Analiza danych niestrukturalnych techniki i metody NLP

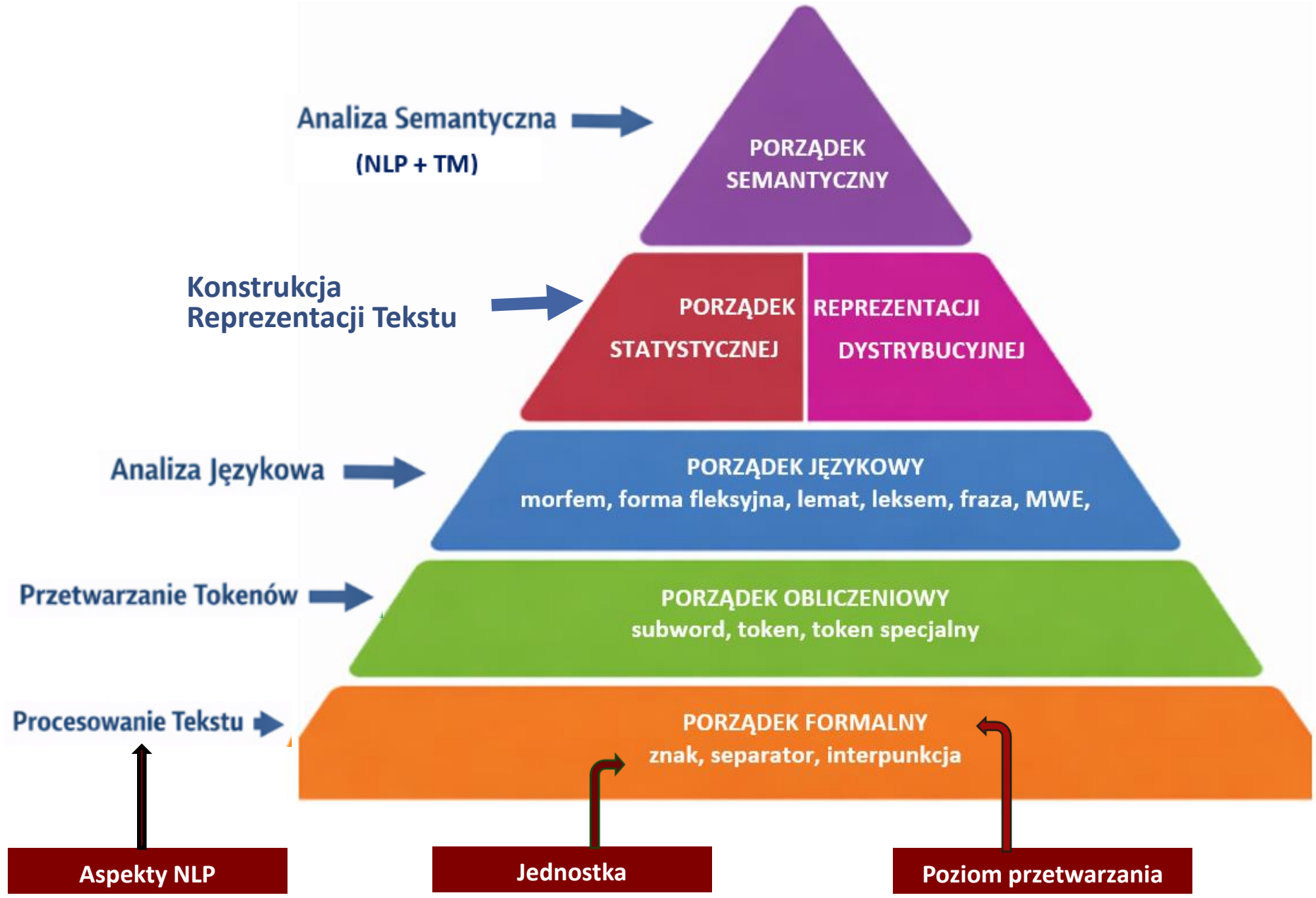
Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Przetwarzania języka naturalnego - NLP

Natural Language Processing
Jest dziedziną sztucznej inteligencji (AI – *Artificial Intelligence*) i lingwistyki komputerowej (*Computational Linguistics*), której podstawowym celem jest umożliwienie maszynom interakcji (komunikacji) z ludźmi w sposób zrozumiały i naturalny, co jest fundamentem takich technologii jak chatboty, asystenci głosowi, systemy rekomendacyjne czy wyszukiwarki inteligentne.

