Politechnika Wrocławska

Metody przetwarzania języka naturalnego na potrzeby systemów informatycznych (S10)

inż. Krystian Lema

Informatyka

Inżynieria systemów informatycznych (INS)

HR EXCELLENCE IN RESEARCH

Politechnika Wrocławska

Plan prezentacji

1. Definicja NLP
2. Zastosowanie NLP
3. Przetwarzanie sygnału mowy
4. Przetwarzanie tekstu
5. Analiza syntaktyczna
6. Analiza semantyczna
7. Źródła
8. Literatura

1. Definicja NLP

Definicja sztucznej inteligencji

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

1. Definicja NLP

Definicja przetwarzania języka naturalnego

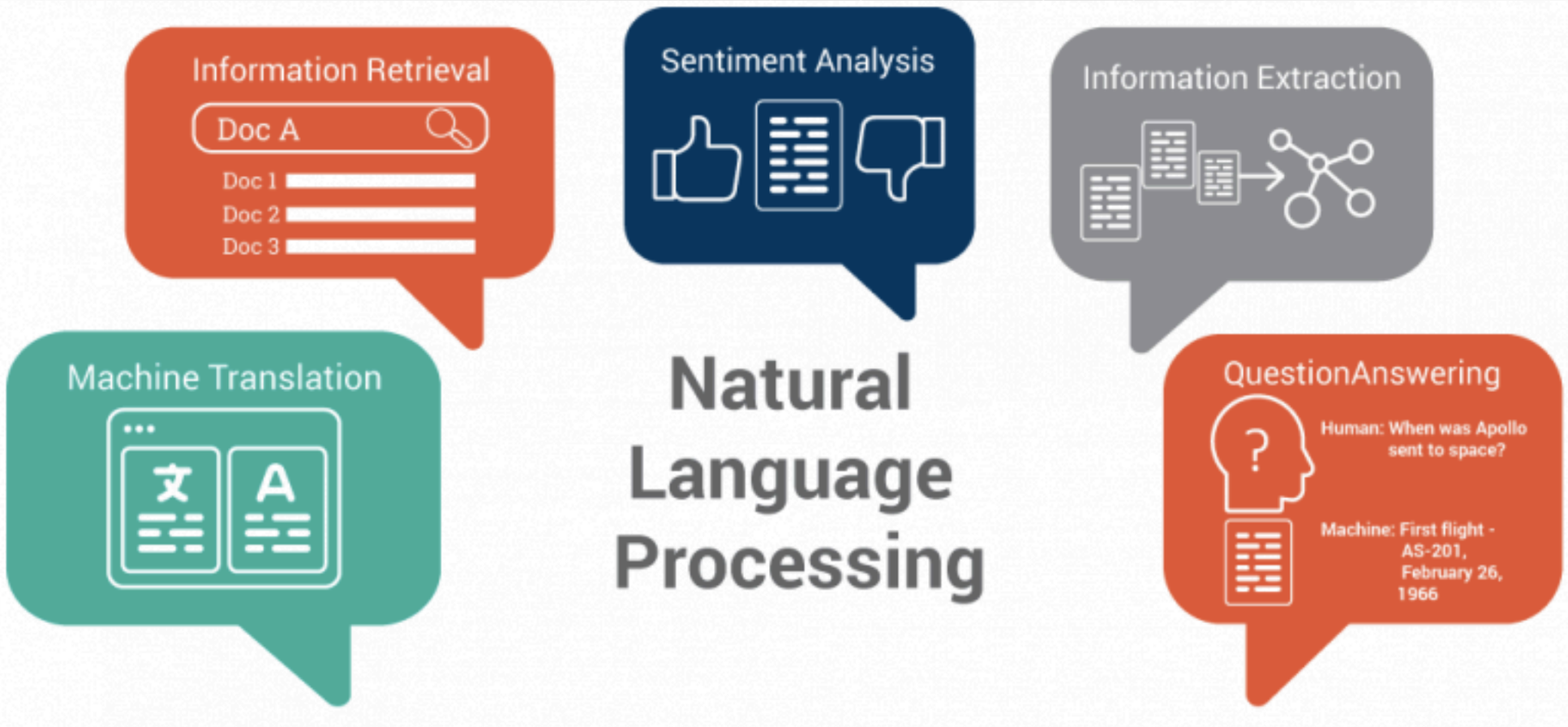
Przetwarzanie języka naturalnego (**NLP**) to dział **AI** zajmujący się poszukiwaniem metod formalnego opisu języka naturalnego oraz reprezentacji wiedzy zawartej w tekstach.

