Metody przetwarzania języka naturalnego na potrzeby systemów informatycznych

Pytanie specjalnościowe nr 10

Opracował:
Krystian Lema

Spis treści

1 Definicja i zastosowanie NLP

1.1 Definicja

Sztuczna inteligencja (AI) to kierunek badań na styku informatyki, neurologii i psychologii. Jego zadaniem jest konstruowanie urządzeń i oprogramowania zdolnego rozwiązać w oparciu o modelowanie wiedzy problemy nie poddające się algorytmizacji w sposób efektywny.

Przetwarzanie języka naturalnego (NLP) to dział **AI** zajmujący się poszukiwaniem metod formalnego opisu języka naturalnego oraz reprezentacji wiedzy zawartej w tekstach. Systemy NLP możemy podzielić na dwie kategorie:

- systemy ułatwiające korzystanie z innych programów,
- systemy służące do wykonywania pewnych operacji na tekstach w języku naturalnym.

1.2 Zastosowanie

Zastosowanie przetwarzania języka naturalnego:

- rozpoznawanie (rozumienie) i synteza mowy (systemy TextToSpeech)
- interfejsy w języku naturalnym (HCI z ang. human-computer interfaces)
- rozumienie i generowanie tekstów
- prowadzenie dialogu (np. inteligentne wyszukiwanie informacji, dokonywanie streszczeń, tworzenie bazy wiedzy, itd.)
- automatyczne tłumaczenie tekstów (np. system JANUS-II)
- inteligentne edytory tekstów
- nauka języków obcych

2 Przetwarzanie sygnału mowy

2.1 Struktura systemu rozpoznawania mowy

W bloku wstępnego przetwarzania następuje odbiór sygnału mowy z mikrofonu oraz jego wstępne przetworzenie: wzmocnienie, filtracja (ograniczenie pasma od dołu i od góry) oraz przetworzenie na postać cyfrową (przetwornik a/c). W obecnych systemach funkcję tą spełnia karta dźwiękowa.

W bloku ekstrakcji parametrów następuje analiza sygnału mowy, w wyniku której otrzymuje się wartości parametrów niosących informację o treści wypowiedzi i niezależnych od indywidualnych cech głosu mówcy. Zbiór tych parametrów tworzy wektor (lub macierz) cech, na podstawie którego dokonuje się klasyfikacji sygnału.

W bloku klasyfikacji następuje porównanie nadchodzących ciągów obrazów wypowiedzi ze znajdującymi się w pamięci wzorcami, które stanowią uogólniony (uśredniony) opis klas dźwięków. Klasą mogą być fonemy, wyrazy lub nawet całe zdania. Obrazy wzorcowe są tworzone w procesie uczenia.

2.2 Metody klasyfikacji sygnałów

Do podstawowych metod rozpoznawania mowy zalicza się:

- algorytmy statystyczne parametryczne (np. algorytm Bayesa)
- algorytmy statystyczne nieparametryczne (np. algorytm NN – najbliższy sąsiad)
- algorytmy oparte o funkcję podobieństwa
- programowanie dynamiczne
- ukryte modele Markowa HMM
- sieci neuronowe

2.3 Parametry sygnału mowy

2.3.1 Parametry w dziedzinie czasu

Parametry w dziedzinie czasu to grupa parametrów, które można określić bezpośrednio na podstawie struktury czasowej sygnału mowy. Parametry te można podzielić na dwie grupy:

- parametry będące funkcjami czasu
  - natężenie sygnału w funkcji czasu
  - obwiednia amplitudy sygnału
- parametry związane z pomiarem punktów, w których sygnał mowy zmienia znak (przejście sygnału mowy przez zero)
  - gęstość przejść przez zero
  - rozkład interwałów pomiędzy kolejnymi przejściami przez zero

2.3.2 Parametry w dziedzinie częstotliwości

Podstawową formą prezentacji wizualnej sygnału mowy w dziedzinie częstotliwości jest trójwymiarowy obraz, który przedstawia amplitudę sygnału w funkcji czasu i częstotliwości.