Zajmuje się projektowaniem systemów i algorytmów pozwalających komputerom rozumieć, analizować, interpretować i generować język naturalny, czyli taki, jakim posługują się ludzie.

Hierarchia przetwarzania danych w NLP



Znak – jednostka zapisu
Wyraz - jednostka powierzchniowa tekstu

- **Znak**

  Pojedynczy symbol tekstowy (character) – najmniejsza jednostka zapisu, z której buduje się inne jednostki w NLP (począwszy od najniższego): subwordy, tokeny, słowa, frazy / MWE i zdanie.

- **Wyraz**

  Dosłowna, surowa postać elementu tekstu, konkretny zapis graficzny, widoczny w zdaniu. Jednostka, którą można deterministycznie wydzielić z tekstu za pomocą separatorów bez konieczności rozumienia znaczenia czy funkcji gramatycznej.

- **Typowe separatory**

  Białe znaki, znaki interpunkcyjne, znaki specjalne, znaki Unicode.

PL	Nie mogę się zalogować. Nie mogę się zalogować
EN	I'm learning AI, and it's great! I’m learning AI and it’s great

Operacje porządku formalnego

Czyszczenie tekstu→ Normalizacja → Segmentacja

- **Czyszczenie tekstu**

  Proces usuwania "szumu" oraz zbędnych informacji z surowego tekstu.

  Usuwanie: tagów HTMK i metadanych, adresów URL i e-maili, duplikatów, stop-words; obsługa specyficznych znaków i symboli; korekta błędów i literówek

- **Normalizacja**

  Przekształcanie tekstu w formę ujednoliconą (kanoniczną) w celu zredukowania niepotrzebnych różnic w zapisie; rozpoznanie, że różne ciągi znaków ("Idę", "idę", "IDE") odnoszą się do tego samego pojęcia.

  Ujednolicenie wielkości liter (Kot → kot), zastępowanie różnych znaków interpunkcyjnych znakiem standardowym („ ; ” ; " ; « ; » ’ → ,), normalizacja liczb (10 000 → 10000), normalizacja emotikonów i emoji (🙂 → <EMO_POS>), korekta kodowania znaków (Unicode) (Ã³ → ó)

- **Segmentacja**

  Dzielenie surowego ciągu znaków (tekstu) na mniejsze jednostki - na sensowne fragmenty dla dalszej analizy, dostarczając: zdania, linie, paragrafu (akapity). Często używa się: regex + znaki interpunkcyjne + spacje. Operuje wyłącznie na znakach, nie interpretuje znaczenia słów.

Jednostki porządku obliczeniowego

Token

Jednostka powierzchniowa w sensie technicznym, na której operują algorytmy NLP, ale jego znaczenie zależy od dalszego przetwarzania; w praktyce token może odpowiadać wyrazowi, liczbie, symbolowi, emotikonowi lub subwordowi.

Subword

Podwyraz, token subwordowy - jednostka tekstu mniejsza niż cały wyraz, ale większy niż pojedynczy znak; to znaczący fragment słowa (przedrostek, przyrostek, rdzeń lub częste łączenie liter). Granice subwordów są statystyczne, a nie lingwistyczne.

Głowna cecha: zdolność do ponownego wykorzystania w wielu różnych wyrazach, co pozwala modelom językowym uczyć się reprezentacji wspólnych rdzeni, prefiksów i sufiksów.

Token specjalny

Specjalny znak lub ciąg znaków (element techniczny) używany w przetwarzaniu tekstu przez modele językowe lub systemy NLP, który nie reprezentuje zwykłego słowa, lecz pełni funkcję techniczną lub sterującą (oznakowanie początku, końca sekwencji, nieznanego słowa).

