1 Przetwarzanie języka naturalnego

Uporzadkować ten rozdział!

Jaka jest motywacja do badań nad komputerową analizą języka naturalnego? Nie tylko możliwości dialogu z komputerem w języku naturalnym, ale również analiza już istniejących tekstów, wydobywanie z nich informacji, która nie jest tylko prostym cytatem.

Analiza języka daje się podzielić na 3 części. Jest to:
Syntaktyka języka: zajmująca się formatem, regułami budowy i formalną składnią. Gramatyka jest zbiorem reguł syntaktycznych.
Semantyka języka zajmuje się znaczeniem wyrażeń i konstrukcji w danym języku.
Rozbiór gramatyczny (parsing) to pojęcie stosowane zarówno w językach naturalnych jak i sztucznych.

Zainteresowanie analizą języka naturalnego wynika z wielu pobudek: wymienię je tu od najprostszych do najbardziej ambitnych.

1. Wspomaganie tworzenia i rozwijania słowników, tezaurusów i innych dużych zbiorów językowych, wspomaganie badań lingwistycznych (analiza dużych zbiorów tekstów, weryfikacja hipotez lingwistycznych).
2. Wspomaganie redagowania tekstów wymaga w najprostszym przypadku korekcji pisowni; w językach fleksyjnych nie wystarczy prosty słownik, konieczne jest rozpoznanie rdzenia wyrazu i analiza morfologiczna wyrazu, której rezultatem ma być określenie, czy dana słowoforma jest dopuszczalna. Zagadnienie to zostało rozwiązane dla języka polskiego dopiero w 1995 roku. Jeszcze bardziej ambitnym zadaniem jest analiza składni - programy do analizy składni dostępne są w dobrych edytorach tekstu dla języka angielskiego, niemieckiego czy francuskiego, ale nie dla języka polskiego. Programy te dalekie są od doskonałości.
3. Wspomaganie nauki języka obcego przez analizę błędów popełnianych przez użytkownika programu i automatyczne tworzenie ćwiczeń.
4. Analiza istniejących tekstów - istnieje dużo książek i innych materiałów pisanych, które komputer rozumiejący język naturalny mógłby analizować bez specjalnego przygotowania danych, automatycznie tworzyć streszczenia czy indeksy dokumentów.
5. Tworzenie systemów dialogu w języku naturalnym. Komputer, z którym można porozmawiać, używać będzie mógł każdy, nawet analfabeta (większy rynek); przy okazji lepiej zrozumiemy na czym polega rozumienie języka, oszczędzimy czas na programowanie.
6. Generacja tekstów w języku naturalnym, np. automatyczne komentowanie analiz statystycznych czy finansowych, tworzenie opisanych raportów z baz danych, tworzenie opisów na podstawie zdjęć.
7. Komputerowe wspomaganie tłumaczenia, lub całkowicie automatyczne tłumaczenie tekstów marzy się ludziom od dawna.
8. Analiza i synteza mowy.

Analiza języka naturalnego wiąże się zarówno z reprezentacją wiedzy jak i próbami rozumienia języka mówionego. W Europie kraje Unii Europejskiej podjęły kilka inicjatyw zmierzających do stworzenia inżynierii języka jako odrębnej dyscypliny. W Polsce pierwsze spotkanie dotyczące tej tematyki odbyło się w kwietniu 1995 roku. Na razie udało się stworzyć poprawnie działające analizatory morfologiczne (wyróżnianie rdzeni i określanie możliwych końcówek wyrazów) dla języka polskiego, trwają prace nad komputerowym modelem gramatyki (analizatorem syntaktycznym). Przez pewien okres prowadzono również prace nad systemem dialogu w języku polskim. W kilku miejscach opracowano syntezery mowy uwzględniając specyfikę polskiej wymowy. Pozostałe zagadnienia są w naszym kraju niestety w powijakach.

1.1 Wprowadzenie

Dziedzina, zwana *lingwistyką komputerową*, zajmująca się zastosowaniem komputerów do studiowania języków naturalnych, pojawiła się wkrótce po zbudowaniu pierwszych komputerów przy końcu lat 40-tych. na początku było to proste porządkowanie danych i robienie konkordancji. Jednym z pierwszych wielkich projektów popieranych przez IBM była praca nad konkordancją dzieł Św. Tomasza - praca trwała ponad 20 lat.

