Inteligentne systemy decyzyjne

Przetwarzanie języka naturalnego

Plan wykładu

- Etapy analizy językowej.
- Rozumienie języka naturalnego.
- Generowanie tekstu.
- Szukanie semantyczne.
- Tłumaczenie maszynowe.
- Rozwiązania dostępnego oprogramowania do przetwarzania języka naturalnego

Natural Language Processing

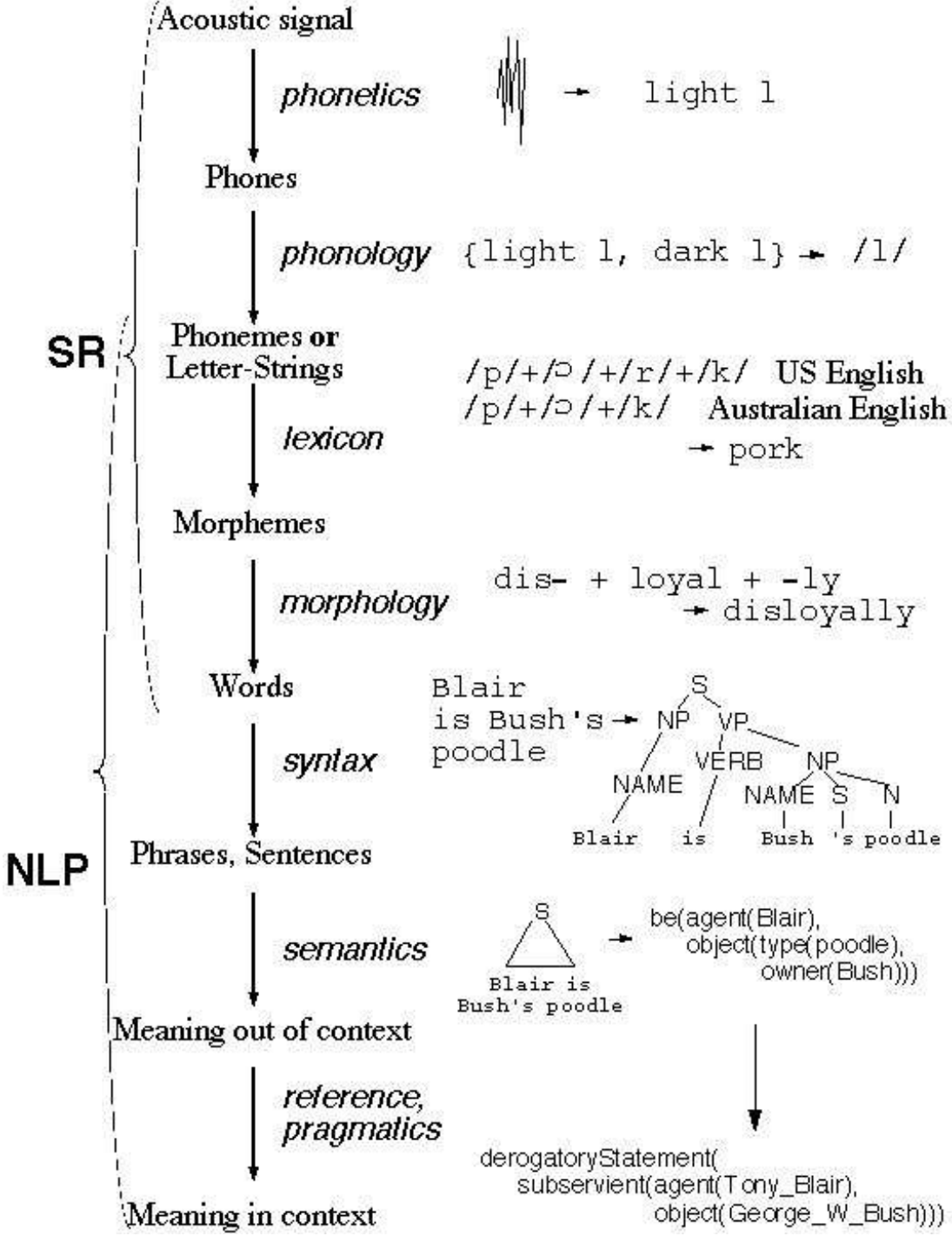
Przetwarzanie języka naturalnego (ang. *Natural Language Processing*) – dziedzina z pogranicza sztucznej inteligencji i językoznawstwa, zajmująca się opracowaniem metod przetwarzania przy pomocy komputera informacji przedstawionej w formie wypowiedzi w języku naturalnym.

