„Jajem, tyjesz, wyjecie” czyli „Dlaczego dialog z komputerem jest taki trudny?”

Aleksander Pohl
http://apohllo.pl

Katedra Informatyki, Akademia Górniczo-Hutnicza

SFI 8. marca 2008

Plan prezentacji

Plan prezentacji

Kilka słów o mnie :-)

- Zapalony programista **Rubiego**
- Entuzjasta **lingwistyki komputerowej**
- Obecnie pracuję głównie w projekcie: **Polska Platforma Bezpieczeństwa Wewnętrznego** (używamy Javy :-/)
- `http://apohllo.pl`

Test Turinga

Alan Turing w 1950 roku sformułował słynny test, który mógłby dać odpowiedź na pytanie: „Czy maszyna jest inteligenta?”.

- Podstawowa idea: ludzka inteligencja przejawia się w *dialogu*

Jeśli komunikując się z komputerem w języku naturalnym nie jesteśmy w stanie zorientować się, że mamy do czynienia z maszyną, to możemy uznać, że jest on inteligentny.
Do dziś żaden system *nie przeszedł testu Turinga*.

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Lingwistyka komputerowa

- obszar wiedzy na przecięciu językoznawstwa i sztucznej inteligencji
- rozwiązanie problemów językowych na potrzeby:
  - **komunikacji człowieka z maszyną** – np. interfejs w języku polskim do bazy danych o filmach
  - **maszynowego tłumaczenie tekstów** – jeden z pierwszych problemów LK: tłumaczenie z rosyjskiego na angielski
  - **ekstrakcji informacji** – np. śledzenie informacji prasowych dotyczących nowych wdzianek Dody
  - **wyszukiwania informacji** – np. użycie zamiast słów kluczowych pytań, opisu pożądanego rezultatu
  - **automatycznej akwizycji wiedzy** – np. robot kuchenny czytający książkę kulinarną

Problemy lingwistyki komputerowej 1

- **rozpoznawanie mowy/pisma** – przekształcanie ciągu dźwięków/obrazu na litery/wyrazy/zdania
- **synteza mowy** – przekształcanie napisów w dźwięki
- **analiza morfologiczna** – przyporządkowanie słowom ich kategorii gramatycznej (np. *koń* – M l. p., r. męski żywotny)
- **ujednoznacznianie sensu** – przyporządkowanie słowu pojęcia (np. *zamek* – określenie czy chodzi o zamek jako budowlę, czy zamek w drzwiach)
- **analiza syntaktyczna** – rozpoznanie gramatycznej struktury zdania, zachodzących w nim związków

Problemy lingwistyki komputerowej 1

- **rozpoznawanie mowy/pisma** – przekształcanie ciągu dźwięków/obrazu na litery/wyrazy/zdania
- **synteza mowy** – przekształcanie napisów w dźwięki
- **analiza morfologiczna** – przyporządkowanie słowom ich kategorii gramatycznej (np. *koń* – M l. p., r. męski żywotny)
- **ujednoznacznianie sensu** – przyporządkowanie słowu pojęcia (np. *zamek* – określenie czy chodzi o zamek jako budowlę, czy zamek w drzwiach)
- **analiza syntaktyczna** – rozpoznanie gramatycznej struktury zdania, zachodzących w nim związków

Problemy lingwistyki komputerowej 1

- **rozpoznawanie mowy/pisma** – przekształcanie ciągu dźwięków/obrazu na litery/wyrazy/zdania
- **synteza mowy** – przekształcanie napisów w dźwięki
- **analiza morfologiczna** – przyporządkowanie słowom ich kategorii gramatycznej (np. *koń* – M l. p., r. męski żywotny)
- **ujednoznacznianie sensu** – przyporządkowanie słowu pojęcia (np. *zamek* – określenie czy chodzi o zamek jako budowlę, czy zamek w drzwiach)
- **analiza syntaktyczna** – rozpoznanie gramatycznej struktury zdania, zachodzących w nim związków

Problemy lingwistyki komputerowej 1

- **rozpoznawanie mowy/pisma** – przekształcanie ciągu dźwięków/obrazu na litery/wyrazy/zdania
- **synteza mowy** – przekształcanie napisów w dźwięki
- **analiza morfologiczna** – przyporządkowanie słowom ich kategorii gramatycznej (np. *koń* – M l. p., r. męski żywotny)
- **ujednoznacznianie sensu** – przyporządkowanie słowu pojęcia (np. *zamek* – określenie czy chodzi o zamek jako budowlę, czy zamek w drzwiach)
- **analiza syntaktyczna** – rozpoznanie gramatycznej struktury zdania, zachodzących w nim związków

Problemy lingwistyki komputerowej 1

- **rozpoznawanie mowy/pisma** – przekształcanie ciągu dźwięków/obrazu na litery/wyrazy/zdania
- **synteza mowy** – przekształcanie napisów w dźwięki
- **analiza morfologiczna** – przyporządkowanie słowom ich kategorii gramatycznej (np. *koń* – M I. p., r. męski żywotny)
- **ujednoznacznianie sensu** – przyporządkowanie słowu pojęcia (np. *zamek* – określenie czy chodzi o zamek jako budowlę, czy zamek w drzwiach)
- **analiza syntaktyczna** – rozpoznanie gramatycznej struktury zdania, zachodzących w nim związków

Problemy lingwistyki komputerowej 2

- **rozpoznawanie obiektów nazwanych** – właściwa interpretacja nazw własnych i ich skrótów (np. Rzeczpospolita Polska, Polska, RP)
- **rozpoznawanie wyrażeń koekstensywnych** – wykrywanie wyrażeń odnoszących się do tych samych obiektów (np. „*Leo* kupił Volkswagena. *On* lubi niemieckie samochody.”)
- **tłumaczenie słów/wyrażeń** – wybór właściwego tłumaczenia spośród wielu możliwych
  - „Jadę samochodem.” – „I *drive* a car.”
  - „Jadę rowerem.” – „I *ride* a bicycle.”

