

Task 3: Entity linking

Aleksander Smywiński-Pohl

AGH University of Science and Technology

apohllo@agh.edu.pl

<http://apohllo.pl> (<http://apohllo.pl>)



Agenda

- problem statement
- participants
- results
- EL vs. NER vs.
disambiguation
- datasets
- SOTA approaches



Entity linking - problem statement

Utworzenie Polskiej Akademii Nauk było związane z postanowieniami, które zapadły na I Kongresie Nauki Polskiej w roku 1951 i połączone było z likwidacją Polskiej Akademii Umiejętności oraz Towarzystwa Naukowego Warszawskiego. Stworzenie tego rodzaju instytucji miało ułatwić władzom politycznym kontakt ze środowiskiem naukowym oraz zapewniać sobie jego wsparcie, przy jednoczesnej kontroli tego środowiska. Miało ponadto ułatwić finansowanie badań naukowych i koncentrowanie ich wokół problemów uznanych za najważniejsze dla państwa, w czym istotną rolę odgrywały zmiany zachodzące w organizacji nauki w krajach zachodnich w związku z powojenną rewolucją naukowo-techniczną. Polska Akademia Nauk została powołana ustawą o PAN z 30 października 1951 roku. Początkowo była tylko korporacją uczonych, jednak w 1960 roku została przekształcona w rządową instytucję centralną, sprawującą ogólną pieczę nad nauką w Polsce i zarządzającą siecią instytutów. W 1990 roku PAN straciła status instytucji rządowej, stając się ponownie zarówno korporacją uczonych, jak i siecią instytutów.



Utworzenie [Polskiej Akademii Nauk \(\)](#) było związane z postanowieniami, które zapadły na [I Kongresie Nauki Polskiej \(\)](#) w roku 1951 i połączone było z likwidacją [Polskiej Akademii Umiejętności \(\)](#) oraz [Towarzystwa Naukowego Warszawskiego \(\)](#). Stworzenie tego rodzaju instytucji miało ułatwić władzom politycznym kontakt ze środowiskiem naukowym oraz zapewniać sobie jego wsparcie, przy jednoczesnej kontroli tego środowiska. Miało ponadto ułatwić finansowanie badań naukowych i koncentrowanie ich wokół problemów uznanych za najważniejsze dla państwa, w czym istotną rolę odgrywały zmiany zachodzące w organizacji nauki w krajach zachodnich w związku z powojenną rewolucją naukowo-techniczną. [Polska Akademia Nauk \(\)](#) została powołana ustawą o [PAN \(\)](#) z 30 października 1951 roku. Początkowo była tylko korporacją uczonych, jednak w 1960 roku została przekształcona w rządową instytucję centralną, sprawującą ogólną pieczę nad nauką w [Polsce \(\)](#) i zarządzającą siecią instytutów. W 1990 roku [PAN \(\)](#) straciła status instytucji rządowej, stając się ponownie zarówno korporacją uczonych, jak i siecią instytutów.



