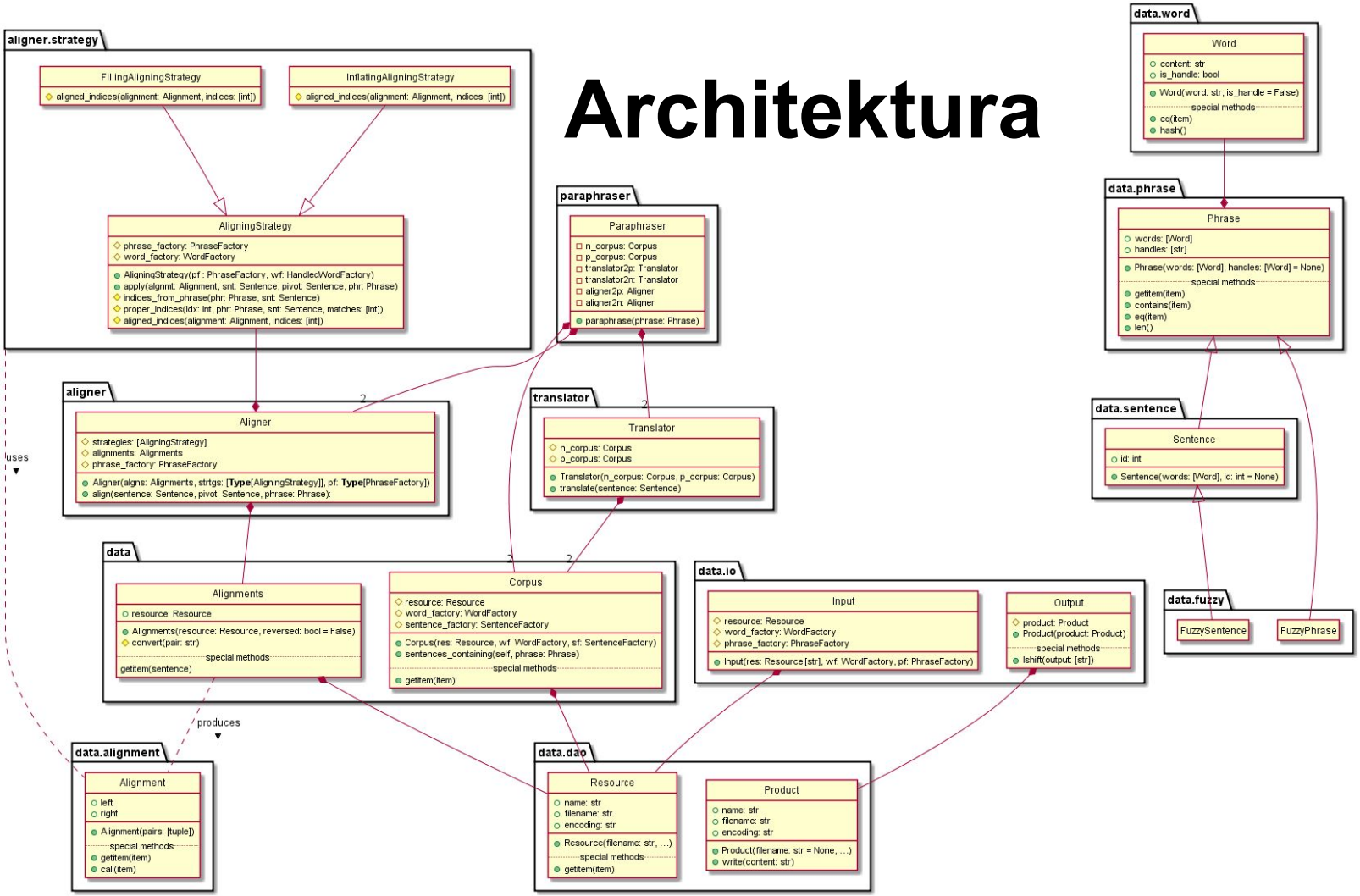
Rozwiązanie — architektura

S *ingle responsibility*	Pojedyncza odpowiedzialność	Klasa powinna mieć tylko jedną odpowiedzialność (nigdy nie powinien istnieć więcej niż jeden powód do zmiany klasy)
O *pen/ closed*	Otwarte/zamknięte	Elementy oprogramowania powinny być otwarte na rozszerzenie, ale zamknięte na zmiany
L *iskov substitution*	Podstawienie Liskov	Podstawienie obiektów egzemplarzami ich klas pochodnych nie powinno wpływać na poprawność
I *nterface segregation*	Oddzielenie interfejsów	Lepiej stosować wiele interfejsów szczegółowych niż jeden ogólnego przeznaczenia
D *ependency inversion*	Odwrócenie zależności	Zależności powinny wynikać z abstrakcji, nie ze szczegółów implementacji

Rozwiązanie — Architektura



Architektura



Pytania…?

Materiały źródłowe

— https://twitter.com/ppdb

— http://techtalks.tv/talks/ppdb-the-paraphrase-database/58473/

— http://hamlet.edu.pl/szyk-skl

— http://zmienic-swiat.blogspot.com/2013/11/skadnia-semantyczna-strukturalna-i-szyku.html

— http://pl.wikipedia.org/wiki/

— http://morfologik.blogspot.com/

— http://glass.ipipan.waw.pl/multiservice/

— http://www.sjikp.us.edu.pl/pliki/ksiazki/piotr_zmigrodzki.pdf

— http://www.kozlowska.uksw.edu.pl/img/skladnia6.pdf