Operacje porządku obliczeniowego – Tokenizacja

Tokenizacja → Embeddingi → Modelowanie języka

Tokenizacja - dzielenie tekstu na jednostki obliczeniowe (tokeny, subwordy) dla modeli językowych.

PL	Nie mogę się zalogować 😂🔥. [Nie] [mogę] [się] [zalogować] [😂] [🔥.]
EN	I'm learning AI, and it's great 👍! [I'm] [learning] [AI] [and] [it's] [great] [👍!] [I] ['m] [learning] [AI] [,] [and] [it] ['s] [great] [👍] [!]

Podejście **subwordowe** - dzielenie na krótsze, statystycznie istotne fragmenty. Subwordy można wielokrotnie wykorzystywać w wielu różnych wyrazach, co pozwala modelom językowym uczyć się reprezentacji wspólnych rdzeni, prefiksów i sufiksów, osiągając lepszą generalizację (zwłaszcza w językach o bogatej morfologii).

PL	nieprzetłumaczalny nie + przetłumacz + alny nie + prze + tłumacz + al + ny ni + e + prze + tłum + acz + al + ny
EN	unbelievability A: un + believ + ability B: un + believ + abil + ity C: un + beli + ev + abil + ity

Operacje porządku obliczeniowego – Embedding

Tokenizacja → Embeddingi → Modelowanie języka

Embeddingi (wektoryzacja/osadzanie słów) - tworzenie wektorowych reprezentacji jednostek językowych w przestrzeni wielowymiarowej, które odzwierciedlają znaczenie i kontekst tych jednostek. Słowo, fraza lub zdanie jest przedstawione jako gęsty wektor liczb rzeczywistych.
Embedding nie patrzy na definicję słowa, lecz na konteksty, w których słowo występuje w tekstach.

samochód	(0.82, 0.11, 0.54, **0.33**, 0.76)
auto	(0.80, 0.10, 0.56, **0.35**, 0.74)
pojazd	(0.78, 0.15, 0.52, **0.30**, 0.72)
ciężarówka	(0.75, 0.20, 0.50, **0.40**, 0.70)
samolot	(0.79, 0.14, 0.48, **0.60**, 0.73)

Samolot ma wektor znajduje się w tej samej części przestrzeni semantycznej (środek transportu).
Pewna wartość (np. 0.60 w 4. wymiarze) może symbolizować cechę odróżniającą (transport powietrzny zamiast lądowego)

Wektory odzwierciedlają znaczenie i kontekst: podobne słowa mają bliskie wektory (punkty w przestrzenie wielowymiarowej są blisko siebie), różne słowa są reprezentowane przez odległe od siebie punkty.

Operacje porządku obliczeniowego – Modelowanie języka

Tokenizacja → Embeddingi → Modelowanie języka

Model językowy - system matematyczny (w tym statystyczny), którego zadaniem jest estymacja/obliczanie prawdopodobieństwa wystąpienia ciągu słów.
W sensie informatycznym to algorytm, który na podstawie dotychczasowego tekstu przewiduje, co powinno pojawić się dalej.

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1})$$

Kot siedzi na stole	wyższe prawdopodobieństwo
Stole kot na siedzi	niższe prawdopodobieństwo

Modelowanie języka polega na estymowaniu prawdopodobieństwa wystąpienia danej sekwencji słów w danym języku naturalnym, czyli realizacji ww. algorytmu

Model nie „rozumie” świata tak jak ludzie, ale na podstawie miliardów przeczytanych zdań „wie”, że po słowach „Ala ma...” z dużym prawdopodobieństwem nastąpi słowo „kota”, a nie „lodówkę”.

Morfem – najmniejsza jednostka znaczeniowa

Najmniejsza jednostka języka niosąca znaczenie leksykalne lub funkcję gramatyczną. Morfemy mają podstawowe znaczenie dla zrozumienia struktury słów i analizy morfologicznej, szczególnie w językach fleksyjnych.