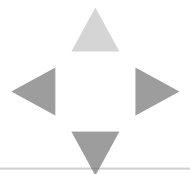
Participants

- baseline
- Cheeky
Mouse
- Zbronk Studio

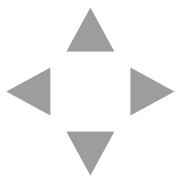


Baseline

```
In [5]: def find_matches(wiki, tokens):
#tokens = [('Ala', "0"), ('ma', "1"), ('kota', "0"), ('i', "1"), ('psa', "0")]
    for size in reversed(range(len(tokens) + 1)):
        if(size == 0):
            continue
        candidates = {}
        for index in range(0, (len(tokens) - size + 1)):
            string = ''
            lemmas = ''
            for token, lemma, space in tokens[index:index+size]:
                if(space == "1"):
                    string += ' '
                    lemmas += ' '
                string += token
                lemmas += lemma
            string = string.strip()
            lemmas = lemmas.strip()
            if(string in wiki.senses):
                candidates[string] = wiki.senses[string]
            if(lemmas in wiki.senses and not lemmas in candidates):
                candidates[lemmas] = wiki.senses[lemmas]
        if(len(candidates) > 0):
            return candidates
    return {}
```



```
In [ ]: output = open(el_path + 'most-probable.csv', 'w')
        for k,v in entities.items():
            candidates = find_matches(wiki, v)
            values = {}
            for candidate, tuples in candidates.items():
                for id, count in tuples.items():
                    values[id] = wiki.sense_probability(candidate, id)
            sorted_values = sorted(values, key=values.get, reverse=True)
            if(len(sorted_values) > 0):
                wiki_id = wiki.pages[sorted_values[0]][0]
                if(wiki_id in tuples_wd):
                    output.write('e{},{}\n'.format(k, tuples_wd[wiki_id]['id']))
        output.close()
```



Cheeky mouse

- page contents used to train word-embedding representation of the (competing) senses
- mention context used to compute word-embedding representation
- cosine similarity used to select the best sense



Zbronk NLP studio

- cascade of heuristics:
 - **dictionary of predefined links**
 - nation heuristic
 - adjectival form heuristic
 - most probable sense of a Wikipedia page
 - most probable lemmatized sense of a Wikipedia page
 - Wikidata search results
- dictionary of predefined links was constructed using the testing dataset



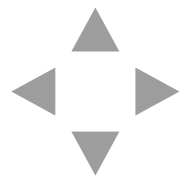
Results

System	Acc
Zbronk NLP studio	91.9
baseline	77.2
Cheeky mouse	26.7



Entity Linking vs Named Entity Recognition

- NER = sequence **tagging**
- EL = mention **disambiguation**
- EL → multiple mentions of the same name are connected
- EL → multiple variants of the same name are connected (*PL, Polska, Rzeczpospolita*)
- EL requires more background knowledge (e.g. a knowledge-base)

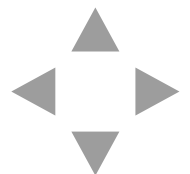


NER in Clarin

XML

Tekst

Woda jest jedną z najpospolitszych substancji we **Wszechświecie**.
Cząsteczka wody jest trzecią najbardziej rozpowszechnioną molekułą w ośrodku międzygwiazdowym, po cząsteczkowym wodorze i tlenku węgla.
Jest również szeroko rozpowszechniona w **Układzie Słonecznym**: stanowi istotny element budowy Ceres i księżyców lodowych krążących wokół planet-olbrzymów, jako domieszka występuje **nam_org** w atmosferach, a przypuszcza się, że duże jej ilości znajdują się we wnętrzach tych planet.
Jako lód występuje także na części planetoid, a zapewne również na obiektach transneptunowych.
Woda jest bardzo rozpowszechniona także na powierzchni **Ziemi**.
Występuje głównie w oceanach, które pokrywają 70,8% powierzchni globu, ale także w rzekach, jeziorach i w postaci stałej w lodowcach.
Część wody znajduje się w atmosferze (chmury, para wodna).
Niektóre związki chemiczne zawierają cząsteczki wody w swojej budowie (hydraty – określa się ją wówczas mianem wody krystalizacyjnej).
Zawartość wody włączonej w strukturę minerałów w płaszczu **Ziemi** może przekraczać łączną zawartość wody w oceanach i innych