Zagadnieniami wchodzącymi w skład NLP są na przykład:

- rozbiór tekstu na jednostki gramatyczne i morfologiczne,
- tworzenie gramatycznych modeli zdań,
- rozumienie język naturalnego,
- generowanie języka naturalnego.
- automatyczne tłumaczenie tekstu,
- stworzenie programów do rozmowy z komputerem (infoboty, chatterboty),
- wyszukiwanie tekstu według zawartości semantycznej.

Rozumienie języka naturalnego

Rozumienie języka naturalnego (*Natural Language Understanding*) – dział NLP zajmujący się zamianą informacji dostępnej w formie języka naturalnego na wiedzę możliwą do zapisania w bazie danych komputera.


Acoustic signal
phonetics
light l
Phones
phonology
{light l, dark l} → /l/
SR
Phonemes or
Letter-Strings
/p/+/ɔ/+/r/+/k/ US English
/p/+/ɔ/+/k/ Australian English
lexicon
→ pork
Morphemes
morphology
dis- + loyal + -ly
→ disloyally
Words
Blair
is Bush's
poodle
S
NP
VP
NAME
VERB
NP
NAME
S
N
Blair
is
Bush
's poodle
syntax
NLP
Phrases, Sentences
semantics
S
Blair is
Bush's poodle
be(agent(Blair),
object(type(poodle),
owner(Bush)))
Meaning out of context
reference,
pragmatics
derogatoryStatement(
subservient(agent(Tony_Blair),
object(George_W_Bush)))
Meaning in context

Etapy analizy językowej

- Analiza fonologiczna – konieczna w przypadku rozpoznawania mowy, analiza dźwięków wchodzących w skład wypowiedzi w celu wyodrębnienia głosek i zamiany ich na litery.
- Rozbiór morfologiczny – podział wypowiedzi na wyrazy, a wyrazów na części takie jak temat i końcówka lub sylaby.
- Analiza syntaktyczna – uwzględnienie reguł tworzenia zdań obowiązujących w języku oraz gramatyki wyrazów. Efektem jest rozbiór wypowiedzi na części mowy i części zdania.
- Analiza semantyczna – uwzględnienie znaczenia słów. Rozróżnienie nazw własnych i rzeczowników pospolitych, odróżnienie homonimów.
- Analiza pragmatyczna– najwyższy poziom analizy. Uwzględnia sens wypowiedzi, typowe połączenia wyrazów i relacje pomiędzy częściami wypowiedzi.

Gramatyka bezkontekstowa

Gramatyka bezkontekstowa (*Context-Free Grammar*) – gramatyka umożliwiająca generację języka w oparciu o elementy nieterminalne (niezależne od znaczenia).

G = {T,N, S,R}

- T – zbiór symboli terminalnych
- N – zbiór symboli nieterminkalnych.
- S – element początkowy
- R – zbiór zasad

Gramatyka G generuje język L

Gramatyka bezkontekstowa

Elementy terminalne:

pies, książka, drzewo, student, uczelnia, ja, w, nad, spać, jeść…

Elementy nieterminalne:

VB – Verb – czasownik (orzeczenie)

NN – Noun – rzeczownik (podmiot, dopełnienie)

ADJ – Adjective – przymiotnik (przydawka)

DT – Determiner – rodzajnik

P – Preposition – przyimek

PRO – Pronoun – zaimek

VP – Verb Phrase – fraza z czasownikiem

NP – Noun Phrase – fraza z rzeczownikiem

Gramatyka bezkontekstowa

Kluczowe znaczenie dla wykorzystania gramatyki bezkontekstowej w NLP, zarówno w procesie rozumienia, jak i generowania, mają **reguły**.

S → NP VP - zdanie składa się z frazy rzeczownikowej i frazy czasownikowej

NP → ADJ NN – fraza rzeczownikowa składa się z przymiotnika i rzeczownika

VP → VV PP – fraza z czasownikiem składa się z czasownika i wyrażenia przyimkowego

PP → P NN - wyrażenie przyimkowe składa się z przyimka i rzeczownika

Poprzez wykorzystanie reguł można generować wypowiedzi w danym języku z użyciem dostępnych symboli terminalnych. Zbiór reguł składniowych nazywa się **syntaktyką** języka.