Problemy lingwistyki komputerowej 2

- **rozpoznawanie obiektów nazwanych** – właściwa interpretacja nazw własnych i ich skrótów (np. Rzeczpospolita Polska, Polska, RP)
- **rozpoznawanie wyrażeń koekstensywnych** – wykrywanie wyrażeń odnoszących się do tych samych obiektów (np. „*Leo* kupił Volkswagena. *On* lubi niemieckie samochody.”)
- **tłumaczenie słów/wyrażeń** – wybór właściwego tłumaczenia spośród wielu możliwych
  - „Jadę samochodem.” – „I *drive* a car.”
  - „Jadę rowerem.” – „I *ride* a bicycle.”

Problemy lingwistyki komputerowej 2

- **rozpoznawanie obiektów nazwanych** – właściwa interpretacja nazw własnych i ich skrótów (np. Rzeczpospolita Polska, Polska, RP)
- **rozpoznawanie wyrażeń koekstensywnych** – wykrywanie wyrażeń odnoszących się do tych samych obiektów (np. „*Leo* kupił Volkswagena. *On* lubi niemieckie samochody.”)
- **tłumaczenie słów/wyrażeń** – wybór właściwego tłumaczenia spośród wielu możliwych
  - „Jadę samochodem.” – „I *drive* a car.”
  - „Jadę rowerem.” – „I *ride* a bicycle.”

Problemy lingwistyki komputerowej 3

- **analiza semantyczna** – interpretacja zdania w pewny języku formalnym (np. logice predykatów, logice deskryptywnej), odniesienie wyrażeń do rzeczywistości pozajęzykowej
- **analiza dyskursu** – uwzględniania pragmatycznych aspektów komunikacji (np. dostosowanie komunikatów do zasobu słownictwa użytkownika)
- **synteza zdań** – tworzenie zdań na podstawie elementarnych danych, notacji logicznej

Problemy lingwistyki komputerowej 3

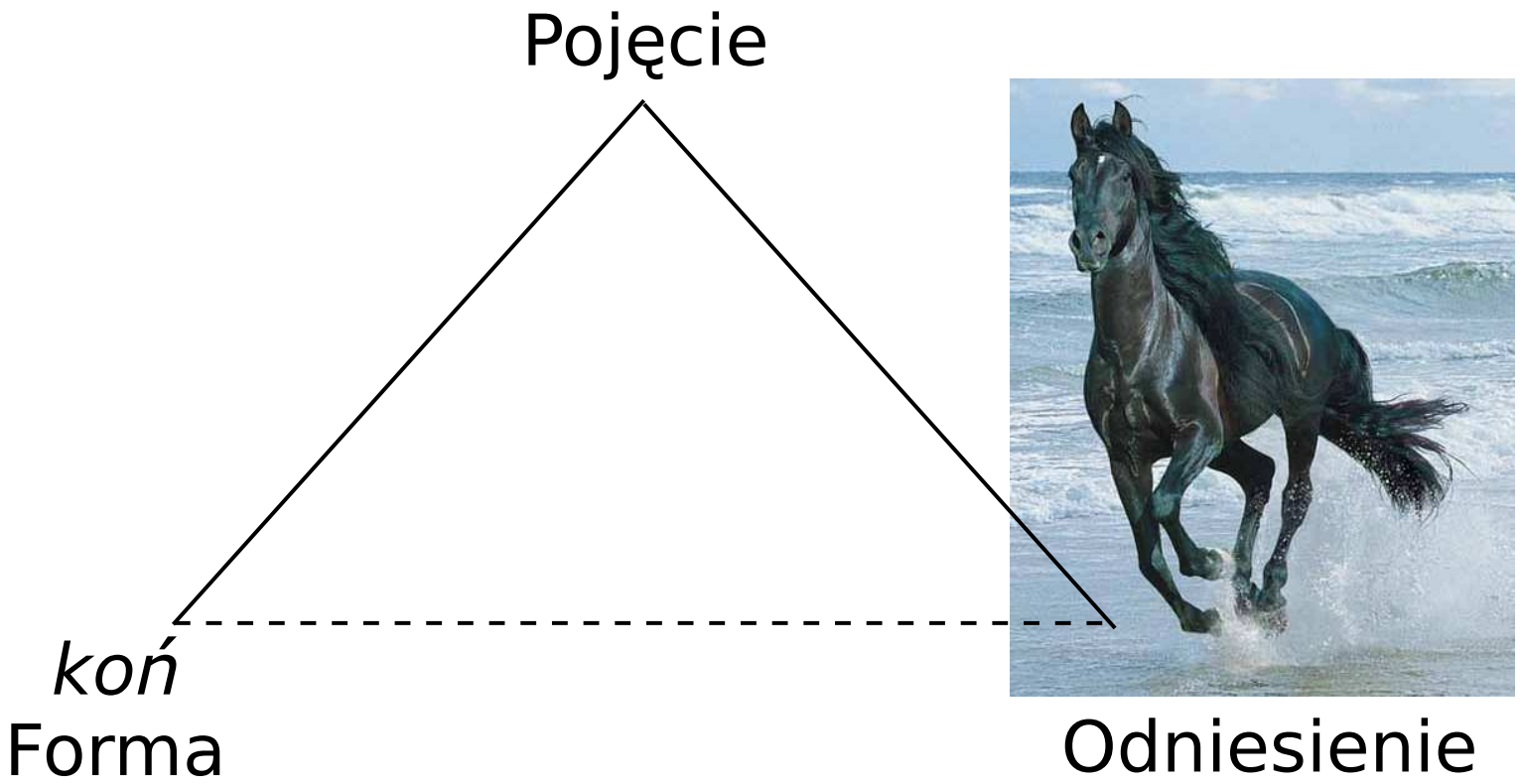
- **analiza semantyczna** – interpretacja zdania w pewny języku formalnym (np. logice predykatów, logice deskryptywnej), odniesienie wyrażeń do rzeczywistości pozajęzykowej
- **analiza dyskursu** – uwzględniania pragmatycznych aspektów komunikacji (np. dostosowanie komunikatów do zasobu słownictwa użytkownika)
- **synteza zdań** – tworzenie zdań na podstawie elementarnych danych, notacji logicznej

