

Zastosowanie Wikipedii w przetwarzaniu języka naturalnego

Aleksander Smywiński-Pohl

Katedra Lingwistyki Komputerowej, Uniwersytet Jagielloński

Plan prezentacji

Wikipedia

Klasyfikacja

Zastosowanie w NLP

Plan prezentacji

Wikipedia

Klasyfikacja

Zastosowanie w NLP



WIKIPEDIA

The Free Encyclopedia

Rysunek :

http://img2.wikia.nocookie.net/__cb20100519143323/logopedia/images/b/b3/Wikipedia-logo-v2-en.svg

Wikipedia

Mike Bergman „Sweetpedia” (2010) – lista 246 artykułów naukowych na temat wykorzystania Wikipedii w problemach z zakresu sztucznej inteligencji i przetwarzaniu języka naturalnego

- ▶ wyszukiwanie semantyczne
- ▶ rozstrzygnięcie wieloznaczności
- ▶ ekstrakcja informacji

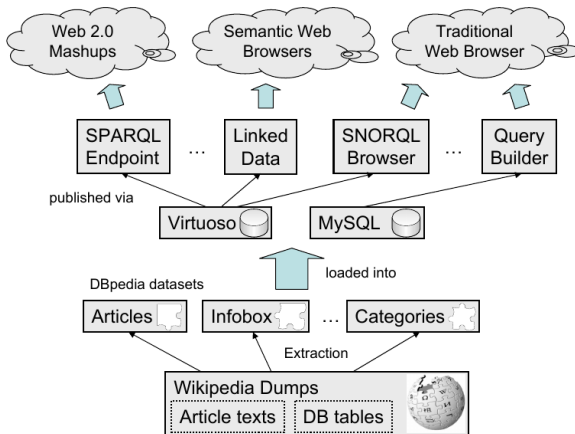
Komercyjne systemy korzystające z Wikipedii

- ▶ IBM Watson
- ▶ Freebase
- ▶ Google Knowledge Graph
- ▶ Apple Siri



DBpedia

- ▶ baza wiedzy zbudowana na podstawie Wikipedii
- ▶ powstała w 2006 roku
- ▶ ekstrakcja z Infoboksów
- ▶ formalizm: RDF
- ▶ ontologia: DBpedia ontology

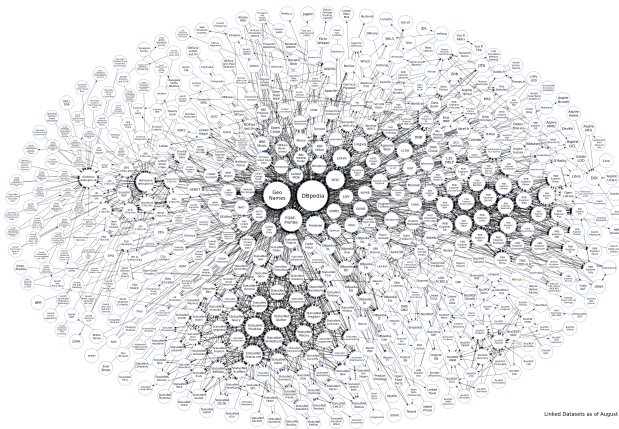


Rysunek : S. Auer, Ch. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives: *DBpedia: A nucleus for a Web of Open Data*

Infobox

```
{ {Państwo infobox
|nazwa_oryginalna      = Rzeczpospolita Polska
|nazwa_polska          =
|flaga_obraz           = Flag of Poland.svg
|godło_obraz           = Herb Polski.svg
...
|głowa_państwa         = [[Bronisław Komorowski]]
|głowa_państwa_opis    = [[Prezydent Rzeczypospolitej Polskiej|prezydent RP]]
|szef_rządu            = [[Donald Tusk]]
|szef_rządu_opis       = [[Premierzy Polski|prezes Rady Ministrów]]
...
|powierzchnia          = 312 679 <ref>...</ref>
|powierzchnia_wód      =
|powierzchnia_miejsce  = 70
|ludność               = 38 485 779<ref>...</ref>
|gęstość               = 123
|gęstość_miejsce       = 90
|ludność_rok           = 2014
... }
```