Systemy NLP możemy podzielić na dwie kategorie:

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

2. Zastosowanie NLP

Zastosowanie przetwarzania języka naturalnego



2. Zastosowanie NLP

Zastosowanie przetwarzania języka naturalnego

- rozpoznawanie (rozumienie) i synteza mowy
- 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

NLP

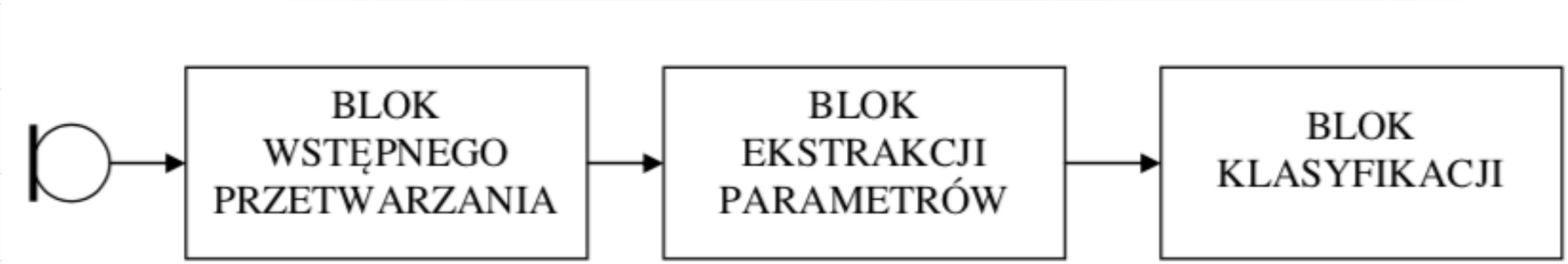
Przetwarzanie sygnału mowy

Przetwarzanie tekstu

3. Przetwarzanie sygnału mowy

Definicja

Struktura systemu rozpoznawania mowy



3. Przetwarzanie sygnału mowy

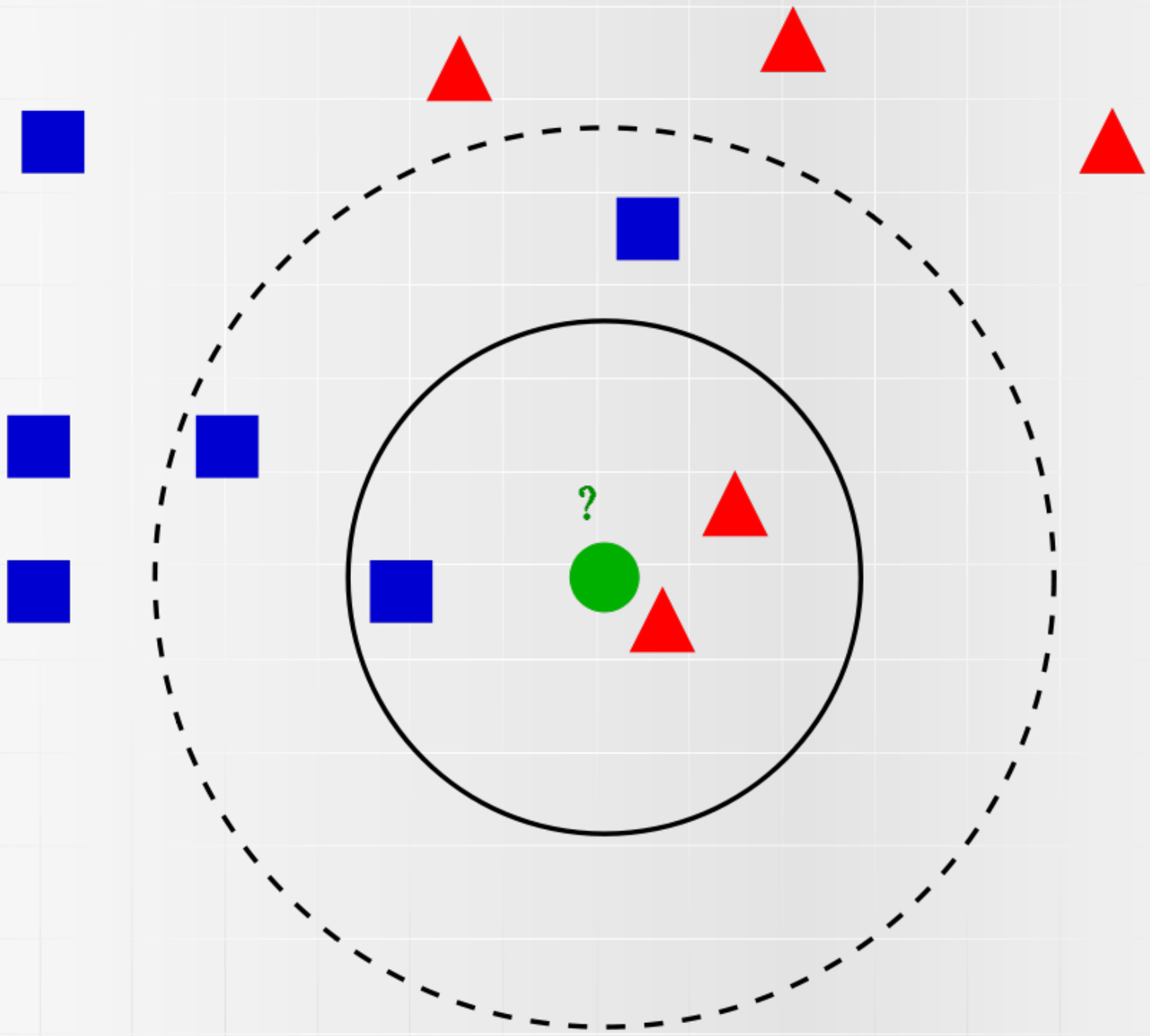
Metody