Obraz sygnału mowy w dziedzinie częstotliwości jest obrazem złożonym i jego bezpośrednie wykorzystanie w procesie rozpoznawania jest właściwie niemożliwe. W związku z tym poszukuje się parametrów, które dobrze odzwierciedlają jego własności. Do podstawowych parametrów w dziedzinie częstotliwości zalicza się:

- częstotliwości formantowe (maksimum obwiedni)
- względne amplitudy poszczególnych formantów
- częstotliwości antyformantowe (minimum obwiedni)
- częstotliwość podstawowa tonu krtaniowego
- wartości średnie widma liczone np. w pasmach tercjowych (1/23 oktawy)

3 Przetwarzanie tekstu

3.1 Podział na zdania

W języku polskim wykorzystuje się regułowy podział tekstu na zdania. Algorytm ten został zaproponowany w ramach prac prowadzonych przy tworzeniu korpusu PWN języka polskiego. Algorytm ten korzysta z następujących danych:

- informacji o przypisanych słowom częściach mowy,
- listy skrótów zakończonych kropką,
- listy słów, które są niejednoznaczne,
- informacji dla każdego skrótu czy może on wystąpić na końcu zdania.

Algorytm znajduje potencjalne znaki początku i końca zdania a następnie za pomocą serii reguł weryfikuje czy potencjalny znak jest znakiem faktycznym początku/końca zdania.

3.2 Trójetapowe wstępne przetwarzanie tekstu

3.2.1 Segmentacja

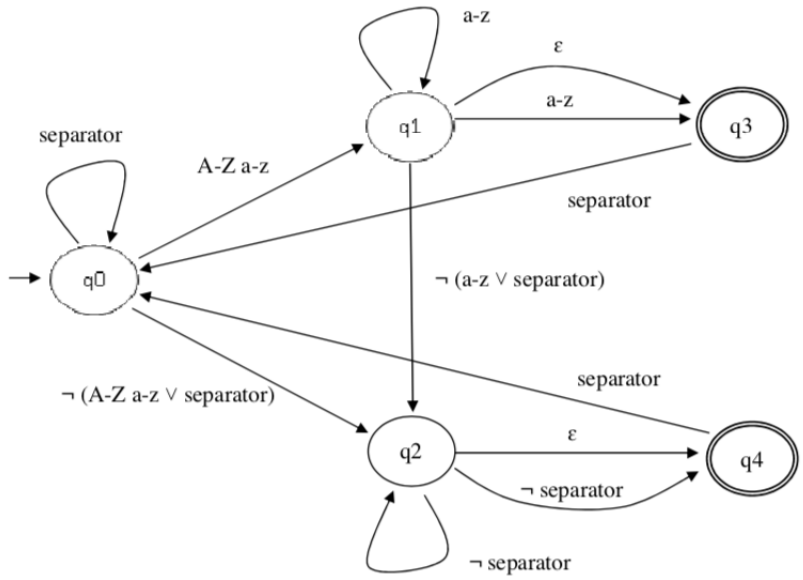
Celem segmentacji tekstu jest jego podział na jednostki zwane segmentami lub tokenami (ang. tokens), którym można przypisać interpretacje morfosyntaktyczne w postaci znaczników morfologicznych (tzw. tagów). Znaczniki te zawierają informację o częściach mowy (np. rzeczownik, czasownik, przymiotnik) oraz wartościach kategorii gramatycznych (np. liczba, rodzaj, przypadek). Tak rozumiana segmentacja nazywana jest również tokenizacją. Najczęściej wyróżnia się następujące klasy segmentów (tokenów):

- ciąg małych liter rozpoczynający się od wielkiej litery np. Wrocław
- ciąg składający się tylko z wielkich liter np. PZU, ZUS
- ciąg składający się tyko z małych liter np. dom, komputer
- ciąg małych i wielkich liter np. PeKaO
- ciąg cyfr np. 123
- ciąg cyfr z wewnętrzną kropką lub przecinkiem np. 12.5
- znak interpunkcyjny np. kropka (.), przecinek (,), średnik (;), myślnik (-)

Dodatkowo można wyróżnić typy segmentów charakterystycznych dla tekstów określonego typu:

- data, godzina adres, numer telefonu
- adres e-mail, adres strony www, tagi języka HTML
- wzory cząsteczek związków chemicznych

Jedną z metod wykorzystywanych do rozpoznawania granic segmentów są automaty skończone. Na ogół zapewniają one wystarczającą skuteczność i efektywność.



