

Problem rozstrzygania wieloznaczności

Aleksander Pohl

Katedra Informatyki
Akademia Górniczo-Hutnicza

27. listopada 2007

Plan prezentacji

- ▶ Charakterystyka problemu
- ▶ Występowanie problemu
- ▶ Metody rozwiązywania problemu
- ▶ Niezbędne narzędzia lingwistyczne
- ▶ Konferencje
- ▶ Propozycja

Charakterystyka problemu

- ▶ Problem z dziedziny przetwarzania języka naturalnego
- ▶ Pierwsze próby podejmowane już w latach 50
- ▶ Brak algorytmów precyzyjnie rozwiązujących problem
- ▶ Konieczność dostosowania rozwiązań do specyfiki danego języka
- ▶ Dla ludzi problem ten jest niedostrzegalny, dla komputerów – bardzo trudny

Na czym polega problem rozstrzygania wieloznaczności?

Problem rozstrzygania wieloznaczności polega na **wyborze właściwego znaczenia** dla każdego ze słów występujących w zdaniu.

W języku występują słowa, które choć mają identyczną formę, posiadają odmienne znaczenie. Przykład: słowo *akcja* może oznaczać m.in. papier wartościowy oraz fabułę książki.

Dla słów tego rodzaju konieczne jest rozstrzygnięcie, w którym znaczeniu zostały one użyte. Przykład: *Kupiłem dziś akcje warte 100 tysięcy. Akcja tej książki rozwija się bardzo wolno.*

Podstawowe pojęcia z zakresu lingwistyki

- ▶ **Homonimiczność** – zjawisko polegające na występowaniu słów o identycznej formie, lecz odmiennym znaczeniu, np. *akcja*.
- ▶ Relacja **paradygmaticzna** – występuje pomiędzy słowami, które mogą nawzajem się zastępować w określonym kontekście nie łamiąc ograniczeń semantycznych, np. *kot/zwierzę*.
- ▶ Relacja **syntagmatyczna** – występuje pomiędzy słowami, które pojawiają się razem w określonych kontekstach, np. *pić/napój*.

Występowanie problemu

- ▶ **Tłumaczenie maszynowe** – słowa, które w jednym języku są homonimiczne, zazwyczaj w innych językach nie są. Przykład: *jechać* – *ride, drive*.
- ▶ **Wyszukiwanie informacji** – słowa wprowadzane do wyszukiwarki często są homonimiczne. Przykład: *akcja*.
- ▶ **Ekstrakcja informacji** – rozpoznanie ról semantycznych często wymaga rozstrzygnięcia wieloznaczności. Przykład: *Sprawy idą w dobrym kierunku./Dzieci idą do szkoły.*

Metody rozwiązywania problemu I

Z wykorzystaniem **korpusu tekstów** – dla każdego wieloznacznego słowa dodawana jest informacja o tym, w jakim znaczeniu ono występuje. Wykorzystując metody automatycznego uczenia rejestruje się słowa należące do kontekstu danego słowa, które najczęściej występują z danym znaczeniem.

Wady:

- ▶ **czasochłonność** – wymaga ręcznego oznaczenia wszystkich wieloznaczných słów w wielu tekstach
- ▶ **ograniczoność** – brak idealnego korpusu

Metody rozwiązywania problemu II

Z wykorzystaniem **baz wiedzy** – w słownikach semantycznych lub ontologiach opisuje się znaczenia poszczególnych słów wykorzystując relacje semantyczne (paradygmatyczne i syntagmatyczne). Na podstawie tego opisu dąży się do rozstrzygnięcia, w którym znaczeniu występuje dane słowo.

Wady:

- ▶ **czasochłonność** – wymaga skonstruowania bazy wiedzy
- ▶ **problematiczność** – trudno rozstrzygnąć gdzie kończy się wiedza leksykalna, a zaczyna wiedza ogólna

Narzędzia lingwistyczne

- ▶ **Słownik fleksyjny** – w języku polskim występuje zjawisko *fleksji*. Jedno słowo może występować w wielu formach, np.: *kot, kota, kotu...* Słownik fleksyjny pozwala odpowiedzieć na pytanie: jaka jest podstawowa forma danego słowa, w jakim przypadku/liczbie/osobie występuje, etc.
- ▶ **Słownik semantyczny** – niezależnie od użytej metody rozwiązywania problemu, trzeba wyodrębnić poszczególne znaczenia słów wieloznacznych. Definicje w języku naturalnym są mało zrozumiałe dla komputera, dlatego opis ten dokonuje się w terminach relacji semantycznych.

Konferencje

Od 9 lat organizowana jest przez *Association for Computational Linguistics* konferencja **Senseval/Semeval**, która jest w całości poświęcona problemowi rozstrzygania wieloznaczności. Jej głównym celem jest wypracowanie metod pozwalających na miarodajną ocenę systemów komputerowych, pod kątem ich zdolności do rozwiązywania omawianego problemu.

Pomimo tego, że odbyły się już 4 takie konferencje, w których brały udział systemy z całego świata, **w żadnej jej edycji** nie występował system, który starałby się rozwiązywać ten problem **dla języka polskiego**. Wynika to m.in. z faktu, że nie istnieje słownik semantyczny języka polskiego, który zawierałby definicje słów, zgodne z wcześniej przedstawionym wzorcem.

Propozycja

- ▶ Zastosowanie podejścia opartego o **bazy wiedzy**
- ▶ Wykorzystanie biblioteki **CLP** jako podstawowego słownika fleksyjnego
- ▶ Wykorzystanie **Polskiego Słownika Semantycznego** jako podstawowej leksykalnej bazy wiedzy (relacje paradygmatyczne i syntagmatyczne).
- ▶ Ewentualne wykorzystanie ontologii **Cyc**, jako dodatkowego źródła wiedzy.