Problemy lingwistyki komputerowej 3

- **analiza semantyczna** – interpretacja zdania w pewny języku formalnym (np. logice predykatów, logice deskryptywnej), odniesienie wyrażeń do rzeczywistości pozajęzykowej
- **analiza dyskursu** – uwzględniania pragmatycznych aspektów komunikacji (np. dostosowanie komunikatów do zasobu słownictwa użytkownika)
- **synteza zdań** – tworzenie zdań na podstawie elementarnych danych, notacji logicznej

Plan prezentacji

Trójkąt semiotyczny



Wieloznaczność

Właściwie *brak jednoznaczności*.
Przykład znany wszystkim informatykom:

- Jak porównujemy liczby całkowite?
  $a == 10$
- Jak porównujemy liczby rzeczywiste?
  $a == 10.0$ Nie!
  $|a - 10.0| \leq DELTA$

Pojedyncza liczba rzeczywista w zapisie binarnym reprezentuje *przedział* liczb rzeczywistych.
Pojedynczy znak może być zinterpretowany na wiele sposobów, może reprezentować *różne pojęcia* a przez to *różne odniesienia*.

Wieloznaczność

Właściwie *brak jednoznaczności*.
Przykład znany wszystkim informatykom:

- Jak porównujemy liczby całkowite?
  $a == 10$
- Jak porównujemy liczby rzeczywiste?
  $a == 10.0$ Nie!
  $|a - 10.0| \leq DELTA$

Pojedyncza liczba rzeczywista w zapisie binarnym reprezentuje *przedział* liczb rzeczywistych.
Pojedynczy znak może być zinterpretowany na wiele sposobów, może reprezentować *różne pojęcia* a przez to *różne odniesienia*.

Wieloznaczność

Właściwie *brak jednoznaczności*.
Przykład znany wszystkim informatykom:

- Jak porównujemy liczby całkowite?
  $a == 10$
- Jak porównujemy liczby rzeczywiste?
  $a == 10.0$ Nie!
  $|a - 10.0| \leq DELTA$

Pojedyncza liczba rzeczywista w zapisie binarnym reprezentuje *przedział* liczb rzeczywistych.
Pojedynczy znak może być zinterpretowany na wiele sposobów, może reprezentować *różne pojęcia* a przez to *różne odniesienia*.

Źródła wieloznaczności

- przetworzenie obrazu/dźwięku na tekst
- fleksja
- wyrażenia złożone
- struktura syntaktyczna
- homonimia
- tłumaczenie

Obraz/dźwięk → tekst

- rozpoznanie fonemów – np. głoski dźwięczne i bezdźwięczne: p/b, t/d, etc.
- rozpoznanie końca wyrazu:
  - jajem / ja jem ?
  - tyjesz / ty jesz ?
  - wyjecie / wy jecie ?
- zapis ortograficzny – *čarny* → czarny
  może → może / morze ?
- **Rozwiązanie**: ukryte modele Markowa (Hidden Markov Models).
- **Przykład zastosowania**: interfejs głosowy komputera Mac (niestety tylko w języku angielskim).

Obraz/dźwięk → tekst

- rozpoznanie fonemów – np. głoski dźwięczne i bezdźwięczne: p/b, t/d, etc.
- rozpoznanie końca wyrazu:
  - jajem / ja jem ?
  - tyjesz / ty jesz ?
  - wyjecie / wy jecie ?
- zapis ortograficzny – *čarny* → czarny
  może → może / morze ?
- **Rozwiązanie**: ukryte modele Markowa (Hidden Markov Models).
- **Przykład zastosowania**: interfejs głosowy komputera Mac (niestety tylko w języku angielskim).

Obraz/dźwięk → tekst

- rozpoznanie fonemów – np. głoski dźwięczne i bezdźwięczne: p/b, t/d, etc.
- rozpoznanie końca wyrazu:
  - jajem / ja jem ?
  - tyjesz / ty jesz ?
  - wyjecie / wy jecie ?
- zapis ortograficzny – *čarny* → czarny
  może → może / morze ?
- **Rozwiązanie**: ukryte modele Markowa (Hidden Markov Models).
- **Przykład zastosowania**: interfejs głosowy komputera Mac (niestety tylko w języku angielskim).

Obraz/dźwięk → tekst

- rozpoznanie fonemów – np. głoski dźwięczne i bezdźwięczne: p/b, t/d, etc.
- rozpoznanie końca wyrazu:
  - jajem / ja jem ?
  - tyjesz / ty jesz ?
  - wyjecie / wy jecie ?
- zapis ortograficzny – *čarny* → czarny
  może → może / morze ?
- **Rozwiązanie**: ukryte modele Markowa (Hidden Markov Models).
- **Przykład zastosowania**: interfejs głosowy komputera Mac (niestety tylko w języku angielskim).