LOD cloud



Rysunek : <http://richard.cyganiak.de/2007/10/lod/>

Polska w DBpedia

dbpedia-owl:anthem	■ dbpedia:Poland_Is_Not_Yet_Lost
dbpedia-owl:areaTotal	■ 312602266860.239075 (xsd:double) ■ 312679000000.000000 (xsd:double)
dbpedia-owl:currency	■ dbpedia:Polish_zloty
dbpedia-owl:demonym	■ Pole Polish
dbpedia-owl:ethnicGroup	■ dbpedia:German_minority_in_Poland ■ dbpedia:Kashubians ■ dbpedia:Ukrainians_in_Poland ■ dbpedia:Poles ■ dbpedia:Belarusian_minority_in_Poland ■ dbpedia:Silesians
dbpedia-owl:ethnicGroupsInYear	■ 2011-01-01 (xsd:date)
dbpedia-owl:flag	■ Flag of Poland.svg
dbpedia-owl:foundingDate	■ 0966-04-14 (xsd:date) ■ 1025-04-18 (xsd:date) ■ 1945-04-08 (xsd:date)
dbpedia-owl:governmentType	■ dbpedia:Parliamentary_republic
dbpedia-owl:language	■ dbpedia:Polish_language
dbpedia-owl:leader	■ dbpedia:Donald_Tusk ■ dbpedia:Bronisław_Komorowski
dbpedia-owl:leaderTitle	■ President ■ Prime Minister
dbpedia-owl:longName	■ Republic of Poland

Rysunek : <http://www.dbpedia.org/page/Poland>

Problemy DBpedii

- ▶ różnorodność formatu i reprezentacji danych źródłowych
- ▶ sprzeczne informacje
- ▶ brak informacji o zakresie ważności stwierdzeń
- ▶ konieczność ręcznej konstrukcji szablonów ekstrakcyjnych

Wikidata/Wikidane

- ▶ powstałe w 2013
- ▶ „konkurencja” dla DBpedii
- ▶ tworzona w duchu Wikipedii (niemiecki oddział MediaWiki)
- ▶ początkowo celem było ujednolicenie odnośników między wersjami językowymi



Strona główna
Portal społeczności
Kafejka
Utwórz nowy element
Element według tytułu
Ostatnie zmiany
Losowa strona
Pomoc
Darowizny

Drukuj lub eksportuj
Utwórz książkę
Pobierz jako PDF
Wersja do druku

Narzędzia

Linkujące
Zmiany w linkowanych
Strony specjalne
Link do tej wersji
Informacje o tej stronie
Koncept URI
Cytowanie tego artykułu

polski Apohilo Dyskusja Preferen

Strona [Dyskusja](#)

[Czytaj](#) [Wyświetl hist](#)

Polska Akademia Nauk (Q651690)

[\[edytuj \]](#)

Nie podano opisu

Nie zdefiniowano aliasów.

▼ W innych językach [Konfiguruj](#)

Język	Etykieta	Opis	Inne nazwy
polski	Polska Akademia Nauk	Nie podano opisu	
angielski	Polish Academy of Sciences	Nie podano opisu	
śląski	Etykieta nie została podana	Nie podano opisu	
białoruski	Etykieta nie została podana	Nie podano opisu	

Deklaracje

kategoria Commons	PAN [edytuj]
	► 1 odnośnik
	[dodaj]
identyfikator NKC	kn20020808069 [edytuj]
	▼ 0 odnośników
	[dodaj przypis]
	[dodaj]

Rysunek : <https://www.wikidata.org/wiki/Q651690>

Cechy Wikidata

- ▶ dane wprowadzane są ręcznie i automatycznie
- ▶ dane muszą posiadać źródło
- ▶ może istnieć kilka „alternatywnych” faktów
- ▶ fakty mogą posiadać meta-dane (np. data urzędowania określonej osoby na stanowisku)
- ▶ dodawanie **faktów** nie jest ograniczone (pod warunkiem akceptacji licencji)
- ▶ dodawanie **relacji** jest ograniczone – propozycja → akceptacja społeczności → utworzenie
- ▶ **klasy** są traktowane jak zwykłe obiekty

Plan prezentacji

Wikipedia

Klasyfikacja

Zastosowanie w NLP



Projekty związane z klasyfikacją

- ▶ DBpedia ontology
- ▶ YAGO
- ▶ WikiTaxonomy
- ▶ BabelNet
- ▶ WikiNet
- ▶ MENTA
- ▶ Linked Hypernoms
- ▶ WiBi
- ▶ Tipalo
- ▶ Airpedia

cycloped.io

- ▶ współpraca: UJ, AGH, Techmo, Structured Dynamics
- ▶ osoby: Krzysztof Wróbel, Mike Bergman, Bartosz Ziółko
- ▶ klasyfikacja artykułów Wikipedii
- ▶ cecha charakterystyczna: wykorzystanie ontologii Cyc
- ▶ aktualnie tylko angielska Wikipedia
- ▶ trwają prace na *polską* oraz *japońską* Wikipedią
- ▶ kod open-source `github.com/cycloped-io`

Dlaczego Cyc?