Wikipedia

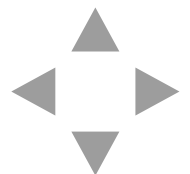
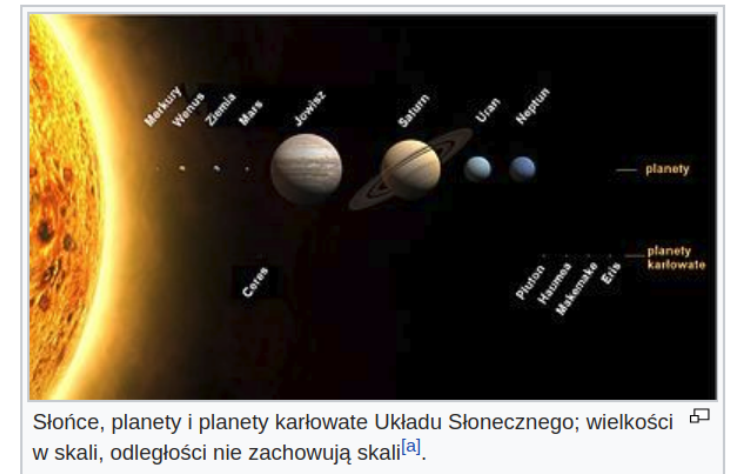
Układ Słoneczny [edytuj]

(Przekierowano z [Układ słoneczny](#))

Układ Słoneczny – [układ planetarny](#) w [galaktyce Drogi Mlecznej](#), składający się ze [Słońca](#) i powiązanych z nim [grawitacyjnie ciał niebieskich](#). Ciała te to osiem [planet](#)^[b], 185 znanych [księżyców](#) [planet](#)^{[c][1]}, pięć [planet karłowatych](#)^[d] i miliardy (a być może nawet biliony)^[2] [małych ciał Układu Słonecznego](#), do których zalicza się [planetoidy](#), [komety](#), [meteoroidy](#) i [pył międzyplanetarny](#).

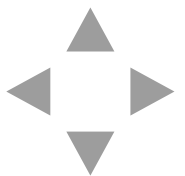
Zbadane regiony Układu Słonecznego zawierają, licząc od Słońca: cztery [planety skaliste](#) ([Merkury](#), [Wenus](#), [Ziemia](#), [Mars](#)), [pas planetoid](#) składający się z małych skalistych ciał, cztery zewnętrzne [gazowe olbrzymy](#) ([Jowisz](#), [Saturn](#), [Uran](#), [Neptun](#)) oraz drugi pas składający się z obiektów skalno-lodowych, tak zwany [Pas Kuipera](#). Za Pasem Kuipera znajduje się [dysk rozproszony](#), dużo dalej [heliopauza](#) i w końcu hipotetyczny [Obłok Oorta](#). Pięć obiektów zaliczonych do klasy planet karłowatych to: [Ceres](#) (największy obiekt w pasie planetoid), [Pluton](#) (do 24 sierpnia 2006 roku uznawany za 9. planetę Układu)^[3], [Haumea](#), [Makemake](#) (drugi co do wielkości obiekt w Pasie Kuipera) i [Eris](#) (największy znany obiekt w dysku rozproszonym).

Sześć z ośmiu planet i trzy z planet karłowatych mają [naturalne satelity](#), zwane księżycami. Każda z planet zewnętrznych jest otoczona [pierścieniami](#) złożonymi z ziaren lodowych i pyłu kosmicznego. Wszystkie planety, z wyjątkiem Ziemi i Urana (który zawdzięcza nazwę greckiemu bóstwu [Uranosowi](#)), noszą imiona bóstw z [mitologii rzymskiej](#).



Entity Linking vs Sense disambiguation (i.e. WordNet)

- EL concentrates on Named Entities and other name-like expressions (e.g. names of species)
- mostly applies to nouns and nominal expressions
- EL does not capture sense nuances
- distribution of mentions is highly skewed



Słownosieć

Typy

[układ planetarny 1](#) 

planety i inne ciała niebieskie, krążące wokó...

Meronimy (części)

[plutoid 1](#) 

domena: obiekty naturalne

[Jowisz 1](#) 

domena: obiekty naturalne

[Wenus 1](#) 

domena: obiekty naturalne

[Saturn 1](#) 

domena: miejsca i umiejscowienie

... Pokaż pozostałe 6 połączeń

Synonimy międzyjęzykowe

[solar system 1](#) 

the sun with the celestial bodies that revolv...