Obraz/dźwięk → tekst

- rozpoznanie fonemów – np. głoski dźwięczne i bezdźwięczne: p/b, t/d, etc.
- rozpoznanie końca wyrazu:
  - jajem / ja jem ?
  - tyjesz / ty jesz ?
  - wyjecie / wy jecie ?
- zapis ortograficzny – *čarny* → czarny
  może → może / morze ?
- **Rozwiązanie**: ukryte modele Markowa (Hidden Markov Models).
- **Przykład zastosowania**: interfejs głosowy komputera Mac (niestety tylko w języku angielskim).

Fleksja

W językach fleksyjnych (np. polskim) słowa odmieniają się przez osoby, czasy, liczby, rodzaje, przypadki, stopnie:

- **jeść**: ja *jem*, ty *jesz*, on/ona/ono *je*, my *jemy*, wy *jecie*, oni *jedzą*, . . . (45 form)
- **kot**: M. *kot*, D. *kota*, C. *kotu*, B. *kota*, N. *kotem*, Ms. *kocie*, W. *kocie*, . . .
- **dobry**: r.m., st. równy M. *dobry*, D. *dobrego*, . . .
  r.m., st. wyższy M. *lepszy*, D. *lepszego*, . . .
  r.ż., st. równy M. *dobra*, D. *dobrej*, . . .
- **bardzo**: st. równy *bardzo*, st. wyższy *bardziej*, st. najwyższy *najbardziej*

Fleksja

W językach fleksyjnych (np. polskim) słowa odmieniają się przez osoby, czasy, liczby, rodzaje, przypadki, stopnie:

- **jeść**: ja *jem*, ty *jesz*, on/ona/ono *je*, my *jemy*, wy *jecie*, oni *jedzą*, . . . (45 form)
- **kot**: M. *kot*, D. *kota*, C. *kotu*, B. *kota*, N. *kotem*, Ms. *kocie*, W. *kocie*, . . .
- **dobry**: r.m., st. równy M. *dobry*, D. *dobrego*, . . .
  r.m., st. wyższy M. *lepszy*, D. *lepszego*, . . .
  r.ż., st. równy M. *dobra*, D. *dobrej*, . . .
- **bardzo**: st. równy *bardzo*, st. wyższy *bardziej*, st. najwyższy *najbardziej*

Fleksja

W językach fleksyjnych (np. polskim) słowa odmieniają się przez osoby, czasy, liczby, rodzaje, przypadki, stopnie:

- **jeść**: ja *jem*, ty *jesz*, on/ona/ono *je*, my *jemy*, wy *jecie*, oni *jedzą*, . . . (45 form)
- **kot**: M. *kot*, D. *kota*, C. *kotu*, B. *kota*, N. *kotem*, Ms. *kocie*, W. *kocie*, . . .
- **dobry**: r.m., st. równy M. *dobry*, D. *dobrego*, . . .
  r.m., st. wyższy M. *lepszy*, D. *lepszego*, . . .
  r.ż., st. równy M. *dobra*, D. *dobrej*, . . .
- **bardzo**: st. równy *bardzo*, st. wyższy *bardziej*, st. najwyższy *najbardziej*

Fleksja

W językach fleksyjnych (np. polskim) słowa odmieniają się przez osoby, czasy, liczby, rodzaje, przypadki, stopnie:

- **jeść**: ja *jem*, ty *jesz*, on/ona/ono *je*, my *jemy*, wy *jecie*, oni *jedzą*, . . . (45 form)
- **kot**: M. *kot*, D. *kota*, C. *kotu*, B. *kota*, N. *kotem*, Ms. *kocie*, W. *kocie*, . . .
- **dobry**: r.m., st. równy M. *dobry*, D. *dobrego*, . . .
  r.m., st. wyższy M. *lepszy*, D. *lepszego*, . . .
  r.ż., st. równy M. *dobra*, D. *dobrej*, . . .
- **bardzo**: st. równy *bardzo*, st. wyższy *bardziej*, st. najwyższy *najbardziej*

Fleksja

W językach fleksyjnych (np. polskim) słowa odmieniają się przez osoby, czasy, liczby, rodzaje, przypadki, stopnie:

- **jeść**: ja *jem*, ty *jesz*, on/ona/ono *je*, my *jemy*, wy *jecie*, oni *jedzą*, . . . (45 form)
- **kot**: M. *kot*, D. *kota*, C. *kotu*, B. *kota*, N. *kotem*, Ms. *kocie*, W. *kocie*, . . .
- **dobry**: r.m., st. równy M. *dobry*, D. *dobrego*, . . .
  r.m., st. wyższy M. *lepszy*, D. *lepszego*, . . .
  r.ż., st. równy M. *dobra*, D. *dobrej*, . . .
- **bardzo**: st. równy *bardzo*, st. wyższy *bardziej*, st. najwyższy *najbardziej*

Opis morfologiczny i forma podstawowa

Podstawowy problem: dla danej formy znaleźć jej *opis morfologiczny* oraz *formę podstawową*:

- »Szli przez **ciemny** las« – biernik liczby pojedynczej przymiotnika *ciemny*
- »Spotkali czarownicę z **kotem**« – narzędnik liczby pojedynczej rzeczownika *kot*
- »Nie **czekając** odrzekł: „Ja biorę czarownicę a ty kota.”« – imiesłów współczesny czasownika *czekać*

Rozwiązanie: wykorzystanie *stemmera*, stworzenie *słownika fleksyjnego*

Opis morfologiczny i forma podstawowa

Podstawowy problem: dla danej formy znaleźć jej *opis morfologiczny* oraz *formę podstawową*:

- »Szli przez **ciemny** las« – biernik liczby pojedynczej przymiotnika *ciemny*
- »Spotkali czarownicę z **kotem**« – narzędnik liczby pojedynczej rzeczownika *kot*
- »Nie **czekając** odrzekł: „Ja biorę czarownicę a ty kota.”« – imiesłów współczesny czasownika *czekać*

Rozwiązanie: wykorzystanie *stemmera*, stworzenie *słownika fleksyjnego*

Opis morfologiczny i forma podstawowa

Podstawowy problem: dla danej formy znaleźć jej *opis morfologiczny* oraz *formę podstawową*:

- »Szli przez **ciemny** las« – biernik liczby pojedynczej przymiotnika *ciemny*
- »Spotkali czarownicę z **kotem**« – narzędnik liczby pojedynczej rzeczownika *kot*
- »Nie **czekając** odrzekł: „Ja biorę czarownicę a ty kota.”« – imiesłów współczesny czasownika *czekać*

Rozwiązanie: wykorzystanie *stemmera*, stworzenie *słownika fleksyjnego*

Słownik fleksyjny języka polskiego

Fleksja – wieloznaczność

Czy to wystarczy?

- »**Wyjecie** jak wyjce afrykańskie!« – forma podstawowa *wyć*
- »Zanim **wyjecie** wszystko z talerza, umyjcie ręce.« – forma podstawowa *wyjeść*

Jednej formie można przypisać *wiele form podstawowych*.

- »Zjadłem dziś dwa **jaja**.« – *biernik liczby mnogiej* rzeczownika jajo
- »Nikt nie mógł znaleźć skradzionego **jaja** Fabergé.« – *dopełniacz liczby pojedynczej* rzeczownika jajo

Jednej formie można przyporządkować *wiele opisów morfologicznych*.

Fleksja – wieloznaczność

Czy to wystarczy?

- »**Wyjecie** jak wyjce afrykańskie!« – forma podstawowa *wyć*
- »Zanim **wyjecie** wszystko z talerza, umyjcie ręce.« – forma podstawowa *wyjeść*

Jednej formie można przypisać *wiele form podstawowych*.

- »Zjadłem dziś dwa **jaja**.« – *biernik liczby mnogiej* rzeczownika jajo
- »Nikt nie mógł znaleźć skradzionego **jaja** Fabergé.« – *dopełniacz liczby pojedynczej* rzeczownika jajo

Jednej formie można przyporządkować *wiele opisów morfologicznych*.

Fleksja – wieloznaczność

Czy to wystarczy?

- »**Wyjecie** jak wyjce afrykańskie!« – forma podstawowa *wyć*
- »Zanim **wyjecie** wszystko z talerza, umyjcie ręce.« – forma podstawowa *wyjeść*

Jednej formie można przypisać *wiele form podstawowych*.

- »Zjadłem dziś dwa **jaja**.« – *biernik liczby mnogiej* rzeczownika jajo
- »Nikt nie mógł znaleźć skradzionego **jaja** Fabergé.« – *dopełniacz liczby pojedynczej* rzeczownika jajo

Jednej formie można przyporządkować *wiele opisów morfologicznych*.

Fleksja – interpretacje

Dla zdania:
Nie ma rzeczy bardziej zwykłej i naturalnej niż to, że ludzie, którzy mają roszczenie, iż odkryli jakąś rzecz nową w świecie filozofii i nauk, sugerują innym, by chwalili ich własne systemy, osławiając jednocześnie wszystkie te, które powstały wcześniej.
otrzymujemy **120960 kombinacj**!

Rozwiązanie i zastosowania

Nie ma prostej odpowiedzi na pytanie jak rozwiązać powyższe problemy.

Można wyróżnić dwa podejścia:

- **morfosyntaktyczne** – wykorzystanie *morfosyntaktycznych* opisów słów występujących w kontekście danego słowa
- **semantyczne** – wykorzystanie *relacji semantycznych* zachodzących pomiędzy danym słowem a innymi słowami w jego kontekście

Zastosowania: wyszukiwarka Google od 2007 roku uwzględnia fleksję języka polskiego, framework Ruby on Rails *zna* fleksję języka angielskiego.

Rozwiązanie i zastosowania

Nie ma prostej odpowiedzi na pytanie jak rozwiązać powyższe problemy.
Można wyróżnić dwa podejścia:

- **morfosyntaktyczne** – wykorzystanie *morfosyntaktycznych* opisów słów występujących w kontekście danego słowa
- **semantyczne** – wykorzystanie *relacji semantycznych* zachodzących pomiędzy danym słowem a innymi słowami w jego kontekście

Zastosowania: wyszukiwarka Google od 2007 roku uwzględnia fleksję języka polskiego, framework Ruby on Rails *zna* fleksję języka angielskiego.

Rozwiązanie i zastosowania

Nie ma prostej odpowiedzi na pytanie jak rozwiązać powyższe problemy.
Można wyróżnić dwa podejścia:

- **morfosyntaktyczne** – wykorzystanie *morfosyntaktycznych* opisów słów występujących w kontekście danego słowa
- **semantyczne** – wykorzystanie *relacji semantycznych* zachodzących pomiędzy danym słowem a innymi słowami w jego kontekście

Zastosowania: wyszukiwarka Google od 2007 roku uwzględnia fleksję języka polskiego, framework Ruby on Rails *zna* fleksję języka angielskiego.

Rozwiązanie i zastosowania

Nie ma prostej odpowiedzi na pytanie jak rozwiązać powyższe problemy.
Można wyróżnić dwa podejścia:

- **morfosyntaktyczne** – wykorzystanie *morfosyntaktycznych* opisów słów występujących w kontekście danego słowa
- **semantyczne** – wykorzystanie *relacji semantycznych* zachodzących pomiędzy danym słowem a innymi słowami w jego kontekście

Zastosowania: wyszukiwarka Google od 2007 roku uwzględnia fleksję języka polskiego, framework Ruby on Rails *zna* fleksję języka angielskiego.