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

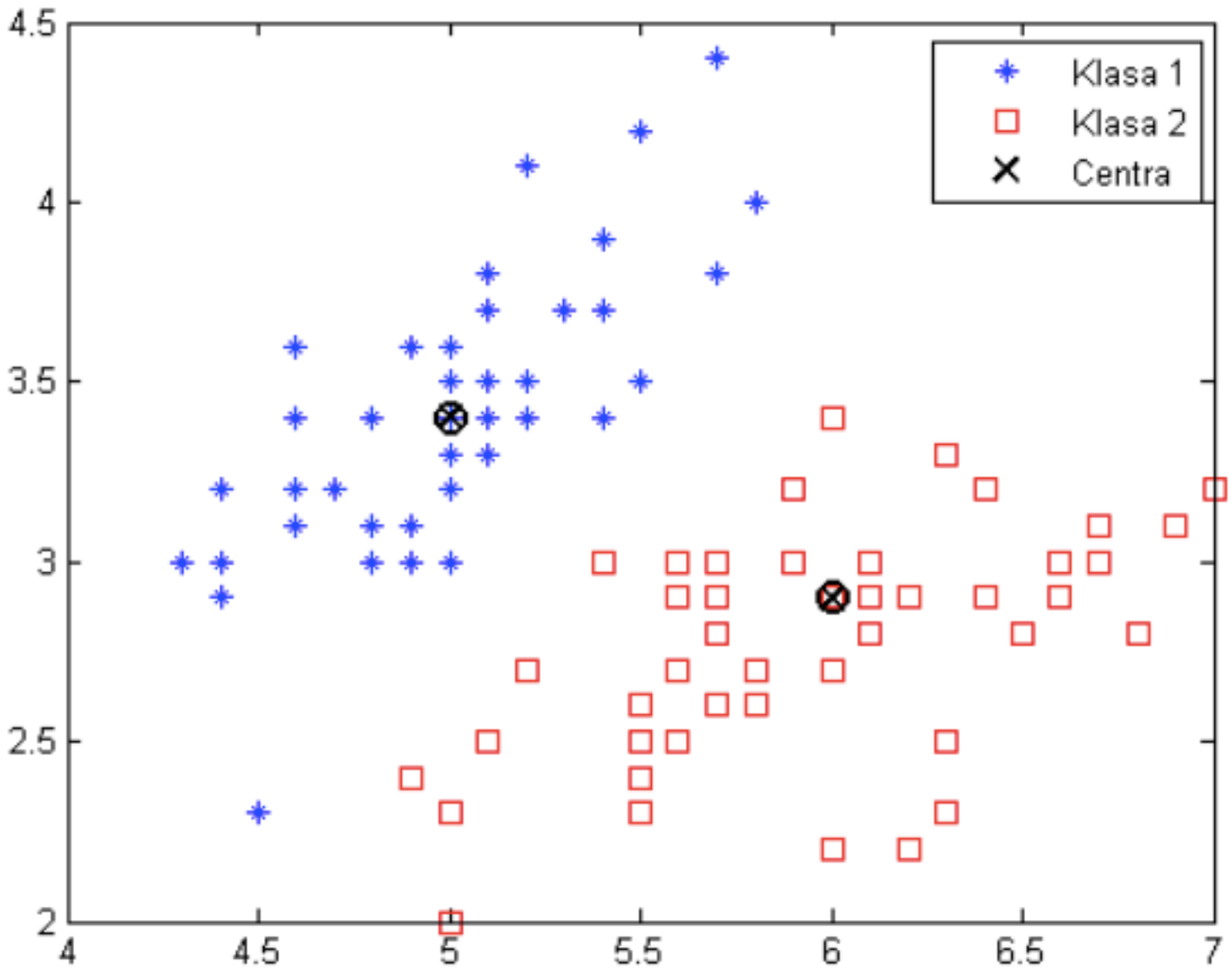
3. Przetwarzanie sygnału mowy

Algorytm k-najbliższych sąsiadów



3. Przetwarzanie sygnału mowy

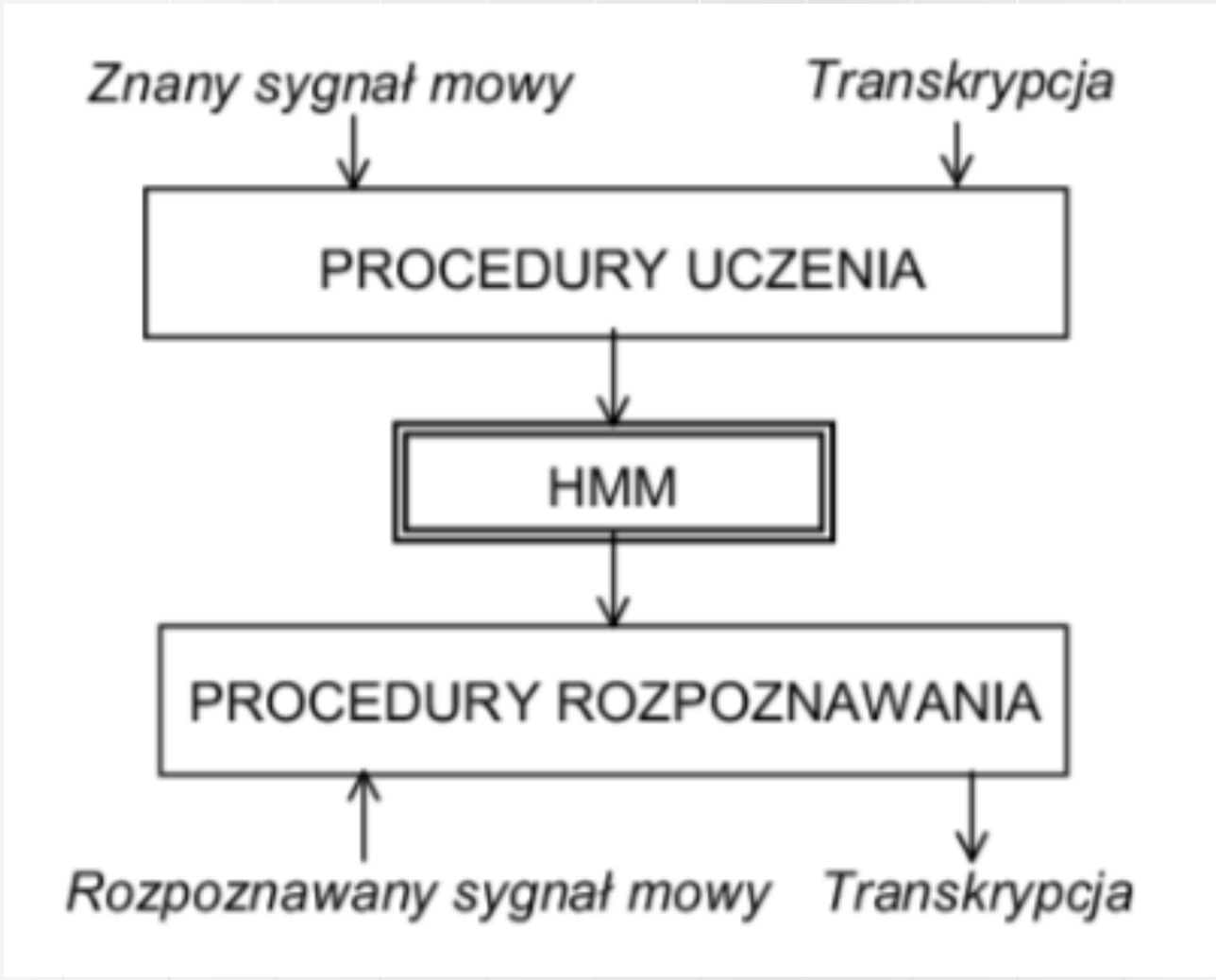
Algorytm najbliższa średnia



3. Przetwarzanie sygnału mowy

Ukryte modele Markowa (HMM)