RZECZOWNIK

Układ Słoneczny 1

układ planetarny składający się ze Słońca i powiązanych z nim grawitacyjnie ciał niebieskich

DOMENA

sytuacje statyczne (stany)

ŹRÓDŁO

Słownosieć

HIPERONIMY

[Pokaż ścieżkę do najwyższego hiperonimu`](#)

GRAPH

[Wyświetl wizualizację graficzną](#)

Fuzzynimy synsetów

[pozasłoneczny 1](#) 

taki, który znajduje się poza Układem Słone...

Holonimy (części)

[kosmos 1](#) 

wszechświat, wszystko, co fizycznie istnieje; ...

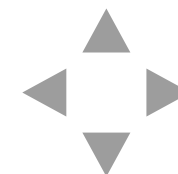
Meronimy (elementy kolekcji)

[planeta wewnętrzna 1](#) 

planeta leżąca pomiędzy Słońcem i pasem p...

[planeta zewnętrzna 1](#) 

w Układzie Słonecznym planeta leżąca poza ...



Wikidata



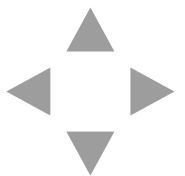
Training dataset - Polish Wikipedia

- tokenized, sentence-split, tagged, lemmatized Wikipedia
- list of selected Wikidata types
- Wikidata items
- some data extracted from Wikipedia:
 - pages
 - page categories
 - links between pages
 - redirects
 - translations
 - templates
 - and more...



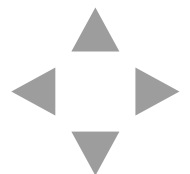
Testing dataset - subcorpus of NKJP

- entity types - "names"
- mention scope - the longest possible
- semantic category of mentions
- adjectives
- specific vs. general links
- ~ 40k tokens
- ~ 4k mentions



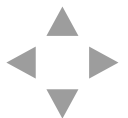
Wikipedia text

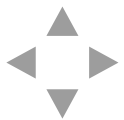
2	AWK	AWK	0	subst:sg:nom:f	_	_	
2	-	-	1	interp	_	_	
2	interpretowany	interpretować	1	ppas:sg:nom:m3:imperf:aff			_
-							
2	język	język	1	subst:sg:nom:m3	język programowania	Q9143	
2	programowania	programować	1	ger:sg:gen:n:imperf:aff	język programow		
ania	Q9143						
2	,	,	0	interp	_	_	
2	którego	który	1	adj:sg:gen:m3:pos			
2	główną	główny	1	adj:sg:inst:f:pos			
2	funkcją	funkcja	1	subst:sg:inst:f	_		
2	jest	być	1	fin:sg:ter:imperf			
2	wyszukiwanie	wyszukiwać	1	ger:sg:nom:n:imperf:aff	_		
2	i	i	1	conj	_		
2	przetwarzanie	przetwarzać	1	ger:sg:nom:n:imperf:aff	_		
2	wzorców	wzorzec	1	subst:pl:gen:m3	_		
2	w	w	1	prep:loc:nwok	_		
2	plikach	plika	1	subst:pl:loc:f	plik		
2	lub	lub	1	conj	_		
2	strumieniach	strumień	1	subst:pl:loc:m3	_		



Wikidata types

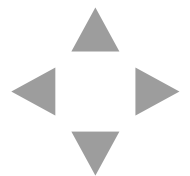
- human (<https://www.wikidata.org/wiki/Q5>)
- geographic location (<https://www.wikidata.org/wiki/Q2221906>)
- academic discipline (<https://www.wikidata.org/wiki/Q11862829>)
- anatomical structure (<https://www.wikidata.org/wiki/Q4936952>)
- occupation (<https://www.wikidata.org/wiki/Q12737077>)
- vehicle model (<https://www.wikidata.org/wiki/Q29048322>)
- construction (<https://www.wikidata.org/wiki/Q811430>)
- written work (<https://www.wikidata.org/wiki/Q47461344>)
- astronomical body (<https://www.wikidata.org/wiki/Q6999>)
- etc.