Gramatyka bezkontekstowa

Przykład wykorzystujący gramatykę bezkontekstową:

symbole terminalne – {*kot, na, czarny, wszedł, płot*}

symbole nieterminalne – {VB,NN,VP,NP,PP ,P,ADJ}

wykonaie reguł:

S → NP VP

S →ADJ NN VP

S →ADJ NN VV PP

S → ADJ NN VV P NN

NP → ADJ NN

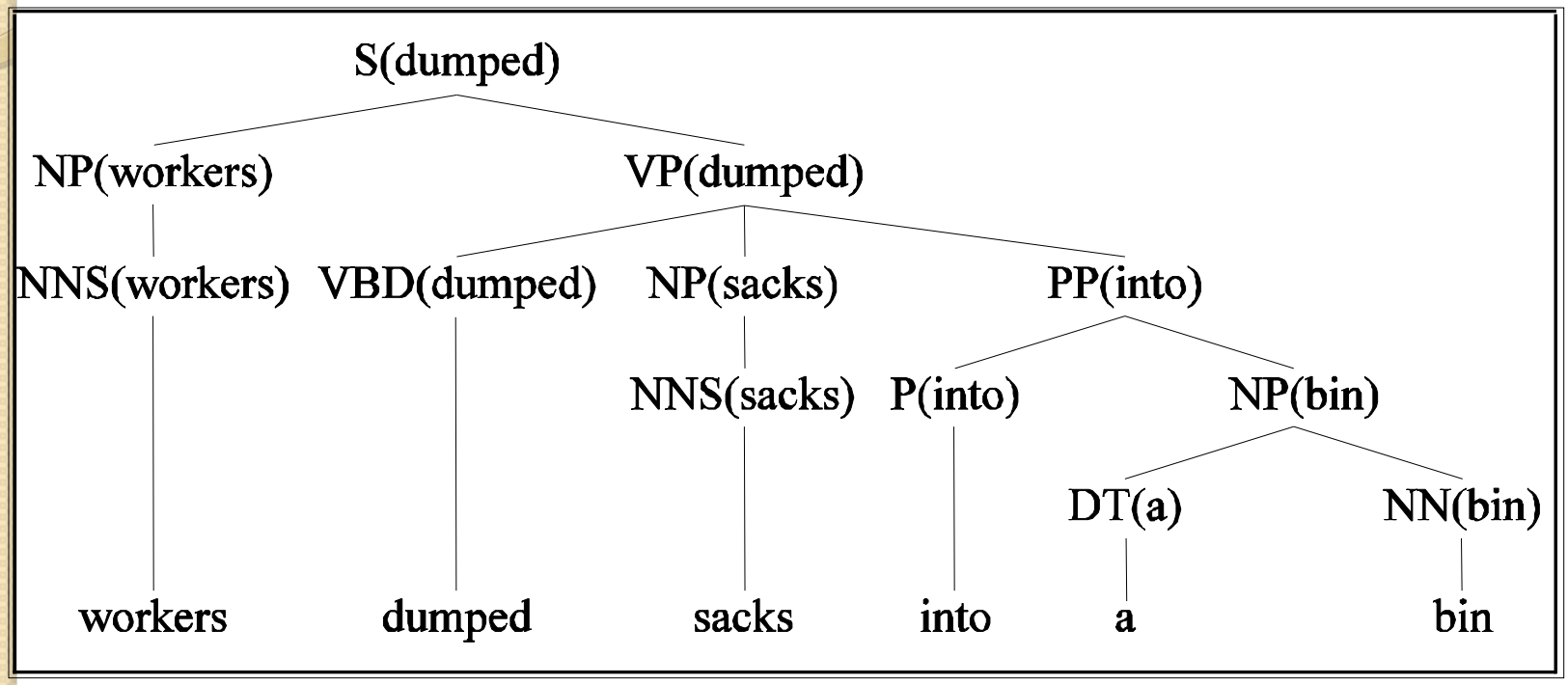
VP →VV PP

PP → P NN

Wynik:

Czarny kot wszedł na płot.

Rozbiór syntaktyczny zdania



Generowanie tekstu

Generowanie języka naturalnego – *Natural Language Generation* – dział nauki zajmujący się zamianą komputerowej reprezentacji wiedzy na tekst w języku naturalnym. Jest to problem dualny do rozumienia języka naturalnego. Pokrewną dziedziną jest automatyczne streszczanie tekstu (*automatic text summarization*), którego zadaniem jest generowanie opisu na podstawie zawartości semantycznej tekstu.