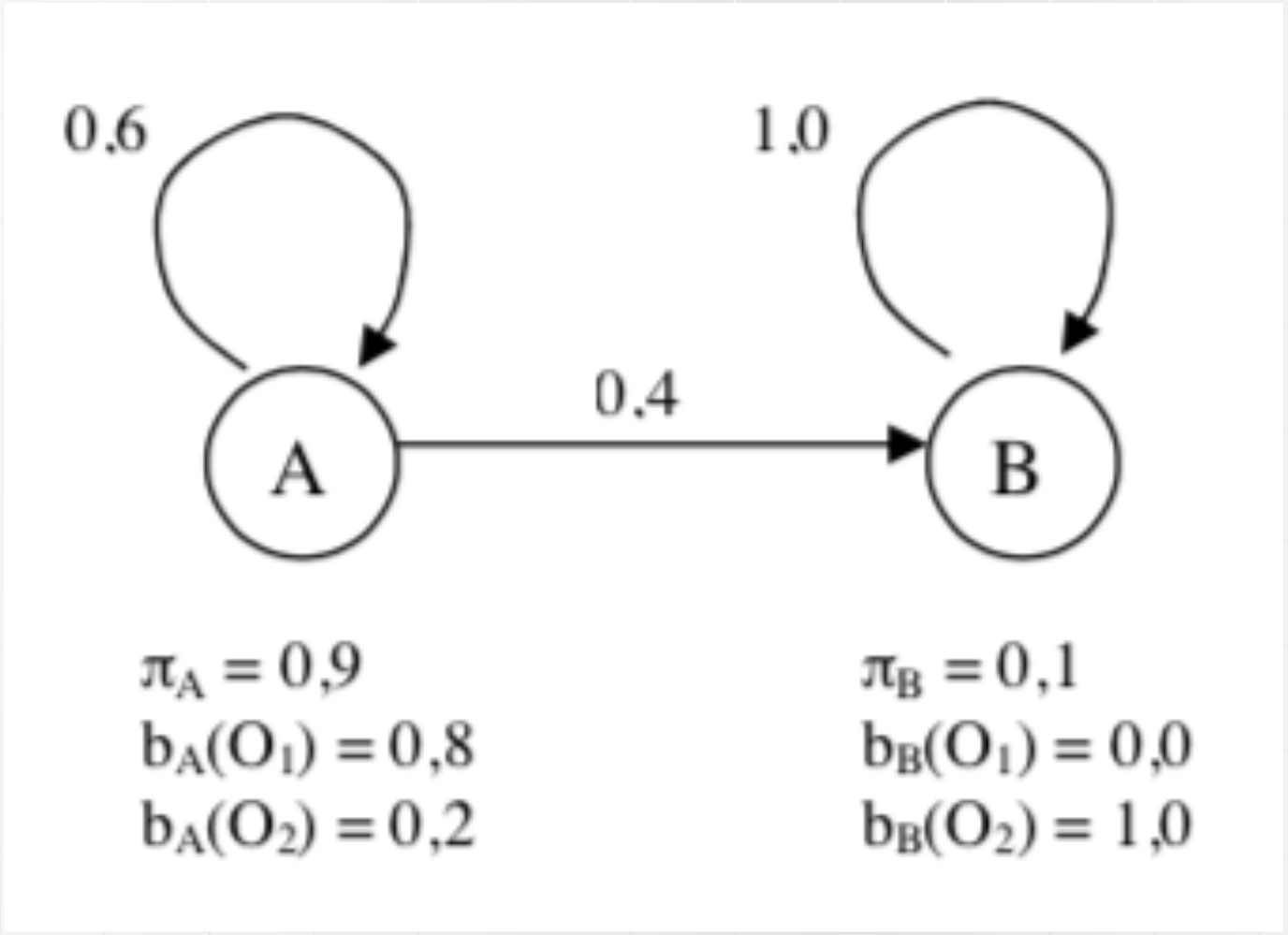
Proces rozpoznawana mowy za pomocą HMM przedstawia poniższy schemat:



3. Przetwarzanie sygnału mowy

Ukryte modele Markowa (HMM)

Przykładowy automat o dwóch stanach reprezentujący model pewnego słowa:



4. Przetwarzanie tekstu

Definicja

Etapy przetwarzania tekstu:

- Segmentacja - podział tekstu na jednostki zwane segmentami co umożliwia przypisanie znaczników zawierających informacje o częściach mowy oraz wartościach kategorii gramatycznych
- Analiza morfologiczna - przypisanie wyróżnionym w tekście segmentom wszystkich możliwych interpretacji morfologicznych (znaczniki)
- Dezambiguacja morfologiczna - ograniczenie zbioru znaczników tylko do takich które nie są sprzeczne z kontekstem użycia wyrazów

4. Przetwarzanie tekstu

Rozpoznawania granic zdań

Problem rozpoznawania granic zdań:

Sesję rozpoczął prof. dr hab. Juliusz P. Kowalski. (1)

Inflacja w 2008 r. wyniosła 3 proc., natomiast w 2009 r. 4 proc. Były to … (2)

Symbol „?” oznacza szukanie dowolnego pojedynczego znaku. (3)

4. Przetwarzanie tekstu

Regułowy podział tekstu na zdania dla języka polskiego

Algorytm ten korzysta z następujących danych:

- informacji o przypisanych słowom częściach mowy,
- listy skrótów zakończonych kropką (uwzględniono skróty, które kończą się kropką tylko w określonych sytuacjach np. dr Jurand ale dr. Juranda),
- listy słów, które są niejednoznaczne (mogą być skrótem zakończonym kropką lub formą rzeczownika, który występuje na końcu zdania np. ul.),
- informacji (dla każdego skrótu), czy może on wystąpić na końcu zdania, w środku zdania lub na obu tych pozycjach np. skróty od nazw tytułów naukowych (prof., doc., dr) muszą wystąpić w środku zdania, gdyż wymagają dopełnień w postaci określonej frazy.

4. Przetwarzanie tekstu

Regułowy podział tekstu na zdania dla języka polskiego

W algorytmie tym wprowadzono następujące oznaczenia:

- **POM** (ang. potential opening marker) – znak, który może rozpoczynać zdanie np. wielka litera,
- **OM** (ang. opening marker) – znak, który rzeczywiście rozpoczyna zdanie,
- **PCM** (ang. potencial closing marker) – znak, który może kończyć zdanie np. kropka,
- **CM** (ang. closing marker) – znak, który rzeczywiście kończy zdanie.

4. Przetwarzanie tekstu

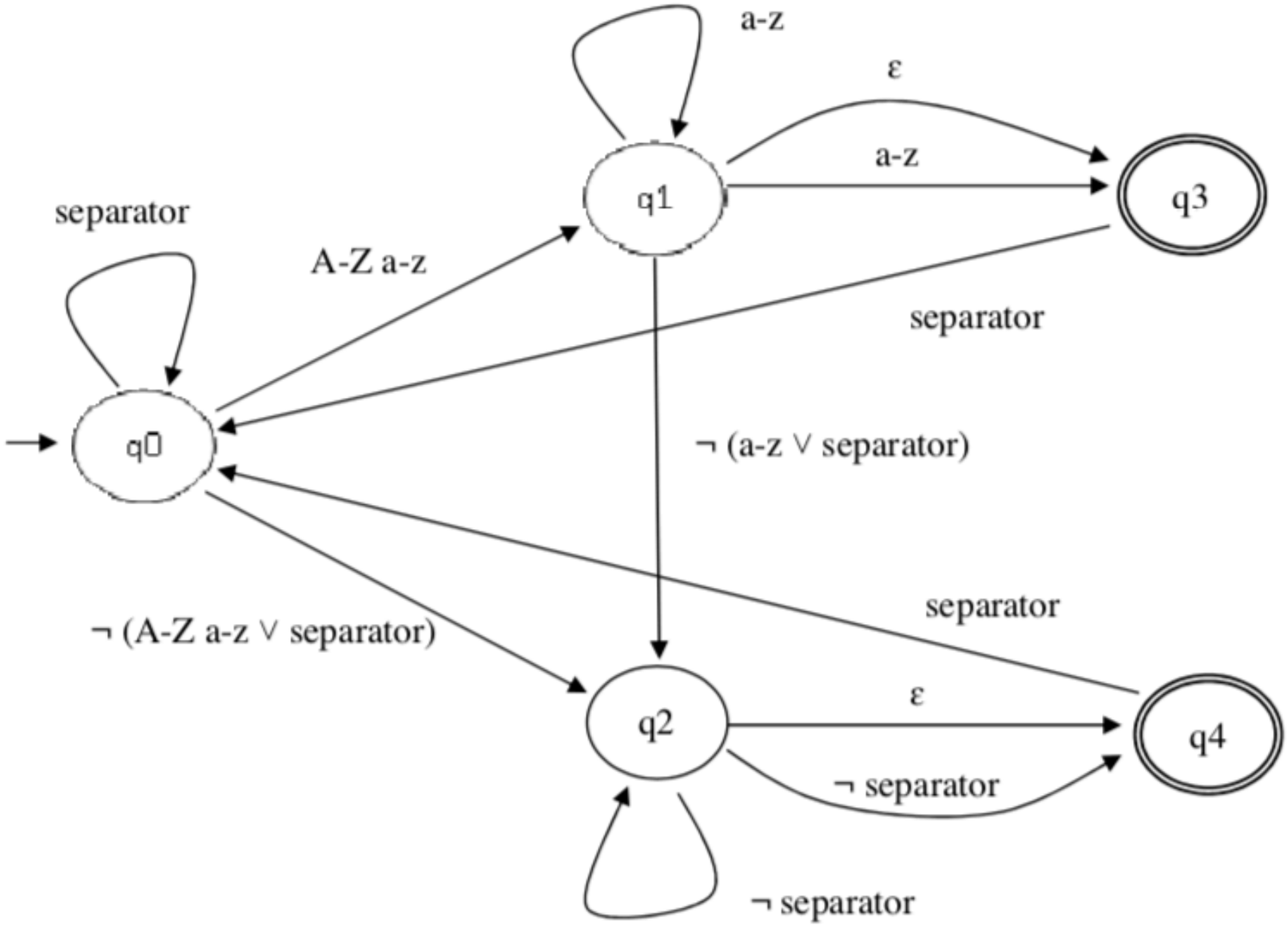
Regułowy podział tekstu na zdania dla języka polskiego