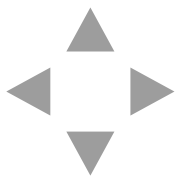
Wikidata items

```
{ "id": "Q31",  
  "labels": { "en": "Belgium", "pl": "Belgia" },  
  "wiki": { "en": "Belgium", "pl": "Belgia" },  
  "P31": [ "Q3624078", "Q43702", "Q6256", "Q20181813" ] }  
{ "id": "Q8",  
  "labels": { "en": "happiness", "pl": "szczęście" },  
  "wiki": { "en": "Happiness", "pl": "Szczęście" },  
  "P31": [ "Q331769", "Q60539479" ], "P279": [ "Q16748867" ] }  
{ "id": "Q23",  
  "labels": { "en": "George Washington", "pl": "George Washington" },  
  "wiki": { "en": "George Washington", "pl": "George Washington" },  
  "P31": [ "Q5" ] }  
{ "id": "Q42",  
  "labels": { "en": "Douglas Adams", "pl": "Douglas Adams" },  
  "wiki": { "en": "Douglas Adams", "pl": "Douglas Adams" },  
  "P31": [ "Q5" ] }  
...
```



Wikipedia pages

2,AWK,0,0,18337
4,Alergologia,0,0,825
6,ASCII,0,0,32937
7,Atom,0,0,105440
8,Aksjomat,0,0,10797
10,Arytmetyka,0,0,6328
12,Alkeny,0,0,5085
13,ActiveX,0,0,2743
14,Interfejs programowania aplikacji,0,0,1517
15,AmigaOS,0,0,8999
16,Association for Computing Machinery,0,0,6926



Page parents

2,Języki skryptowe

4,Alergologia,Specjalności lekarskie

6,Formaty plików komputerowych,Kodowania znaków

7,Artykuły na medal,Atomy,Wyróżnione artykuły

8,Logika matematyczna

10,Artykuły z brakującymi przypisami,Artykuły z wyrażeniami zwodniczymi od 2013-12,Arytmetyka

11,Artykuły w obszarze zainteresowań Projektu Chemia

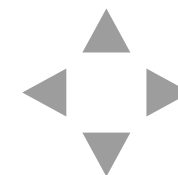
12,Alkeny

13,Internet Explorer,Multimedia,Windows

14,Interfejsy programistyczne

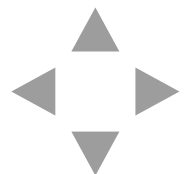
15,Amiga,Systemy operacyjne

16,Międzynarodowe organizacje pozarządowe



Links between pages

2,1977,2GL,3GL,4GL,5GL,ABAP,ALGOL,APL (język programowania),...
4,Alergen,Alergia,Anestezjologia,Angiochirurgia,...
6,(3568) ASCII,ACK,ASCII-Art,American National Standards Institute,...
7,1624,1964,Akcelerator cząstek,Aksjon,Albert Einstein,Alchemia,...
8,A priori,Aksjomatyka Hilberta,Alfred Tarski,Arytmetyka,...
10,Algebra,Algebra abstrakcyjna,Algebra liniowa,...
12,"1,3-Butadien",Addycja,Alkany,Alkiny,Alkohole,...
13,.NET Remoting,.Net Framework,ADO.NET,ASP.NET,Active Directory,...
14,Biblioteka Kongresu,Biblioteka Narodowa Francji,...
15,1 listopada,1985,1985 w informatyce,24 grudnia,...
16,Audio Compression Manager,Azja,Columbia University,Dyrektor general...



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

TABLE IV
 ARTICLE PROBABILITIES OF *zamek* LINK ACCORDING TO THE
 POLISH WIKIPEDIA. (THE ARTICLES WITH LESS THAN 2 LINK
 OCCURRENCES ARE NOT SHOWN.)