PL: **niezalogowaliśmy**	
nie-	prefiks (negacji; zaprzeczenie lub przeciwieństwo)
za-	prefiks słowotwórczy (tworzenie czasownika dokonany; rozpoczęcie lub wykonanie czynności
log	rdzeń leksykalny (podstawowe znaczenie)
-owa-	sufiks słowotwórczy (tworzy czasownik od rzeczownika)
-li-	sufiks czasu przeszłego (1 os. l. mn)
-śmy	sufiks fleksyjny (czas przeszły, l. mn.)

EN: **unhappiness**	
un-	prefiks (negacji)
happy	rdzeń leksykalny
ness	sufiks (tworzenie rzeczownika od przymiotnika)

Forma fleksyjna, leksem, lemat

Forma fleksyjna - postać wyrazu powstała w wyniku odmiany, wyrażająca określone cechy gramatyczne (np. przypadek, liczbę, rodzaj, czas).

PL: kot → kot, kota, kotu, kotem, koty

EN: mouse → mouse, mice, mouse's, mice's

Leksem (słowo) - podstawowa jednostka słownikowa (leksykalna) języka niosąca jedno znaczenie lub zestaw znaczeń. Obejmuje wszystkie formy fleksyjne danego słowa (całą rodzinę form gramatycznych). Pisany wielką literą.

PL: **IŚĆ**	EN: **BE**
idę	am
szedłem	are
pójdę	were
idźcie	been

Lemat - jednostka operacyjna (techniczna); unikalna forma tekstowa (ciąg znaków, string), która służy do indeksowania, reprezentowania leksemu w słowniku lub w systemie NLP. Przykład: iść, be

System NLP operuje na lemacie jako ciągu znaków, ale jego interpretacja semantyczna odnosi się do leksemu.

Fraza, jednostka wielowyrazowa

Fraza - grupa słów funkcjonująca w zdaniu jako spójna jednostka składniowa i znaczeniowa (tworzy składniową i znaczeniową całość wewnątrz zdania).

PL	Pójdę na spacer, gdy przestanie padać deszcz. Fraza 1 (zdanie główne): Pójdę na spacer Łącznik: gdy Fraza 2 (Zdanie podrzędne): przestanie padać deszcz
EN	I wanted to call you, but I lost my phone. Fraza 1 (zdanie współrzędne): I wanted to call you Łącznik: but Fraza 2 (zdanie współrzędne): I lost my phone

Jednostka wielowyrazowa (MWE – Multi-Word Expression) - każda jednostka językowa składająca się z więcej niż jednego wyrazu (grupa słów), którą traktuje się jak jedno słowo (jeden leksem)

idiom (frazeologizm)	bić pianę, biały kruk
kolokacja	mocna kawa ←→ *silna kawa* silny wiatr ←→ *mocny wiatr*
wyrażenie ustalone	mimo wszystko, raz na zawsze, od stóp do głów
termin złożony	karta graficzna, Stany Zjednoczone
czasownik frazowy (EN)	give up
wyrażenie funkcyjne	w związku z, w celu

Części mowy - POS

Part of Speech - kategoria, do której zalicza się wyrazy na podstawie ich znaczenia, budowy i funkcji w zdaniu. Najpopularniejsze etykiety POS (zestaw tagów):