Semantyka

Semantyka to dział językoznawstwa, który koncentruje się na badaniu znaczenia wyrażeń języka.

Jak można opisać znaczenie słów?

- za pomocą **definicji** – w języku naturalnym lub formalnym: **jajko** 1. »żeńska komórka rozrodcza ptaka, zwykle kury, zawierająca białko i żółtko, otoczone skorupką, wykorzystywana jako produkt spożywczy«[1]
- za pomocą **relacji semantycznych** – np. *hipernimii, hiponimii, holonimii, meronimii, synonimii, sprawstwa*, etc.: **jajko** *hipernimy*: komórka, *meronimy*: żółtko, białko, . . .

[1] *Uniwersalny słownik języka polskiego* Wydawnictwo Naukowe PWN, Warszawa 2003.

Semantyka

Semantyka to dział językoznawstwa, który koncentruje się na badaniu znaczenia wyrażeń języka.
Jak można opisać znaczenie słów?

- za pomocą **definicji** – w języku naturalnym lub formalnym: **jajko** 1. »żeńska komórka rozrodcza ptaka, zwykle kury, zawierająca białko i żółtko, otoczone skorupką, wykorzystywana jako produkt spożywczy«[1]
- za pomocą **relacji semantycznych** – np. *hipernimii, hiponimii, holonimii, meronimii, synonimii, sprawstwa*, etc.: **jajko** *hipernimy*: komórka, *meronimy*: żółtko, białko, . . .

[1] *Uniwersalny słownik języka polskiego* Wydawnictwo Naukowe PWN, Warszawa 2003.

Semantyka

Semantyka to dział językoznawstwa, który koncentruje się na badaniu znaczenia wyrażeń języka.
Jak można opisać znaczenie słów?

- za pomocą **definicji** – w języku naturalnym lub formalnym: **jajko** 1. »żeńska komórka rozrodcza ptaka, zwykle kury, zawierająca białko i żółtko, otoczone skorupką, wykorzystywana jako produkt spożywczy«[1]
- za pomocą **relacji semantycznych** – np. *hipernimii, hiponimii, holonimii, meronimii, synonimii, sprawstwa*, etc.: **jajko** *hipernimy*: komórka, *meronimy*: żółtko, białko, . . .

[1] *Uniwersalny słownik języka polskiego* Wydawnictwo Naukowe PWN, Warszawa 2003.

Semantyka – wieloznaczność I

- »Kupiłem wczoraj **akcje** warte 100 tysięcy.« – akcja jako *papier wartościowy*
- »**Akcja** tej książki rozwijała się niemrawo.« – akcja jako *fabuła*

Czysta *homonimia* – jeden wyraz (w znaczeniu leksemu) posiadający wiele znaczeń

- »**Lekarz** zalecił podanie zastrzyku.«
- »**Lekarka** zaleciła podanie zastrzyku.«

Dwa wyrazy posiadające (niemal) identyczne znaczenie.

Semantyka – wieloznaczność I

- »Kupiłem wczoraj **akcje** warte 100 tysięcy.« – akcja jako *papier wartościowy*
- »**Akcja** tej książki rozwijała się niemrawo.« – akcja jako *fabuła*

Czysta *homonimia* – jeden wyraz (w znaczeniu leksemu) posiadający wiele znaczeń

- »**Lekarz** zalecił podanie zastrzyku.«
- »**Lekarka** zaleciła podanie zastrzyku.«

Dwa wyrazy posiadające (niemal) identyczne znaczenie.

Rozwiązanie

Można wyróżnić dwa podejścia:

- **statystyczne** – wykorzystanie algorytmów automatycznego uczenia do wyekstrahowania związków semantycznych na podstawie dużego korpusu tekstów.
  Wady: żaden korpus nie jest idealny, konieczne jest ręczne oznaczenie sensu wszystkich słów w korpusie
- **symboliczne** – stworzenie słownika zawierającego relacje semantyczne lub definicje formalne
  Wady: czasochłonność, brak zgodności co do tego, które relacje powinny być uwzględnione

Rozwiązanie

Można wyróżnić dwa podejścia:

- **statystyczne** – wykorzystanie algorytmów automatycznego uczenia do wyekstrahowania związków semantycznych na podstawie dużego korpusu tekstów.
  Wady: żaden korpus nie jest idealny, konieczne jest ręczne oznaczenie sensu wszystkich słów w korpusie
- **symboliczne** – stworzenie słownika zawierającego relacje semantyczne lub definicje formalne
  Wady: czasochłonność, brak zgodności co do tego, które relacje powinny być uwzględnione

Rozwiązanie – cd.

Słowniki semantyczne:

- **WordNet** – najbardziej znany słownik semantyczny języka angielskiego. Na jego podstawie tworzone są słowniki dla innych języków, np. polskiego – *Słowosieć*
  - Zalety: liczba słów, uwzględnienie wyrażeń złożonych i nazw własnych.
  - Wady: brak pokrycia w *korpusie tekstów*, brak relacji *syntagmatycznych*.
  - Przykład: **egg** *hypernyms*: ovum, egg cell; *hyponyms*: nit, spawn, roe, . . .
- **FrameNet** – słownik zawierający opis sytuacji i czynności. Na razie nie doczekał się ostatecznej wersji.

Rozwiązanie – cd.