W algorytmie zaproponowano następujące reguły rozstrzygające o granicach zdań:

- PCM, po którym występuje inny znak przystankowy (z wyjątkiem myślnika) nie jest CM,
- PCM, po którym występuje mała litera (poprzedzona ewentualnie odstępem lub myślnikiem) nie
  jest CM,
- kropka pomiędzy cyframi nie jest CM,
- kropka poprzedzona skrótem bez kropki jest CM,
- kropka poprzedzona skrótem, który wymaga dopełnienia nie jest CM,
- inicjały nigdy nie kończą zdania,
- jeżeli słowo zaczynające się wielką literą następuje po PCM i może być jednoznacznie rozpoznane jako czasownik, partykuła, przysłówek lub spójnik, to PCM jest CM.

4. Przetwarzanie tekstu

Segmentacja tekstu - automaty skończone



4. Przetwarzanie tekstu

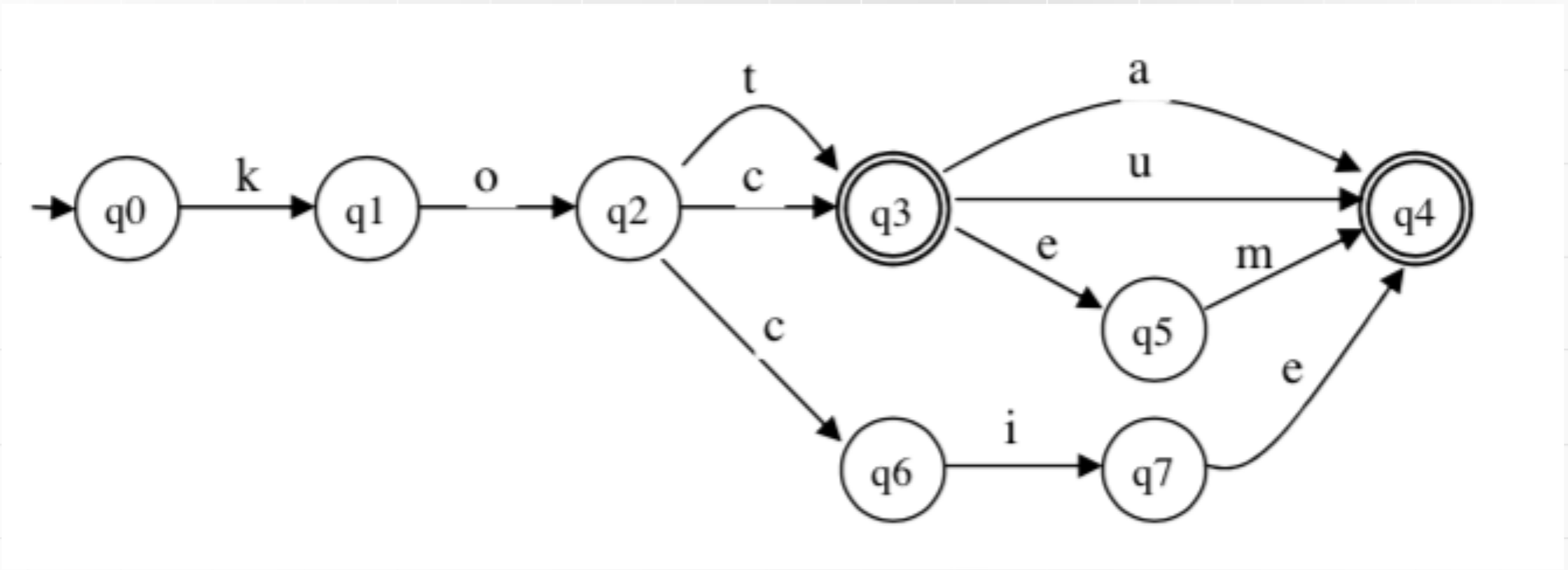
Segmentacja tekstu - klasy segmentów

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 (-)

4. Przetwarzanie tekstu

Analiza morfologiczna - automaty skończone



4. Przetwarzanie tekstu

Analiza morfologiczna - analizator morfologiczny Morfeusz

Ala ma kota.