- ▶ dobrze zdefiniowana struktura
- ▶ duża liczba pojęć (porównywalna z liczbą rzeczowników w angielskim WordNecie)
- ▶ formalne definicje pojęć
- ▶ formalne definicje relacji
- ▶ silnik inferencyjny
- ▶ budowa modułowa (mikroteorie)
- ▶ klasyfikacja wieloaspektowa
- ▶ meta-modelowanie i meta-assertje
- ▶ niewielka liczba pojęć złożonych
- ▶ role
- ▶ wykrywanie sprzeczności

Metody klasyfikacji

- ▶ pierwsze zdanie traktowane jak definicja
- ▶ kategorie Wikipedii
- ▶ mapowanie do DBpedii
- ▶ bezpośrednie mapowanie do Cyc
- ▶ wzorce nazw kategorii

Pierwsze zdanie

Artykuł Dyskusja

Czytaj Tekst źródłowy Historia i autorzy Szukaj

 Wiki lubi Przyrodę – odkrywaj piękno natury, rób zdjęcia dla Wikipedii i wygrywaj nagrody!

Polska

[edytuj] Na mapach: 52°N 19°E (mapa)

 Zobacz też: [inne znaczenia](#). Na tę stronę wskazuje także [przekierowanie „RP”](#). Zobacz też: [RP](#) (ujednoznacznienie)

Polska, **Rzeczpospolita Polska** (RP) – państwo unitarne w Europie Środkowej położone między Morzem Bałtyckim na północy a Sudetami i Karpatami na południu, w przeważającej części w dorzeczu Wisły i Odry. Powierzchnia administracyjna Polski wynosi 312 679 km²^[b]^[1], co daje jej 70. miejsce na świecie i 9. w Europie. Zamieszкана przez prawie 38,5 miliona ludzi (2014), zajmuje pod względem liczby ludności 34. miejsce na świecie^[5], a 6. w Unii Europejskiej.

Od północy Polska graniczy z Rosją (z jej obwodem kaliningradzkim) i Litwą, od wschodu z Białorusią i Ukrainą, od południa ze Słowacją i Czechami, od zachodu z Niemcami. Większość północnej granicy Polski wyznacza wybrzeże Morza Bałtyckiego. Polska wyłączna strefa ekonomiczna na Bałtyku graniczy ze strefami Danii i Szwecji.

Rzeczpospolita Polska



Kategorie

Artykuł Dyskusja

Czytaj

Tekst źródłowy

Historia i autorzy



Szukaj



Wiki Lub! Przyrodę – odkrywaj piękno natury, rób zdjęcia dla Wikipedii i wygrywasz nagrody!



Polska [edytuj]

Na mapach: 52°N 19°E (mapa)Zobacz też: [inne znaczenia](#). Na tę stronę wskazuje także [przekierowanie „RP”](#). Zobacz też: [RP](#) ([ujednoznacznienie](#))

Polska, **Rzeczpospolita Polska** (RP) – państwo unitarne w Europie Środkowej położone między Morzem Bałtyckim na północy a Sudetami i Karpatami na południu, w przeważającej części w dorzeczu Wisły i Odry. Powierzchnia administracyjna Polski wynosi 312 679 km²^[b]^[1], co daje jej 70. miejsce na świecie i 9. w Europie. Zamieszkała przez prawie 38,5 miliona ludzi (2014), zajmuje pod względem liczby ludności 34. miejsce na świecie^[5], a 6. w Unii Europejskiej.

Od północy Polska graniczy z [Rosją](#) (z jej obwodem kaliningradzkim) i [Litwą](#), od wschodu z [Białorusią](#) i [Ukrainą](#), od południa ze [Słowacją](#) i [Czechami](#), od zachodu z [Niemcami](#). Większość północnej granicy Polski wyznacza wybrzeże Morza Bałtyckiego. Polska wyłączna strefa ekonomiczna na Bałtyku graniczy ze strefami [Danii](#) i [Szwecji](#).