Generowanie tekstu

Większość systemów NLG działa na zasadzie prezentacji informacji o konkretnych danych w formie tekstowej, np.:

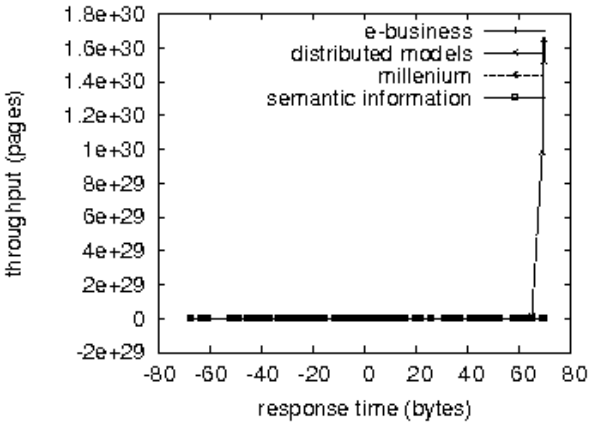
- prognoz pogody,
- danych technicznych,
- rozkładów jazdy,
- terminarzy,
- procedur postępowania.

Wykorzystując dane i związane z nimi symbole terminalne, systemy generacji tekstu wykorzystują gramatykę bezkontekstową do generowania wypowiedzi w języku naturalnym.

Generowanie tekstu

Humorystyczny przykład: generator publikacji naukowych SCIGen:

We ran our application on commodity operating systems, such as GNU/Debian Linux Version 9.2 and NetBSD Version 8a. all software components were hand hex-editted using Microsoft developer's studio built on Timothy Leary's toolkit for lazily refining Bayesian dot-matrix printers [8,29,19]. Our experiments soon proved that making autonomous our partitioned 2400 baud modems was more effective than interposing on them, as previous work suggested. Our experiments soon proved that microkernelizing our randomly distributed dot-matrix printers was more effective than reprogramming them, as previous work suggested. This concludes our discussion of software modifications.



Generator dostępny jest pod adresem:

http://pdos.csail.mit.edu/scigen/

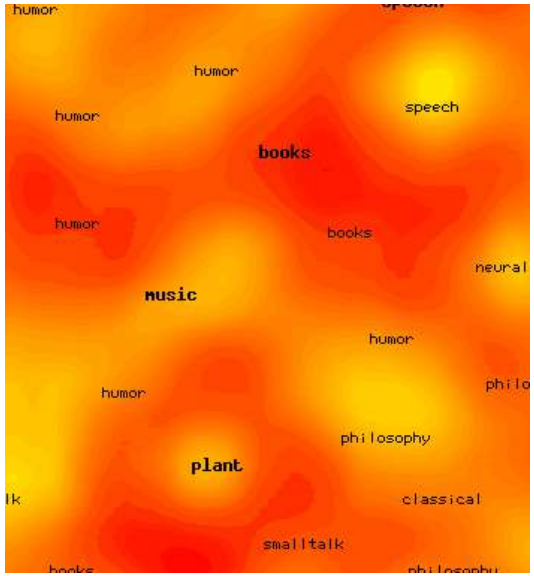
Wykorzystuje gramatykę bezkontekstową (*Context-Free Grammar)*

