sages

Rozpoznawanie jednostek nazewniczych

(*ang.* named entity recognition — NER)

Jednostki nazewnicze

Jednostki nazewnicze to nazwy (fragmenty tekstu) wskazujące na **konkretne obiekty** (osoby, miejsca, budowle, organizacje, itp.)

Henryk Sienkiewicz urodził się we wsi **Wola Okrzejska**.

https://pl.wikipedia.org/wiki/Henryk_Sienkiewicz

https://pl.wikipedia.org/wiki/Wola_Okrzejska

Nazwa pospolita vs. nazwa własna

W przetwarzaniu języka naturalnego **jednostki nazewnicze** to przede wszystkich **nazwy własne**. W odróżnieniu od nazw pospolitych, **nazwy własne** odnoszą się do **konkretnego obiektu**, a nie **klasy obiektów,** np.:

- „Jutro mam wolne i planuję wybrać się do **muzeum**."
- „W wakacje jadę do Paryża i w końcu będę mógł zwiedzić **Luwr**."

https://pl.wikipedia.org/wiki/Luwr

Zjawiska związane z jednostkami nazewniczymi

Deskrypcje określone

Deskrypcje określone to frazy lub zdania, które jednoznacznie wskazują konkretny obiekt poprzez wskazanie **relacji względem innych obiektów**.

Np.:

- Autor powieści „**Ogniem i mieczem**” → **Henryk Sienkiewicz**
- Stolica **Polski** do 1795 → **Kraków**

Deskrypcje określone nie są traktowane jako jednostki nazewnicze.

Metonimia

Metonimia to figura stylistyczna polegająca na zastąpieniu jednej nazwy inną, będącej w relacji części do całości.

Np.:

- „W 2014 **Brazylia** przegrała z **Niemcami** 1:7.”
  - **Brazylia** — reprezentacja Brazylii w piłce nożnej,
  - **Niemcy** — reprezentacja Niemiec w piłce nożnej,
- „Podejrzany odjechał **Fordem** z miejsca zdarzenia.”
  - **Ford** — samochód marki Ford

Najczęściej identyfikowane jest pierwotne znaczenie nazwy własnej.

Niejednoznaczność pomiędzy kategoriami

Paryż (ujednoznacznienie) [edytuj]

To jest strona ujednoznaczniająca. Poniżej znajdują się różne znaczenia hasła: **Paryż**.

- **Paryż** – miasto, stolica Francji

Spis treści [ukryj]

Miejscowości i ich części w Polsce [edytuj | edytuj kod]

Wg TERYT jest ich 11, w tym 1 podstawowa

- Paryż – wieś w woj. kujawsko-pomorskim, w pow. żnińskim, w gminie Żnin
- Paryż – przysiółek wsi Nowa Góra w woj. małopolskim, w pow. krakowskim, w gminie Krzeszowice
- Paryż – część wsi Ochojno w woj. małopolskim, w pow. krakowskim, w gminie Świątniki Górne
- Paryż – część wsi Wola Piskulina w woj. małopolskim, w pow. nowosądeckim, w gminie Łącko
- Paryż – część wsi Janików w woj. mazowieckim, w pow. kozienickim, w gminie Kozienice
- Paryż – część wsi Krępna w woj. opolskim, w pow. krapkowickim, w gminie Zdzieszowice
- Paryż – przysiółek wsi Domaradz w woj. opolskim, w pow. namysłowskim, w gminie Pokój
- Paryż – część wsi Łęki Dolne w woj. podkarpackim, w pow. dębickim, w gminie Pilzno
- Paryż – część wsi Długołęka w woj. podlaskim, w pow. monieckim, w gminie Krypno
- Paryż – część wsi Albertów w woj. śląskim, w pow. kłobuckim, w gminie Lipie
- Paryż – część wsi Pawełki w woj. śląskim, w pow. lublinieckim, w gminie Kochanowice

Miejscowość na Białorusi [edytuj | edytuj kod]