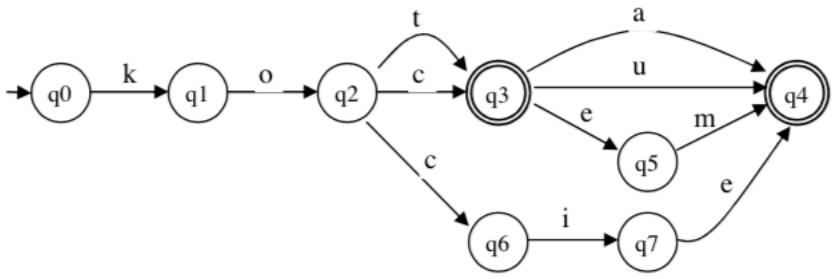
Rysunek 1: Automat, który dzieli tekst na dwa typu segmentów: ciągi liter, w których jako pierwsza litera może być duża (pozostałe litery małe) oraz wszystkie pozostałe typy segmentów

3.2.2 Analiza morfologiczna

Morfologia to nauka o budowie słów. W zakres morfologii wchodzą dwie powiązane ze sobą, ale jednak odrębne dziedziny:

- fleksja, która opisuje różne formy tego samego wyrazu np. kot -> kotem, być -> byłem
- słowotwórstwo, która opisuje zasady tworzenia wyrazów pochodnych np. zebrać -> zebranie,chodzić -> chodzenie ->chód.

Analiza morfologiczna polega na przypisaniu wyróżnionym w tekście segmentom wszystkich możliwych (niezależnych od kontekstu) interpretacji morfologicznych. Pierwszym etapem tej analizy jest stwierdzenie, czy analizowane segmenty są poprawnymi słowami danego języka (ustalenie, czy dany napis jest formą jakiegoś leksemu znajdującego się w słowniku). Następnie określane są własności morfologiczne danej formy wyrazu. Jedną z metod stosowanych do reprezentacji komputerowych słowników są automaty skończone.



Rysunek 2: Automat reprezentujący fragment słownika polskiego uwzględniający odmianę wyrazów kot i koc

3.2.3 Dezambiguacja morfologiczna

Dezambiguacja - uściślenie, ujednolicenie. W wyniku analizy morfologicznej poszczególnym wyrazom w tekście zostały przypisane odpowiednie zbiory tagów. Zadaniem dezambiguacji morfologicznej jest ograniczenie tych zbiorów tylko do tagów które nie są sprzeczne z kontekstem użycia wyrazów. Proces ten polega więc na wybraniu spośród wszystkich możliwych interpretacji wyrazów tylko tych, które są właściwe w danym kontekście.

Dezambiguacja jest jednym z etapów przetwarzania języka naturalnego, a następuje po analizie morfologicznej. Przez to na swoim wejściu ma już przypisane tagi do odpowiednich wyrazów. Jej celem jest ich ujednolicenie, poprzez ograniczenie zbiorów tagów, które nie są sprzeczne z kontekstem użytego wyrazu.

Metody stosowane w procesie ujednoznaczniania morfologicznego można podzielić na dwie klasy:

- oparte na wiedzy lingwistycznej (metody regułowe, przykładowe systemy: Constraint Grammar, XIP, LanGR, JOSKIPI, INTEX, TAGGIT)
- oparte na danych treningowych
  - metody indukcyjne (metody zdobywające wiedze w postaci symbolicznej, metody oparte na modelach probabilistycznych)
  - metody statystyczne

Dezambiguatory regułowe są tworzone na podstawie wiedzy ekspertów – językoznawców. Wiedza ta zostaje zakodowana w postaci odpowiednich reguł. Przykładem może być następująca reguła: Jeżeli po jednoznacznym przyimku występuje segment o interpretacjach rzeczownikowych i czasownikowych, to odrzuć interpretacje czasownikowe.