Słowniki semantyczne:

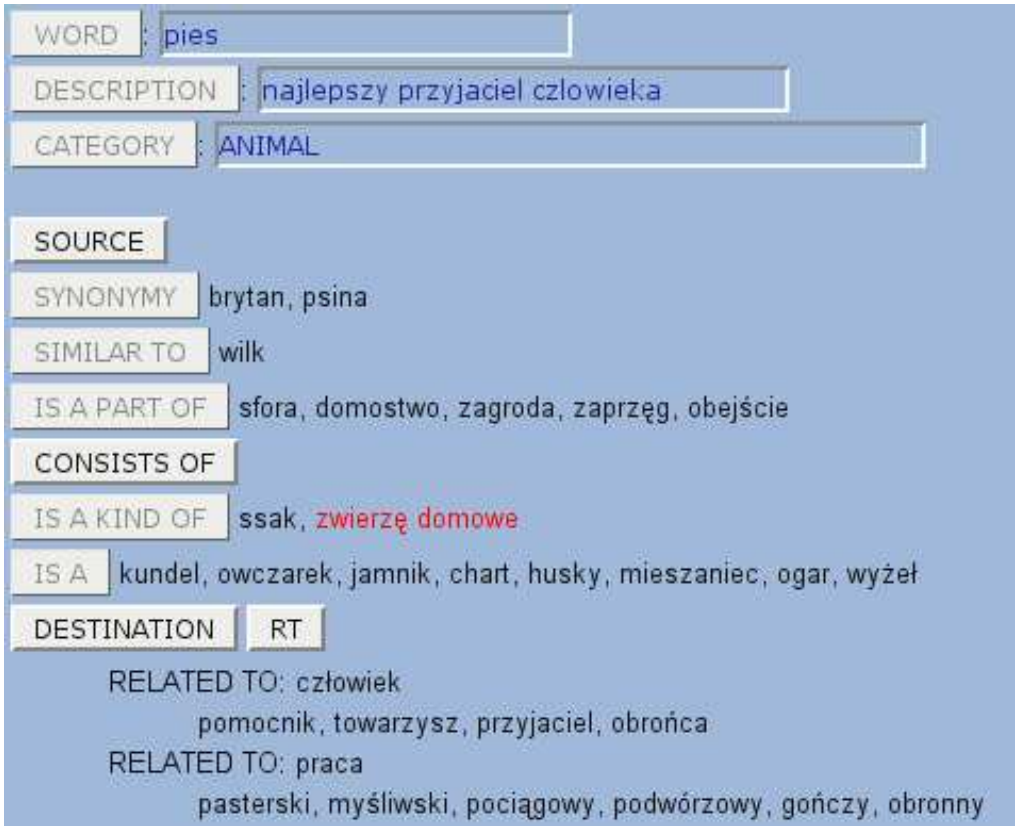
- **WordNet** – najbardziej znany słownik semantyczny języka angielskiego. Na jego podstawie tworzone są słowniki dla innych języków, np. polskiego – *Słowosieć*
  - Zalety: liczba słów, uwzględnienie wyrażeń złożonych i nazw własnych.
  - Wady: brak pokrycia w *korpusie tekstów*, brak relacji *syntagmatycznych*.
  - Przykład: **egg** *hypernyms*: ovum, egg cell; *hyponyms*: nit, spawn, roe, . . .
- **FrameNet** – słownik zawierający opis sytuacji i czynności. Na razie nie doczekał się ostatecznej wersji.

Rozwiązanie – cd.

Słowniki semantyczne:

- **WordNet** – najbardziej znany słownik semantyczny języka angielskiego. Na jego podstawie tworzone są słowniki dla innych języków, np. polskiego – *Słowosieć*
  - Zalety: liczba słów, uwzględnienie wyrażeń złożonych i nazw własnych.
  - Wady: brak pokrycia w *korpusie tekstów*, brak relacji *syntagmatycznych*.
  - Przykład: **egg** *hypernyms*: ovum, egg cell; *hyponyms*: nit, spawn, roe, . . .
- **FrameNet** – słownik zawierający opis sytuacji i czynności. Na razie nie doczekał się ostatecznej wersji.

Słownik semantyczny języka polskiego



Słownik semantyczny języka polskiego – zintegrowany z SFJP

Semantyka – wieloznaczność II

Czy to wystarczy?

- »I saw *clouds flying over* Zurich.«
  »Widziałem *chmury lecące nad* Zurychem.«
- »I saw *buildings flying over* Zurich.«
  »Widziałem *budynki lecąc nad* Zurychem.«

Tylko na podstawie wiedzy zdroworozsądkowej można właściwie przetłumaczyć powyższe przykłady.
Rozwiązanie: wykorzystanie Cyc lub Sumo.

Semantyka – wieloznaczność II

Czy to wystarczy?

- »I saw *clouds flying over* Zurich.«
  »Widziałem *chmury lecące nad* Zurychem.«
- »I saw *buildings flying over* Zurich.«
  »Widziałem *budynki lecąc nad* Zurychem.«

Tylko na podstawie wiedzy zdroworozsądkowej można właściwie przetłumaczyć powyższe przykłady.
Rozwiązanie: wykorzystanie Cyc lub Sumo.

Semantyka – wieloznaczność II

Czy to wystarczy?

- »I saw *clouds flying over* Zurich.«
  »Widziałem *chmury lecące nad* Zurychem.«
- »I saw *buildings flying over* Zurich.«
  »Widziałem *budynki lecąc nad* Zurychem.«

Tylko na podstawie wiedzy zdroworozsądkowej można właściwie przetłumaczyć powyższe przykłady.

Rozwiązanie: wykorzystanie Cyc lub Sumo.

Semantyka – wieloznaczność II

Czy to wystarczy?