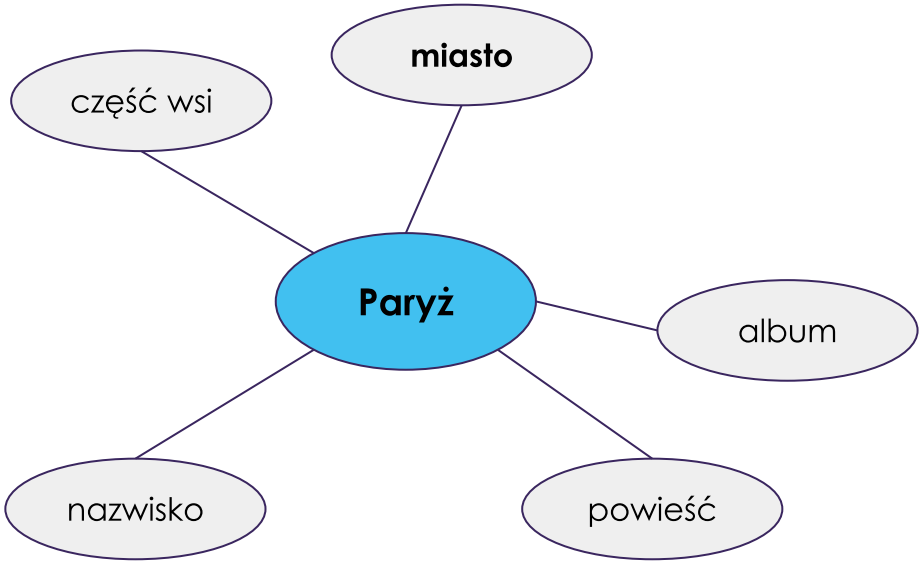
- Paryż – agromiasteczko w obw. witebskim, w rej. postawskim, w sielsowiecie Kozłowszczyzna

Inne [edytuj | edytuj kod]

- Paryż – polskie nazwisko
- *Paryż* – powieść Émila Zoli z cyklu *Trzy miasta*

Zobacz też [edytuj | edytuj kod]

- Paryż 2010: Wielka powódź – francuski film katastroficzny z roku 2006
- Paryż (Moskwa) – album łódzkiej grupy Rezerwat



https://pl.wikipedia.org/wiki/Pary%C5%BC_(ujednoznacznienie)

Niejednoznaczność o obrębie kategorii

Jan Nowak [edytuj]

To jest strona ujednoznaczniająca. Poniżej znajdują się różne znaczenia hasła: **Jan Nowak**.

- Jan Nowak (1906–1987) – polski żołnierz KOP, ZWZ, AK, KWP[1]
- Jan Nowak (1880–1940) – polski geolog, paleontolog
- Jan Nowak (1887–1959) – mierniczy; działacz narodowy; księgarz; pisarz
- Jan Nowak (1895–1940) – kapitan piechoty Wojska Polskiego zamordowany w Katyniu, oficer Batalionu KOP „Łużki"
- Jan Nowak-Jeziorański (1914–2005) – polski polityk, dziennikarz, żołnierz AK
- Jan Nowak (1928–2018) – polski działacz sportowy, członek honorowy Polskiego Związku Piłki Nożnej
- Jan Nowak (1948–2007) – polski trener lekkoatletyczny
- Jan Nowak (ur. 1955) – polski żużlowiec
- Jan Nowak (ur. 1951) – prezes Urzędu Ochrony Danych Osobowych
- Jan Wiktor Nowak (1931–2002) – polski duchowny katolicki, biskup siedlecki

https://pl.wikipedia.org/wiki/Jan_Nowak

Koreferencja

Koreferencja to zabieg stylistyczny polegający na odniesieniu się do wcześniej wspomnianego obiektu poprzez użycie innego środka niż powtórzenie pierwszego odniesienia.

„**Henryk Sienkiewicz** jest autorem ponad 20 nowel. (...) **Pisarz** jest także autorem powieści oraz reportaży.”

Pisarz → Henryk Sienkiewicz

Odniesienia koreferencyjne inne niż nazwy własne nie są rozpatrywane w ramach zadania wykrywania jednostek nazewniczych.

Główne podzadania

Główne podzadania

1. Rozpoznanie jednostek w tekście
2. Klasyfikacja jednostek
3. Lematyzacja
4. Ujednoznacznianie

Przykładowy tekst

Hristo Zlatanov (ur. 21 kwietnia 1976 roku w Sofii) – siatkarz reprezentacji Włoch bułgarskiego pochodzenia. Obecnie występuje w Serie A, w drużynie Copra Berni Piacenza. Gra na pozycji przyjmującego.