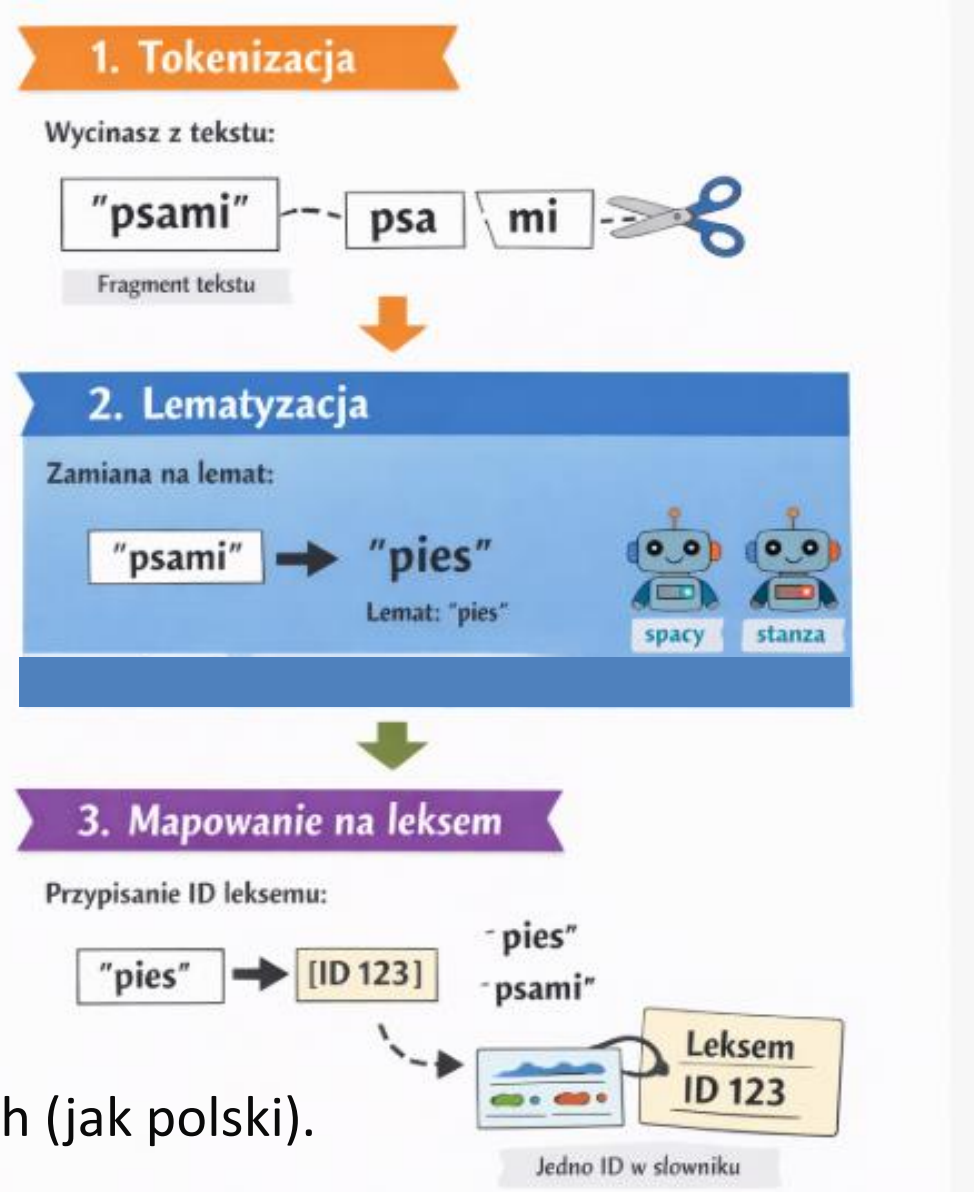
- rzeczownik (**NOUN**), czasownik (**VERB**), przymiotnik (**ADJ** - adjective), przysłówek (**ADV**) - adverb,
- zaimek (**PRON** - pronoun), rodzajnik (EN; **DET** – determiner, ale też article), przyimek (**ADP** – adposition, ale też preposition), liczebnik (**NUM** - numeral),
- spójnik (**CCONJ** - Coordinating Conjunction, dla spójników współrzędnych, np. I, a, ale, lub; **SCONJ** - Subordinating Conjunction dla spójników podrzędnych, np. że, bo, ponieważ, gdy), partykuły (**PART** – Particle, często służy do wyrażania emocji, intencji mówiącego lub akcentowania wybranych treści, np. czy, nie, chyba, niech, by, -że, np. Nie idę do kina, Chodźże szybciej, Oby jutro nie padało), wykrzykniki (**INTJ** – interjection, np. Och, jak tu pięknie!; Ach, jak miło cię widzieć!; Fuj, co to za zapach!), interpunkcja (**PUNCT** - punctuation).