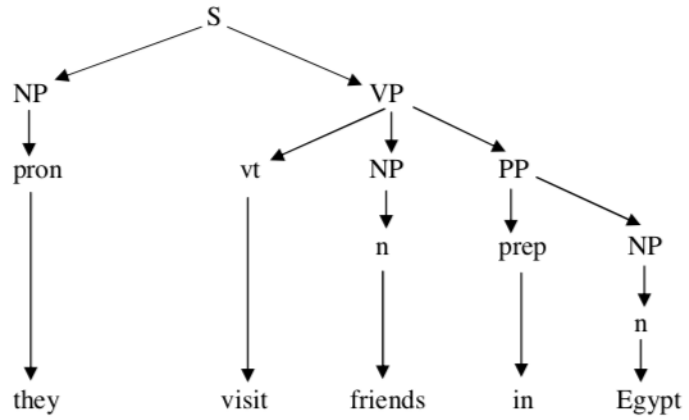
3.3 Analiza syntaktyczna

Syntaktyka zajmuje się opisem reguł budowy zdań z wyrazów.

Zbiór reguł syntaktyczna dla danego języka nazywamy **gramatyką** tego języka.

Proces analizy syntaktycznej nazywany jest **parsowaniem**, którego zadaniem jest transformacja tekstu na strukturę zawierającą informację o związkach i zależnościach między wyrazami i częściami zdania.

Wynik analizy syntaktycznej możemy przedstawic w postaci drzewa struktury frazowej:



Rysunek 3: Drzewo parsowania

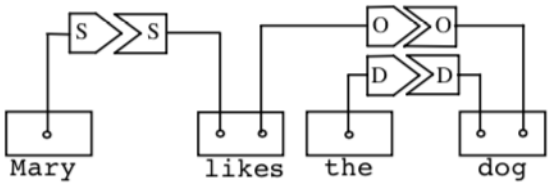
Do analizy syntaktyczne używa się:

- algorytm CYK

  Algorytm ten oparty jest na metodzie programowania dynamicznego. Wykorzystuje gramatykę bezkontekstową w postaci normalnej Chomsky'ego. Bazuje na opisie zbioru symboli terminalnych i nieterminalnych, zbiorze reguł oraz elemencie początkowym. Algorytm polega na sprawdzeniu czy według reguł jesteśmy w stanie uzyskać dla sprawdzanego słowa ciąg symboli nieterminalnych.

- gramatykę łączeń

  Klasyczna gramatyka łączeń składa się ze zbiory słów, które są terminalnymi symbolami gramatyki, z których każde ma tzw. wymagania łączeniowe. Ciąg słów jest zdaniem języka opisywanego przez gramatykę łączeń, jeśli można poprowadzić łuki pomiędzy poszczególnymi wyrazami zdania w ten sposób, że spełnione są wymagania łączeniowe dla danej gramatyki.



Rysunek 4: Przykład gramatyki łączeń

3.4 Analiza semantyczna

Semantyka zajmuje się opisem znaczenia. W zależności od przeznaczenia systemu NLP celem analizy semantycznej może być:

- sprawdzenie poprawności zdania pod względem zrozumiałości i sensowności (np. odrzucenie zdań nielogicznych)
- usunięcie niejednoznaczności w strukturze zdania (np. właściwa interpretacja wyrazów, wybór właściwego rozbioru gramatycznego zdania)
- określenie reprezentacji znaczenia zdania (np. opis znaczenia zdania w celu dalszej analizy)

Do analizy i opisu semantyki w systemach NLP stosowane są różne formalizmy. Do najczęściej stosowanych należą:

- rachunek predykatów pierwszego rzędu
- sieci semantyczne
- gramatyki semantyczne (ang. semantic grammars)
- teoria zależności pojęciowych Schanka (ang. Conceptual Dependancy Theory)
- gramatyka przypadków głębokich Fillmore'a (ang case grammar)

4 Źródła

[1] dr inż. Dariusz Banasiak, Materiały z wykładu Projektowanie systemów z dostepem w jezyku naturalnym