Szukanie semantyczne

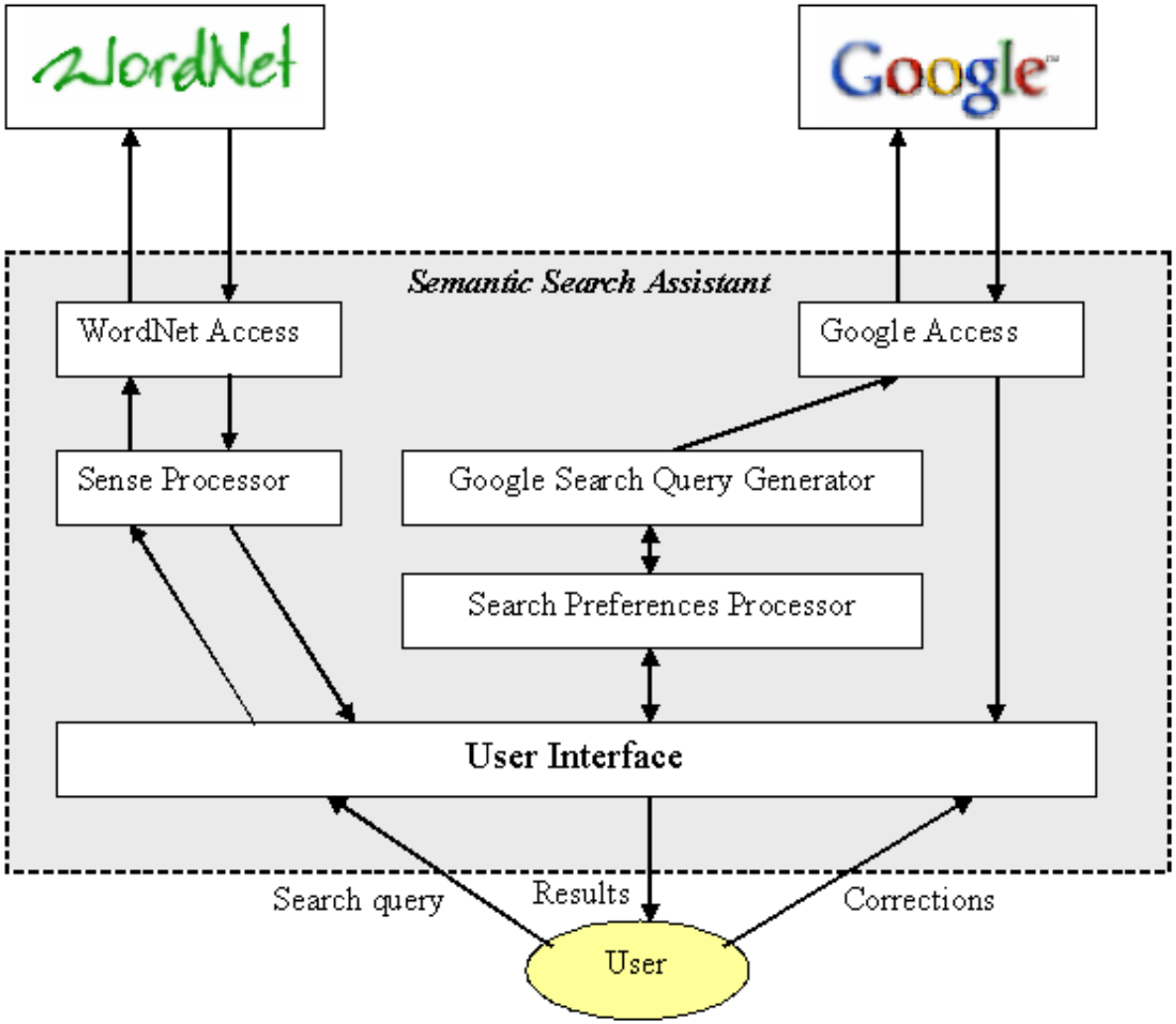
Jest to nowa metoda, wchodząca w skład technologii Web 3.0 polegająca na szukaniu informacji nie na zasadzie porównywania tekstu, a cech znaczeniowych.

Zadaniem wyszukiwania semantycznego jest „zrozumienie” zapytania i szukanie odpowiedzi na konkretny problem.

Jedną z technik jest tworzenie samoorganizujących się map (SOM), w których zawartość sieci podzielona jest tematycznie.