Operacje porządku językowego – Lematyzacja

Lematyzacja → POS tagging → Rozpoznawanie MWE → Parsowanie składniowe

Lematyzacja - sprowadzanie odmienionych form wyrazów do ich formy podstawowej (kanonicznej), znanej jako **lemat**. W przeciwieństwie do prostszego usuwania końcówek (stemming), lematyzacja uwzględnia zasady gramatyczne i kontekst zdania - wykorzystuje słownik morfologiczny, (mapę form gramatycznych) i reguły języka do znalezienia formy bazowej .

W NLP granica między pojęciami **lemat** i **leksem** jest czysto operacyjna. Lemat staje się leksemem (lub jego reprezentantem), gdy przechodzi się z warstwy **tekstowej** (ciąg znaków) dowarstwy **semantycznej** (znaczenia).



Duże znaczenia dla języków fleksyjnych (jak polski).

Operacje porządku językowego – POS tagging

Lematyzacja → POS tagging → Rozpoznawanie MWE → Parsowanie składniowe

Automatyczne przypisywania każdemu słowu w tekście odpowiedniej części mowy. System sprawdza definicję słowa, analizuje jego rolę w zdaniu (w tym rozstrzyganie wieloznaczności, np. „flies”: „latać”, „muchy”).

Wyraz	POS	Objaśnienie
Mój	PRON	zaimek dzierżawczy
stary	ADJ	przymiotnik
pies	NOUN	rzeczownik
szybko	ADV	przysłówek
biegał	VERB	czasownik, forma przeszła
po	ADP	przyimek
ogrodzie	NOUN	rzeczownik, miejscownik
,	PUNCT	przecinek
gdy	SCONJ	spójnik podrzędny
nagle	ADV	przysłówek
zobaczył	VERB	czasownik, forma przeszła
trzy	NUM	liczebnik
kolorowe	ADJ	przymiotnik
ptaki	NOUN	rzeczownik
i	CCONJ	spójnik współrzędny
zaczął	VERB	czasownik, forma przeszła
szczekać	VERB	bezokolicznik
.	PUNCT	kropka

Word	POS	Explanation
My	PRON	possessive pronoun
old	ADJ	adjective
dog	NOUN	noun
ran	VERB	past tense verb
quickly	ADV	adverb
across	ADP	preposition
the	DET	determiner
garden	NOUN	noun
when	SCONJ	subordinating conjunction
he	PRON	personal pronoun
suddenly	ADV	adverb
saw	VERB	past tense verb
three	NUM	numeral
colorful	ADJ	adjective
birds	NOUN	noun
and	CCONJ	coordinating conjunction
started	VERB	past tense verb
barking	VERB	gerund/participle
loudly	ADV	adverb
.	PUNCT	punctuation

Operacje porządku językowego – Rozpoznawanie MWE

Lematyzacja → POS tagging → Rozpoznawanie MWE → Parsowanie składniowe

Identyfikowanie w tekście wyrażeń wielowyrazowych, które tworzą jedną jednostkę znaczeniową (leksem), zamiast analizować je jako oddzielne wyrazy. Ich znaczenie lub funkcja nie wynika wprost z pojedynczych wyrazów.

Metody rozpoznawanie MWE

- Statystyczne (asocjacyjne); analiza częstości współwystępowania słów w korpusie.

  biały dom - możliwe częste wstępowanie razem, znaczenie przewidywalne
  → nie jest MWE

  kupić kota w worku - rzadkie występowanie razem w innych kontekstach
  → potencjalne MWE

- Regułowe (słownikowe); porównanie tekstu ze słownikami MWE, ze stałymi gotowymi zwrotami.

  Stany Zjednoczone

- Uczenie maszynowe i analiza kontekstowa; modele uczone na oznaczonych korpusach MWE, używa się cech leksykalnych (słowa), morfologicznych (POS, przypadek, liczba, rodzaj, czas, osoba, stopień), składniowych (relacje między wyrazami w drzewie składniowym).