Rzeczpospolita Polska



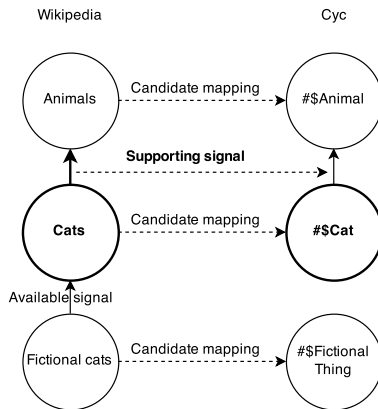
Kontrola autorytatywna (obiekt geograficzny):
[VIAF: 141810140](#) | [LCCN: n79131071](#) | [GND: 4046496-9](#) | [NDL: 00569130](#) | [BnF: 11880842g](#) | [SUDOC: 02658994X](#)
[WorldCat](#)

Kategorie: [Polska](#) | [Członkowie Organizacji Narodów Zjednoczonych](#) | [Państwa członkowskie Unii Europejskiej](#) | [Państwa należące do NATO](#)

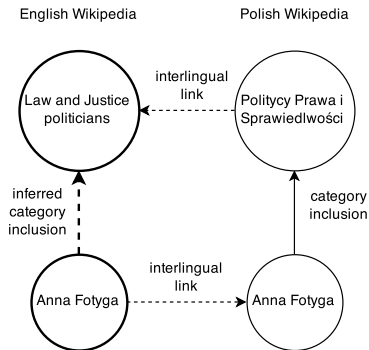
Mapowanie kategorii

- ▶ identyfikacja kategorii, których syntaktyczna głowa jest w liczbie mnogiej, np. *People from New York*
- ▶ generowanie listy kandydatów
 - ▶ # \$Person
 - ▶ # \$PeopleMagazine-TheMagazine
- ▶ filtrowanie listy
- ▶ ujednoznacznianie sensu

Podobieństwo strukturalne



Wykorzystanie innych edycji językowych



Wzorce nazw kategorii

- ▶ Ludzie związani z Warszawą
- ▶ Ludzie związani z Krakowem
- ▶ Ludzie związani z Wrocławiem
- ▶ Ludzie związani z $X \rightarrow \# \$ Person$

Wyniki klasyfikacji dla angielskiej Wikipedii

	A	C_V	F_1	C_T	F'_1
cc	89.9	65.5	75.8	49.2	63.6
cd	91.7	60.0	72.5	44.2	59.6
db	90.1	84.7	87.3	73.9	81.2
di	92.2	18.0	30.0	1.0	2.0
fs	91.1	68.7	77.7	55.5	69.0
lh	87.5	94.4	90.8	83.8	85.6
all	90.8	100.0	95.2	96.2	93.4

Zbiór walidacyjny: 3600 artykułów

Plan prezentacji

Wikipedia

Klasyfikacja

Zastosowanie w NLP



Entity linking/WSD

- ▶ wykrywanie wystąpień odniesień do konkretnych obiektów (coś więcej niż NER)
- ▶ referencyjnym zasobem jest Wikipedia – stworzenie stosownego słownika i opisanie wszystkich znaczeń jest czasochłonne, *vide* słowosieć
- ▶ otagowanie tekstu odnośnikami do sensów/obiektów jest jeszcze bardziej czasochłonne

Wykorzystanie Wikipedii

- ▶ zbiór sensów dostajemy „za darmo” – strony ujednoznaczniające
- ▶ zbiór synonimów dostajemy „za darmo” – przekierowania, nazwy odnośników
- ▶ zbiór odnośników dostajemy „za darmo” – na podstawie tekstu Wikipedii możemy wygenerować miliony wektorów uczących

Wyniki ujednoznaczniania – dla danych wieloznacznych

Metoda ujednoznaczniania	Pr [%]	Rc [%]	F1 [%]
Losowy sens	39,7	26,4	31,7
Losowy sens o $P > 0,5\%$	47,0	47,3	47,2
Najczęstszy sens	81,6	82,2	81,9
SR_G	82,5	83,5	83,0
SR_G + dodatkowe cechy	84,9	83,2	84,0
SR_J	85,4	89,8	87,6
SR_J + dodatkowe cechy	90,4	93,0	91,7

Wyniki dla różnych zbiorów testowych

Zbiór testowy	Rozmiar	Pr [%]	Rc [%]	F1 [%]
Wektory cech z Wikipedii	281714	97,0	94,4	96,7
Odtworzenie odnośników	31092	94,1	90,3	92,2
Notatki PAP	500	79,6	—	—

Ujednoznacznianie sensu (entity linking)

A. Pohl (2012), *Improving the Wikipedia Miner Word Sense Disambiguation Algorithm*, [w:] Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, s. 241-248, ISBN: 978-1-4673-0708-6

Rozpoznawanie jednostek nazewniczych

Połączenie entity linking i klasyfikacji.