Szukanie semantyczne

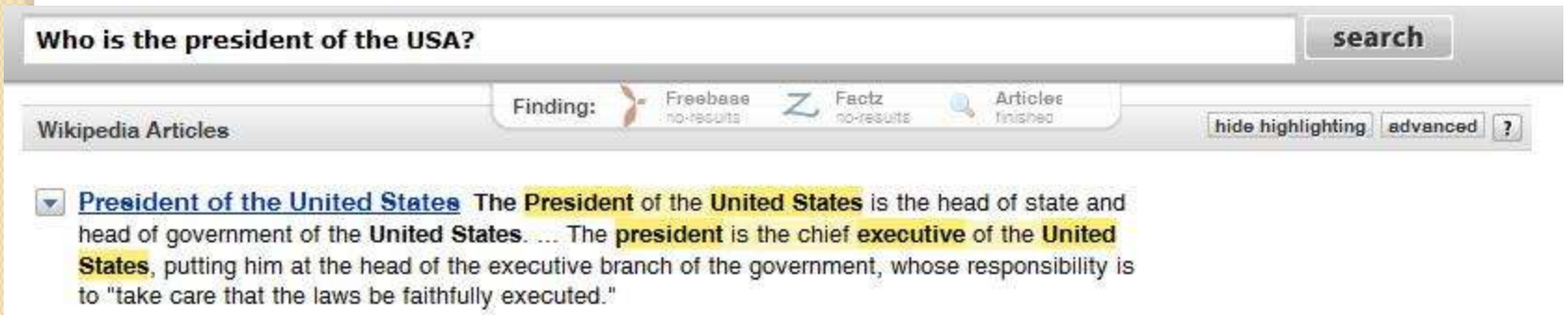


Szukanie semantyczne

Pierwszą wyszukiwarką, która uruchomiła w sieci wyszukiwanie semantyczne jest PowerSet Microsoftu - www.powerset.com

Przykład:

z większością pytań (nawet prostych) wyszukiwarka jednak sobie nie radzi:

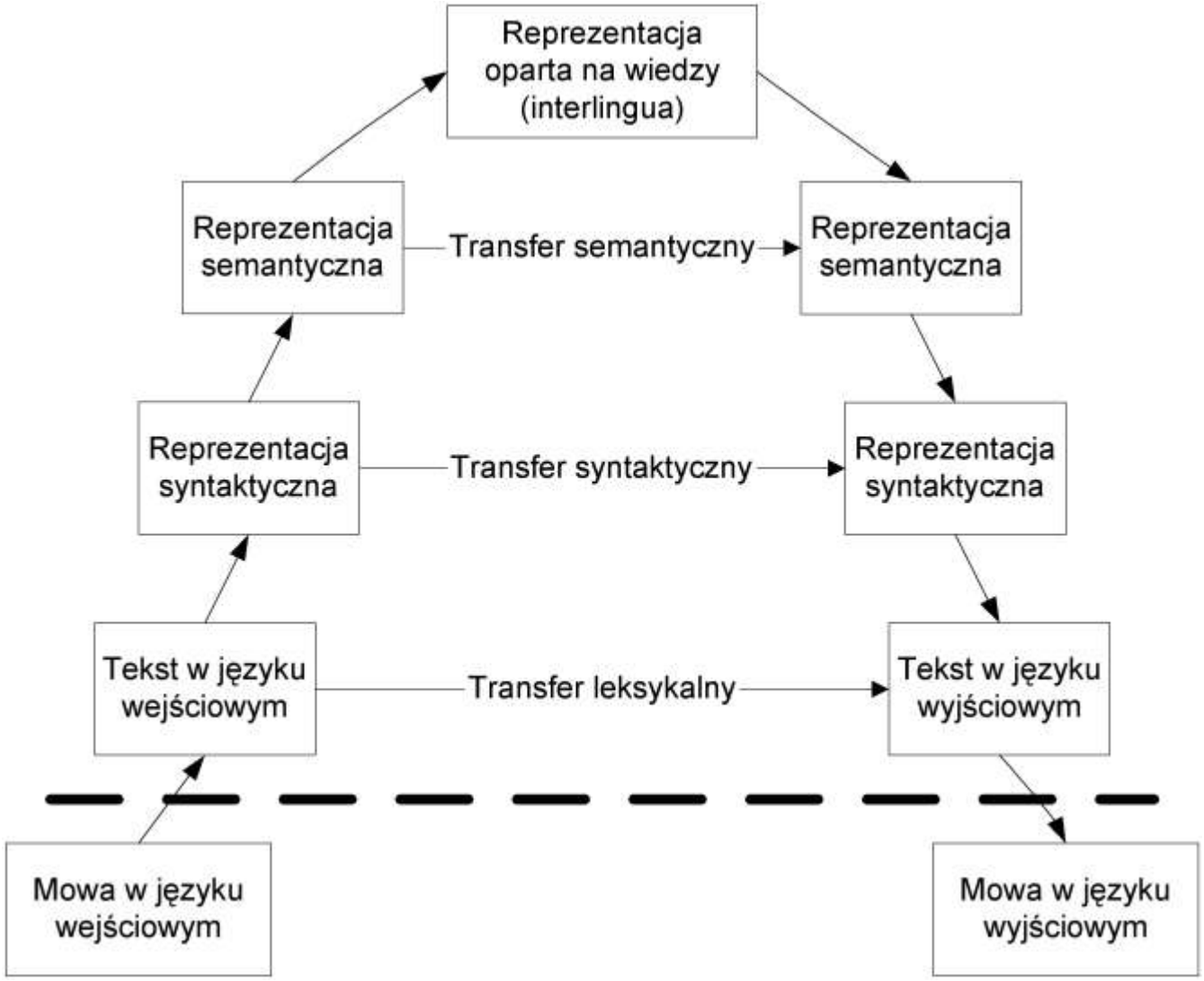


Tłumaczenie maszynowe

Trzy podejścia:

- *Machine Aided Human Translation (MAHT)* – tłumaczenie przez człowieka wspomagane maszynowo – tłumacz korzysta z zestawu narzędzi programowych ułatwiających tłumaczenie (*Computer Aided Translation*).
- *Human Aided Machine Translation (HAMT)* – tłumaczenie maszynowe wspomagane przez człowieka – człowiek podaje komputerowi fragmenty tekstu w formie łatwiejszej do przetworzenia przez komputer i „poprawia" tekst wygenerowany przez maszynę.
- *Fully Automated Machine Translation (FAMT)* – tłumaczenie całkowicie maszynowe – komputer otrzymuje tekst w niezmienionej formie i automatycznie zwraca tekst w innym języku.

Metody tłumaczenia maszynowego



źródło: http://kf.mish.uw.edu.pl/kog/kog_seb.pdf

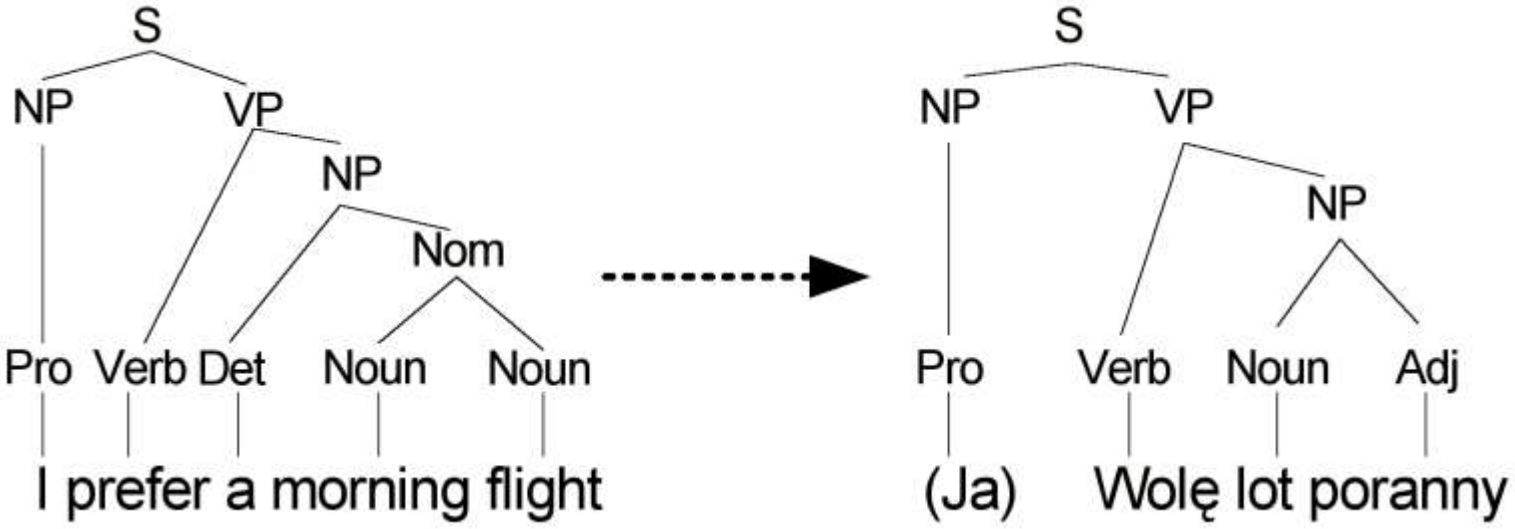
Metody tłumaczenia maszynowego

Transfer leksykalny – tłumaczenie „słowo po słowie”. Z rozwiązaniem tym wiąże się bardzo dużo problemów:

- różna postać wyrazu w zależności od formy fleksyjnej,
- braki leksykalne – gdy wyraz nie ma odpowiednika w innym języku należy się posłużyć peryfrazą (omówieniem),
- homonimia – wyrazy wyglądające identycznie, ale mające różne znaczenia,
- konieczność zamiany szyku wyrazów w przetłumaczonym tekście,
- konieczność transliteracji, gdy w językach występują różne systemy znaków.