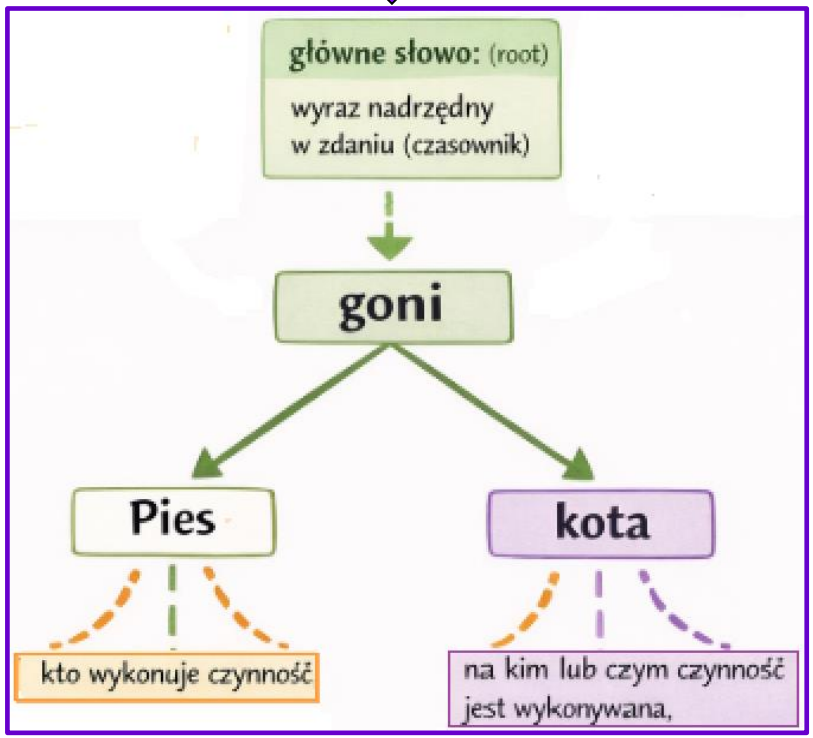
Operacje porządku językowego – Parsowanie składniowe

Lematyzacja → POS tagging → Rozpoznawanie MWE → Parsowanie składniowe

Zdanie przykładowe: **Pies goni kota.**		
Cecha	POS-tagging	**Parsowanie składniowe**
Poziom analizy	Pojedyncze słowo	Całe zdanie
Wynik	Kategoria gramatyczna	Relacja składniowa
Dla ww. zdania	NOUN, VERB, ADJ	Podmiot, orzeczenie, dopełnienie
Ww. przykład	Pies (NOUN) goni (VERB) kota (NOUN) . (PUNCT)	

Proces automatycznej analizy zdania w języku naturalnym w celu określenia jego struktury gramatycznej oraz relacji między poszczególnymi słowami.

Polega na rozbiciu zdania na mniejsze składowe (frazy, części mowy) i ustaleniu, jak te części łączą się ze sobą zgodnie z zasadami gramatyki – buduje strukturę całego zdania, zazwyczaj w formie drzewa.



Porządek reprezentacji statystycznej tekstu

Jednostki/obiekty

- Termin
- Stop-lista
- Słownik
- Tezaurus

Reprezentacje statystyczne

- Worek słów
- Macierz termin-dokument

Modelowanie tematyczne

Porządek reprezentacji dystrybucyjnej tekstu

Jednostki/obiekty

- Embedding słów
- Embedding zdania
- Embedding dokumentów

Word2Vex

GloVec

Modele typu Transformer

Jednostki porządku semantycznego

Zdanie

Najmniejsza jednostka tekstu w NLP stanowiąca samodzielną wypowiedź językową, zwykle zakończoną znakiem interpunkcyjnym i traktowaną jako podstawowa jednostka analizy składniowej lub semantycznej.

Dokument

Większa jednostka tekstu składająca się z jednego lub wielu zdań, stanowiąca pojedynczy obiekt danych w zadaniach NLP (np. artykuł, książka, wiadomość (komunikat), e-mail, post, transkrypcje rozmów).

Korpus

Uporządkowany zbiór wielu dokumentów lub tekstów, zgromadzonych w celu analizy języka lub trenowania i testowania modeli oraz analizy języka.

Operacje porządku semantycznego

Klasyfikacja tekstu

Analiza sentymentu

Streszczenia

Odpowiedzi na pytania

NER (Named Entity Recognition)

WSD (Word Sense Disambiguation)

Modelowanie tematów

I inne