System	Pr [%]	Rc [%]	F ₁ [%]
SProUT	78.0	38.0	51.0
NERF	80.5	74.5	77.3
„Ulepszony” Wikipedia Miner	90.0	41.8	57.1

Rozpoznawanie jednostek nazwniczych

A. Pohl (2013), *Knowledge-based Named Entity Recognition in Polish*, [w:] Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS)

Ekstrakcja relacji semantycznych

1. wybór relacji (np. całość-część)
2. wygenerowanie par uczących (np. rekin-płetwa)
3. odnalezienie par uczących w korpusie tekstów
4. utworzenie *formalnych* szablonów ekstrakcyjnych (np. *-dat – „płetwa rekina”, „płetwy rekina”, etc.)
5. statystyczna analiza szablonów
6. odnalezienie zdań pasujących do szablonów w korpusie tekstów
7. **określenie ograniczeń semantycznych dla odnalezionych przykładów uczących**
8. utworzenie *semantycznych* szablonów ekstrakcyjnych

Metody określania ograniczeń semantycznych

- ▶ ręczne przeglądanie wyników – ekstrakcja ograniczeń
- ▶ wykorzystanie ograniczeń semantycznych z Cyc
- ▶ wykrycie ograniczeń semantycznych w DBpedii

Ręczne określanie ograniczeń

- ▶ wyszukanie zdań pasujących do wzorców formalnych
- ▶ znakowanie znalezionych zdań jako zawierających bądź niezawierających zadaną relację
- ▶ ekstrakcja ograniczeń ze zdań zawierających wybraną relację

Wykorzystanie ograniczeń Cyc

- ▶ nie wymaga żadnej pracy manualnej
- ▶ relacje w Cyc tworzą hierarchię
- ▶ wybranie relacji
- ▶ selekcja wszystkich pod-relacji
- ▶ automatyczne utworzenie zbioru ograniczeń

Wykorzystanie DBpedii

- ▶ wybór relacji w ontologii DBpedii odpowiadających zadanej relacji semantycznej
- ▶ znalezienie wszystkich faktów zawierających tę relację
- ▶ określenie kategorii semantycznych encji (artykułów Wikipedii) występujących w odnalezionych relacjach
- ▶ określenie ograniczeń semantycznych na podstawie analizy statystycznej

Dopasowanie wzorców formalnych relacji

całość-część

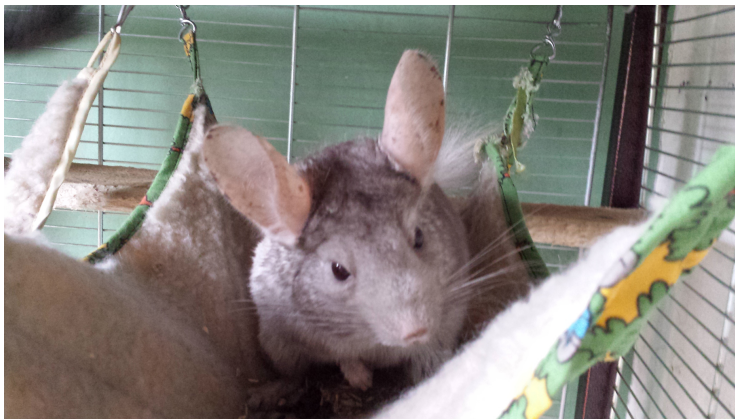
Typ zdania	Liczba	Udział %
zawierające relację	110	10,2
niezawierające relacji	416	38,6
niepoprawne dopasowanie	447	41,5
przykład problematyczny	105	9,7
w sumie	1078	100,0

Wyniki dopasowania wzorców semantycznych

Źródło	Gen.	Id. kat.	<i>Pr</i> [%]	<i>Rc_{rel}</i> [%]	<i>F₁</i> [%]
Weryf. ręczna	—	—	89,2	56,9	69,5
	—	+	92,8	55,1	69,2
	+	—	77,7	64,8	70,7
	+	+	84,3	61,8	71,3
Cyc	—	—	82,4	9,4	17,0
	—	+	81,9	8,7	15,7
	+	—	76,6	41,2	53,5
	+	+	87,6	37,5	52,5
DBpedia	—	—	81,0	20,6	33,5
	—	+	84,4	16,4	27,5
	+	—	70,6	80,3	75,2
	+	+	76,4	74,0	75,2

Podsumowanie

- ▶ `cycloped.io`
 - ▶ klasyfikacja artykułów
 - ▶ ekstrakcja relacji
 - ▶ wersje dla innych języków (w szczególności polskiego)
- ▶ `apohllo.pl/publikacje`



Rysunek : Gagarin

Dziękuję!