0	1	Ala	**Ala**	subst:sg:nom:f	imię
			Al	subst:sg:gen.acc:m1	imię
			Alo	subst:sg:gen.acc:m1	imię
1	2	ma	**mój:a**	adj:sg:nom.voc:f:pos	
			mieć	fin:sg:ter:imperf	
2	3	kota	**kota**	subst:sg:nom:f	nazwa_pospolita
			kot:s1	subst:sg:gen.acc:m2	nazwa_pospolita
			kot:s2	subst:sg:gen.acc:m1	nazwa_pospolita pot., środ.
3	4	.	**.**	interp	

4. Przetwarzanie tekstu

Dezambiguacja morfologiczna - metody

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 wiedzę w postaci symbolicznej, metody oparte na modelach probabilistycznych)
  - metody statystyczne

5. Analiza syntaktyczna

Definicja

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.

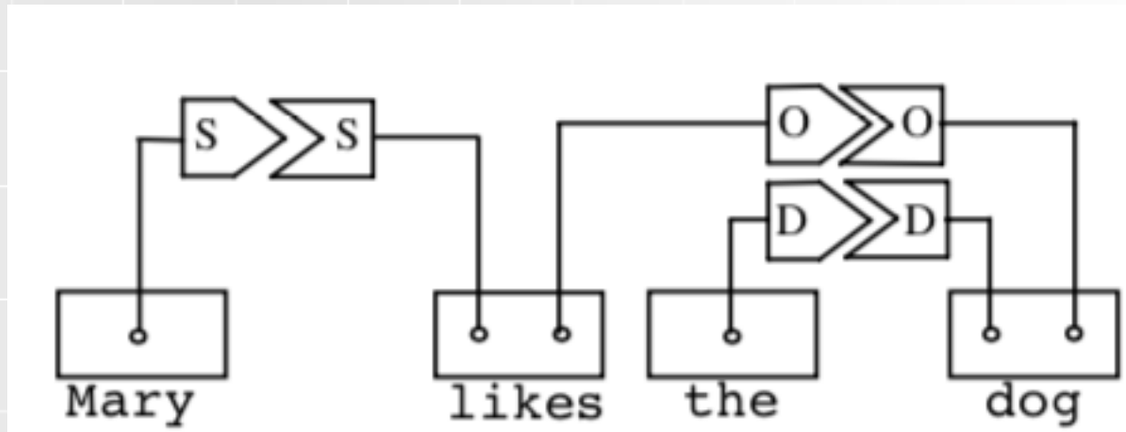
5. Analiza syntaktyczna

Metody

Do sprawdzania poprawności gramatycznej zdań wykorzystuje się:

- Algorytm CYK
- Gramatykę łączeń

1. $S \rightarrow SS$
2. $S \rightarrow AB$
3. $A \rightarrow AS$
4. $A \rightarrow AA$
5. $A \rightarrow a$
6. $B \rightarrow SB$
7. $B \rightarrow BB$
8. $B \rightarrow b$



6. Analiza semantyczna

Definicja

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)

6. Analiza semantyczna

Metody

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)

NLP

Przetwarzanie sygnału mowy

- Klasyfikator Bayesa
- k-NN
- NM
- HMM
- Sieci Neuronowe

Przetwarzanie tekstu

Analiza morfologiczna

Segmentacja

Regułowy podział na zdania

Dezambiguacja morfologiczna

Morfeusz

- Metody regułowe
- Metody statystyczne i indukcyjne

Automaty skończone

Analiza syntaktyczna

Analiza semantyczna

7. Źródła

Źródła obrazów

[1] https://www.ontotext.com/top-5-semantic-technology-trends-2017/, slajd: 5

[2] https://pl.wikipedia.org/wiki/K_najbliższych_sąsiadów, slajd: 10

[3] http://urszula.libal.staff.iiar.pwr.wroc.pl/docs/aro/aro6.pdf, slajd: 11

[4] http://sgjp.pl/morfeusz/demo/?text=Ala+ma+kota., slajd: 22

8. Literatura

Przedstawienie literatury użytej do stworzenia prezentacji

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

Politechnika Wrocławska

Dziękuję za uwagę

HR EXCELLENCE IN RESEARCH