Wikipedia article	# of occurrences	P
Zamek (<i>castle</i>)	425	0.698
Zamek (<i>of rifle</i>)	60	0.099
Zamek w Bydgoszczy	28	0.046
Zamek w Bolkowie	11	0.018
Zamek w Szydłowcu	5	0.008
Zamek Świny	4	0.007
Zamek Kapituły Warmińskiej ...	4	0.007
Zamek (<i>for locking</i>)	4	0.007
Zamek Królewski w Poznaniu	3	0.005
Zamek w Malborku	2	0.003
Zamek w Rzeszowie	2	0.003
Zamek w Suchej Beskidzkiej	2	0.003
Zamek w Kowalu	2	0.003
Zamek Królewski na Wawelu	2	0.003
Zamek w Edynburgu	2	0.003



TABLE I
LINKS REFERRING TO *Poland* IN THE POLISH WIKIPEDIA. (THE
LINKS WITH LESS THAN 100 OCCURRENCES ARE NOT SHOWN.)

Link name	# of occurrences
Polsce	74196
polski	11564
Polski	5528
Polska	3271
polska	2234
Polskę	763
Polską	647
polskiego	605
polskiej	510
polsko	458
polskich	432
polskie	295
polskim	246
polską	130
Polak	108
RP	101
Polaków	100
Polacy	100



TABLE III
THE 15 ARTICLES MOST RELATED TO *Warszawa* THAT LINK TO
OR ARE LINKED FROM THIS ARTICLE ACCORDING TO sr_J
MEASURE. (NOTE: $\log_{10}$ IS USED AS THE $\log$ FUNCTION)

Wikipedia article	sr_J
Kraków	0.5054
Uniwersytet Warszawski	0.4853
Łódź	0.4753
Poznań	0.4740
II wojna światowa	0.4707
Wrocław	0.4595
Powstanie warszawskie	0.4561
1945	0.4449
Lublin	0.4446
Lwów	0.4440
Gdańsk	0.4423
Polska	0.4409
Paryż	0.4327
2007	0.4302
2006	0.4285



TABLE V

NUMBER OF OCCURRENCES AS LINK, TOTAL NUMBER OF
OCCURRENCES AND *link probability* MEASURE FOR SEVERAL
TERMS USED AS LINKS.

Term	Category	# as link	# all	P
Jerzy Buzek	<i>person</i>	108	236	0.458
Katowicach Giszowcu	<i>district</i>	1	4	0.250
Kraków	<i>city</i>	2756	29305	0.094
Miechów	<i>city</i>	93	359	0.259
Polska	<i>country</i>	3720	32164	0.116
Utraty	<i>river</i>	33	82	0.402
internetowy	<i>adj</i>	4	1281	0.003
literatury	<i>noun</i>	291	10472	0.028
małego	<i>adj</i>	2	3083	0.001
miastem	<i>noun</i>	121	8122	0.015
nie	<i>part/pron</i>	3	714190	0.000
polskiego harcerstwa	<i>adj+noun</i>	1	47	0.021
sascy	<i>adj</i>	5	46	0.109
ulica	<i>noun</i>	149	5776	0.026
zakopiański	<i>adj</i>	4	51	0.078
zamek	<i>noun</i>	609	8787	0.069



TABLE VIII
PERFORMANCE OF THE ORIGINAL ALGORITHM AND ITS
IMPROVED VERSION FOR THE POLISH WIKIPEDIA. THE
THRESHOLD FOR THE MINIMAL NUMBER OF LINKS IN THE
TESTING ARTICLES IS 5 AND THE MAXIMAL IS 100.

	precision	recall	F ₁ -measure
Random sense	39.7	26.4	31.7
Random sense with $P > 0.5\%$	47.0	47.3	47.2
Most common sense	81.6	82.2	81.9
sr_G	82.5	83.5	83.0
sr_G + new features	84.9	83.2	84.0
sr_J	85.4	89.8	87.6
sr_J + new features	90.4	93.0	91.7



SOTA algorithms

- classifier based on various features (most approaches)
- entity type prediction



OpenAI

For each possible type a different binary classifier is trained: here we decide whether a token refers to an "Animal" article.