Metody tłumaczenia maszynowego

Transfer syntaktyczny - przekładanie słów z jednego języka na drugi z zachowaniem form gramatycznych. Polega na analizie syntaktycznej zdania wejściowego i przetłumaczenie odpowiednich elementów drzewa syntaktycznego. Pozwala na zastosowanie do przetłumaczonej wypowiedzi reguł syntaktycznych właściwych dla danego języka.



Metody tłumaczenia maszynowego

Transfer semantyczny – w niektórych językach forma wyrazów zależy od kontekstu znaczeniowego. Transfer semantyczny uzależnia treść przetłumaczonej wypowiedzi od zaawansowanych cech znaczeniowych tłumaczonego tekstu.

Tłumaczenie przez reprezentację wiedzy – polega na generacji tekstu w obcym języku na bazie wiedzy pozyskanej z wypowiedzi w języku macierzystym.

Przykładowe systemy NLP

Istnieje szereg bibliotek programistycznych zawierających funkcję z dziedziny przetwarzania języka naturalnego.

AlchemyAPI – C, C++, C#, Java, Python, Perl

OpenNLP – Java,

Stanford NLP – Java,

Natural Language Toolkit – Python,

Programy do rozmowy z komputerem

Infoboty – programy udzielające w formie rozmowy informacji na konkretny temat,

chatterboty – programy zaprojektowane do rozmowy „na każdy temat”.

Najczęściej działają na zasadzie nieskomplikowanej analizy tekstu i schematów konwersacyjnych, zazwyczaj zapisanych w standardzie AIML.

AIML

AIML (*Artificial Intelligence Markup Language*) to sposób opisu tekstu ułatwiający automatyczne rozumienie tekstu przez systemy sztucznej inteligencji.

Elementy:

<aiml> znacznik początku i końca dokumentu AIML.

<category> znacznik obejmujący „jednostkę wiedzy" dostępnej w systemie.

<pattern> znacznik obejmujący zapytanie, które może wprowadzić użytkownik.

<template> znacznik obejmujący możliwe odpowiedzi systemu na dane zapytanie.

AIML

Przykład (A.L.I.C.E):

```
<aiml>
<category>
        <pattern>WHAT ARE YOU</pattern>
        <template>
            <think><set name="topic"> Me </set>
            </think>
             I am the latest result in artificial intelligence,
            which can reproduce the capabilities of the
            human brain with greater speed and accuracy.
        </template>
</category>
</aiml>
```

Test Turinga

W 1950 roku Alan Turing zaproponował test, będący sprawdzianem możliwości komputera w zakresie rozumienia i generowania języka naturalnego. Maszyna przechodzi pozytywnie test, jeżeli sędzia nie jest w stanie odróżnić, czy rozmawia z komputerem, czy z człowiekiem.

Programy do rozmowy z komputerem

ELIZA – program stworzony w 1967 roku, symulujący zachowania psychoanalityka. ELIZA dokonuje prostej analizy wypowiedzi, wyodrębnia z niej słowo kluczowe (nie analizując jego znaczenia) i odpowiada najczęściej pytaniem związanym z tym słowem lub nic niewnoszącym do rozmowy otwartym zwrotem. Cały kod programu zajmuje 240 linii.

http://www.cyberpsych.org/eliza/

Programy do rozmowy z komputerem

A.L.I.C.E. – (*Artificial Linguistic Internet Computer Entity*) – chatterbot uznawany za jeden z bardziej zaawansowanych. Wielokrotnie uzyskiwał najlepszy wynik w teście Turinga. Teoretycznie potrafi rozmawiać na 40 tysięcy tematów. Wykorzystuje *Context-Free Grammar*, język AIML oraz potrafi wykorzystywać wiedzę zdobytą w rozmowie z użytkownikiem.

A.L.I.C.E. wykorzystuje animowane awatary SitePal.

http://alice.pandorabots.com/