1. Rozpoznanie jednostek w tekście

- typowe podejście do rozpoznawania wystąpień jednostek wykorzystuje metody do **tagowania sekwencji** — Naive Bayes, CRF, BiLSTM,
- tekst reprezentowany jest jako sekwencja tokenów,
- do każdego **tokenu** zostaje przypisana **etykieta** reprezentująca wystąpienie lub brak jednostki,
- różne sposoby kodowania IOB, IOB2, BIEOS

Token	IOB	IOB2	BIEOS
Jan	I-NAM	B-NAM	B-NAM
Nowak	I-NAM	I-NAM	E-NAM
z	O	O	O
Warszawy	I-NAM	B-NAM	S-NAM

→ Jan Nowak Warszawy

1. Rozpoznanie jednostek w tekście

„**Hristo Zlatanov** (ur. 21 kwietnia 1976 roku w **Sofii**) – siatkarz reprezentacji **Włoch** bułgarskiego pochodzenia. Obecnie występuje w **Serie A**, w drużynie **Copra Berni Piacenza**. Gra na pozycji przyjmującego.”

Forma tekstowa
Hristo Zlatanov
Sofii
Włoch
Serie A
Copra Berni Piacenza

2. Klasyfikacja jednostek

- typowym podejściem jest wykonanie klasyfikacji jednostek już na **etapie ich rozpoznawania**,

  {O, B-NAM, I-NAM} → {O, B-PER, I-PER, B-LOC, I-LOC, ...}

- dla drobnoziarnistej kategoryzacji liczba możliwych etykiet dla tokenów może być relatywnie duża, przez co **niektóre metody stają się zbyt złożone obliczeniowo**, np. CRF,
- alternatywnym podejściem jest budowa **osobnego klasyfikatora do kategoryzacji jednostek** — na bazie kontekstu budowany jest wektor cech, który przepuszczany jest przez klasyfikator,

2. Klasyfikacja jednostek

„**Hristo Zlatanov** (ur. 21 kwietnia 1976 roku w **Sofii**) – siatkarz reprezentacji **Włoch** bułgarskiego pochodzenia. Obecnie występuje w **Serie A**, w drużynie **Copra Berni Piacenza**. Gra na pozycji przyjmującego."

Forma tekstowa	Kategoria
Hristo Zlatanov	PER / osoba
Sofii	LOC / miasto
Włoch	LOC / państwo
Serie A	EVE / wydarzenie
Copra Berni Piacenza	ORG / drużyna

3. Lematyzacja

Lematyzacja jednostek nazewniczych różni się od lematyzacji tokenów:

1. Nazwy wielowyrazowe

   (do) Brytyjskiej Izby Handlowej → **Brytyjska Izba Handlowa** (Brytyjski Izba Handlowy)

2. Zależna od kategorii jednostki

   (utwór) Mickiewicza [osoba] → **Mickiewicz**
   (ul.) Mickiewicza [ulica] → **Mickiewicza**

3. Lematyzacja

„**Hristo Zlatanov** (ur. 21 kwietnia 1976 roku w **Sofii**) – siatkarz reprezentacji **Włoch** bułgarskiego pochodzenia. Obecnie występuje w **Serie A**, w drużynie **Copra Berni Piacenza**. Gra na pozycji przyjmującego."

Forma tekstowa	Kategoria	Lemat
Hristo Zlatanov	osoba	Hristo Zlatanov
Sofii	miasto	Sofia
Włoch	państwo	Włochy
Serie A	wydarzenie	Serie A
Copra Berni Piacenza	drużyna	Copra Berni Piacenza

4. Ujednoznacznianie

- wymaga bazy wiedzy, która będzie podstawą ujednoznaczniania jednostek, np. **Wikipedia** — link do strony będzie identyfikatorem obiektu,
- dla toponimów alternatywą może być np. baza **Geonames**,
- **zadanie o wysokim poziomie złożoności** — główną trudnością jest **dezambiguacja** (ujednoznacznienie) jednostek o takich samych nazwach w obrębie tych samych kategorii, np. **Jan Nowak** — żołnierz, geolog, kapitan, trener, …?
- dodatkową trudnością może być **niekompletność bazy wiedzy**, np. nie każdy Jan Nowak posiada swoją stronę w Wikipedii,

4. Ujednoznacznianie

„**Hristo Zlatanov** (ur. 21 kwietnia 1976 roku w **Sofii**) – siatkarz reprezentacji **Włoch** bułgarskiego pochodzenia. Obecnie występuje w **Serie A**, w drużynie **Copra Berni Piacenza**. Gra na pozycji przyjmującego."