- »I saw *clouds flying over* Zurich.«
  »Widziałem *chmury lecące nad* Zurychem.«
- »I saw *buildings flying over* Zurich.«
  »Widziałem *budynki lecąc nad* Zurychem.«

Tylko na podstawie wiedzy zdroworozsądkowej można właściwie przetłumaczyć powyższe przykłady.
Rozwiązanie: wykorzystanie Cyc lub Sumo.

Cyc?

Sumo?

Ontologie

- En**CYC**lopedia
- **S**uggested **U**pper **M**erged **O**ntology

Ontologia (w informatyce) to formalna specyfikacja konceptualizacji wybranej dziedziny wiedzy:

- pojęcia
- indywidua
- relacje
- funkcje
- reguły

Ontologie

- En**CYC**lopedia
- **S**uggested **U**pper **M**erged **O**ntology

Ontologia (w informatyce) to formalna specyfikacja konceptualizacji wybranej dziedziny wiedzy:

- pojęcia
- indywidua
- relacje
- funkcje
- reguły

Ontologie – cd.

Cyc oraz Sumo stanowią formalizację zdroworozsądkowej wiedzy obejmującej najbardziej ogólne własności świata.

```
CloudInSky
Mt: UniversalVocabularyMt
  isa: SpatiallyDisjointObjectType
    ExistingObjectType
Mt: TopicMt
  isa: WeatherObjects-Weather-Topic
  genls: Outdoors-ExposedToWeather
    CloudlikeObject Opaque Airborne
    TopAndBottomSidedObject
```

Ontologie – cd.

Cyc oraz Sumo stanowią formalizację zdroworozsądkowej wiedzy obejmującej najbardziej ogólne własności świata.

CloudInSky

```
Mt: UniversalVocabularyMt
  isa: SpatiallyDisjointObjectType
    ExistingObjectType
Mt: TopicMt
  isa: WeatherObjects-Weather-Topic
  genls: Outdoors-ExposedToWeather
    CloudlikeObject Opaque Airborne
    TopAndBottomSidedObject
```

Zastosowania

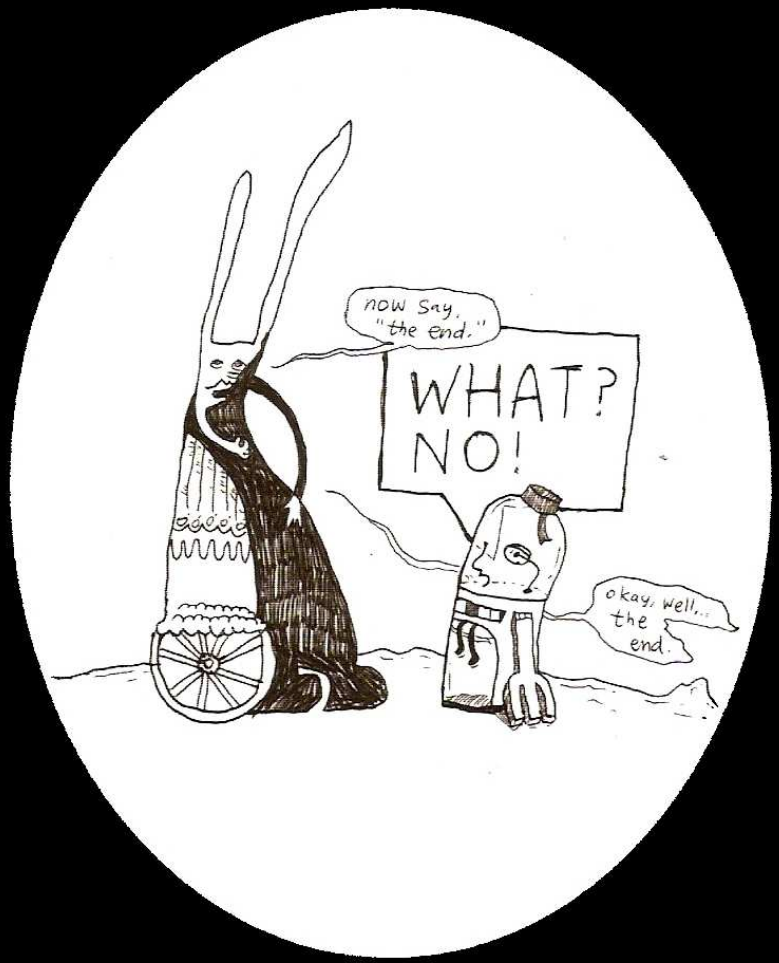
- hakia.com – „semantic search”
- „Prawdziwe” Web 2.0 (języki RDF, OWL)
- Moduł dla Lucene bazujący na WordNecie – uwzględnienie synonimów
- Cycorp – wykorzystuje Cyc m.in. integrowania heterogenicznych baz danych, „inteligentnego” wyszukiwania informacji, rozproszonego AI, analizowania bezpieczeństwa sieci komputerowych
- ABB – ścisła kontrola procesu wytwarzania transformatorów

Zasoby

- WordNet – `wordnet.princeton.edu`
- FrameNet – `framenet.icsi.berkeley.edu`
- Słownik fleksyjny języka polskiego – `winnie.ics.agh.edu.pl/proj_uk/sfjp/index.html`
- Słownik semantyczny języka polskiego (demo) – `wierzba.wzks.uj.edu.pl/~dernow/smddemo`
- OpenCyc – `www.opencyc.org`
- SUMO – `www.ontologyportal.org`
- Strona Włodzisława Ducha – `www.is.umk.pl/~duch/IR.html`

Źródła

- Obrazek „Cyc” – „Everything You Always Wanted To Know About Sex” ;-)
- Obrazek „Sumo” – http://fallingsky.blogs.com
- Obrazek „Nobody knows shoes” – http://hackety.org/press/nks-17.html


now say, "the end."
WHAT? NO!
okay, well... the end.