Forma tekstowa	Kategoria	Lemat	Obiekt
Hristo Zlatanov	osoba	Hristo Zlatanov	https://pl.wikipedia.org/wiki/Hristo_Zlatanov
Sofii	miasto	Sofia	https://pl.wikipedia.org/wiki/Sofia
Włoch	państwo	Włochy	https://pl.wikipedia.org/wiki/W%C5%82ochy
Serie A	wydarzenie	Serie A	https://pl.wikipedia.org/wiki/Serie_A
Copra Berni Piacenza	drużyna	Copra Berni Piacenza	https://pl.wikipedia.org/wiki/Pallavolo_Piacenza

Klasyfikacje i kategorie jednostek nazewniczych

Założenia

1. Nie ma jednego, ugruntowanego zestawu kategorii i reguł anotacji, przez co występują różnice między **schematami anotacji**.

2. Główne różnice schematów anotacji:
   a. **liczba i definicja kategorii** jednostek,
   b. **granice anotacji** — np. czy skrót "ul." powinien być częścią nazwy czy też nie,
   c. **ziarnistość** — główne kategorie i podkategorie,
   d. **płaska lub zagnieżdżona struktura —** *Uniwersytet im. Adama Mickiewicza*,
   e. **ciągłość anotacji —** *Adam i Ewa Nowak*.

CoNLL 2003 (j. angielski)

https://www.aclweb.org/anthology/W03-0419.pdf

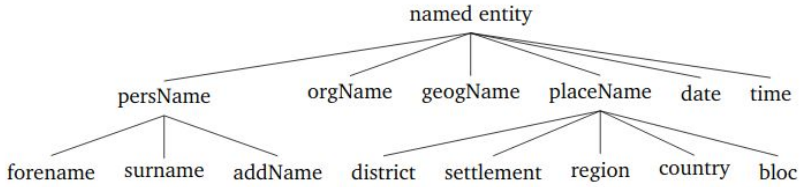
Kategorie:

- **PER** — osoby (PERSON),
- **LOC** — miejsca (LOCATION),
- **ORG** — organizacje (ORGANIZATION),
- **MISC** — pozostałe nazwy, które nie należą do PER, LOC oraz ORG.

NKJP (j. polski)

Narodowy Korpus Języka Polskiego

http://nkjp.pl



Charakterystyka:

1. **4 głównych** oraz **10 szczegółowych** kategorii oraz **wyrażenia temporalne** (date, time),
2. Zawiera wiele poziomów zagnieżdżenia, w tym anotacje tych samych kategorii,
3. Anotacje nieciągłe (ok. 1%)
4. Korpus zawiera ponad **87 tys.** jednostek.

Główne kategorie

1. persName
2. orgName
3. geogName
4. placeName
5. date
6. time

KPWr (j. polski)

Korpus Politechniki Wrocławskiej

https://clarin-pl.eu/dspace/handle/11321/294

Charakterystyka:

1. **8 głównych** oraz **ponad 80 szczegółowych** kategorii,
2. Tylko anotacje ciągłe,
3. Zawiera zagnieżdżenia,
4. Ponad **34 tys.** anotacji w **KPWr** oraz **15 tys.** w korpusie wiadomości gospodarczych (**CEN**).

Główne kategorie

1. nam_adj
2. nam_eve
3. nam_fac
4. nam_liv
5. nam_loc
6. nam_org
7. nam_oth
8. nam_pro

KPWr (j. polski)

Kategorie szczegółowe **nam_loc**

- nam_loc_astronomical — naturalne ciała niebieskie,
- nam_loc_country_region — regiony geograficzne w obrębie kraju,
- nam_loc_gpe — jednostki geopolityczne,
  - nam_loc_gpe_admin — podział administracyjny,
  - nam_loc_gpe_city — miasta, wioski,
  - …
- nam_loc_hydronym — naturalne obiekty wodne,
  - nam_loc_hydronym_river — rzeki,
  - nam_loc_hydronym_lake — jeziora,
  - …
- nam_loc_land — ziemne obiekty geograficzne
  - nam_loc_land_cape — przylądki
  - nam_loc_land_continent — kontynenty
  - ...

KPWr (j. polski)

Kategorie szczegółowe **nam_fac**

- nam_fac_bridge — mosty,
- nam_fac_geo — kina, restauracje, sklepy, itp,
  - nam_fac_geo_stop — przystanki autobusowe, tramwajowe i kolejowe,
  - ...
- nam_fac_park — parki miejskie, krajobrazowe i przyrodnicze,
- nam_fac_road — ulice, drogi i autostrady,
- nam_fac_square — place i rynki,
- nam_fac_system — systemy posiadające własną infrastrukturę,