Już w 1949 roku zaproponowano (W. Weaver), by zastosować komputery do problemów tłumaczenia z różnych języków. Tłumaczenie maszynowe wyobrażano sobie jako przekładanie słowa za słowem i porządkowanie powstałych zdań zgodnie z regułami gramatycznymi docelowego języka. Pojawiły się jednak nieprzewidziane problemy, zarówno w doborze słów jak i w ich porządkowaniu zgodnie z regułami gramatycznymi. Celem AI stało się rozumienie języka, które pozwoliłoby na odpowiedzi na pytania i tłumaczenia maszynowe. Pierwsze programy do analizy języka pojawiły się w latach 60-tych. Spojrzenie na język naturalny znacznie się zmieniło: używanie mowy to skomplikowana działalność poznawcza, zakładająca wiedzę o znaczeniu słów, strukturze zdań, regułach konwersacji, pewien model słuchacza i dużo ogólnej wiedzy o świecie.

W tej dziedzinie powstawały bardzo różnorodne projekty, trudno jest więc je klasyfikować. Winograd (1972) wprowadził klasyfikację opartą na sposobie reprezentacji wiedzy. Można w ten sposób wyróżnić 4 grupy programów.

Najstarsze programy próbowały osiągnąć ograniczone rezultaty w wąskim zakresie używania języka. Programy takie jak BASEBALL (Green), SAD-SAM (Lindsay), STUDENT (Bobrow), ELIZA (Weizenbaum), używały przypadkowo dobranych struktur danych by przechowywać fakty o określonym, wąskim obszarze. Proste zdania wejściowe przeszukiwane były w celu identyfikacji słów kluczowych, reprezentujących znane obiekty czy związki. Odpowiedzi otrzymywano z gotowych zdań zawartych w bazie danych, dobieranych zależnie od słów kluczowych, w oparciu o heurystyczne reguły. Rezultaty działania niektórych z tych programów były imponujące (i w znacznej mierze mylące).

Druga grupa to programy przechowujące tekst rozmów w bazie danych, poindeksowany na różne sprytne sposoby, by dotrzeć do materiału zawierającego słowa lub frazy kluczowe. Przykładami jest tu SEMANTIC MEMORY (Quillian 1968), PROTOSYNTEX (Simmons, Burger, Long 1966). Dzięki bazie danych systemy te nie były przywiązane swoją budową do danej dziedziny, przez zmianę bazy danych można było zmienić dziedzinę. W dalszym ciągu nie można jednak tu mówić o rozumieniu języka, odpowiedzi musiały być zawarte w bazie danych, brak było możliwości wnioskowania.

Trzecia grupa programów oparta była na logice. W programach SIR (Raphael 1968), TLC (Quillian 969), DEACON (1966) i CONVERSE (1968) informacja zapisana była w bazie wiedzy przy użyciu reprezentacji logicznej wiedzy. w pierwszym kroku dokonywana była analiza semantyczna - tłumaczenie zdań na wewnętrzną reprezentację logiczną. W ograniczonym zakresie pozwoliło to na odpowiedzi wymagające prostego wnioskowania, np. stwierdzenie: „Baja to dalmatyńczyk” wraz z informacją w bazie wiedzy: „dalmatynczyk = podzbiór psów”, pozwala na odpowiedź: „Czy Baja to pies?”. Systemy te nie pozwalały jednak na wyciąganie bardziej skomplikowanych wniosków logicznych ze względu na nieefektywność logicznej reprezentacji wiedzy.

Czwarta grupa programów oparta jest na nowszych bazach wiedzy. Równolegle z rozwojem schematów reprezentacji wiedzy nastąpił duży postęp w lingwistyce. Teoria gramatyk generatywnych Chomsky'ego (1957) wywarła duży wpływ na lingwistykę komputerową. Wprowadzono różne rodzaje gramatyk: transformacyjne, frazeologiczne, przypadków, systemiczne i semantyczne.

Rozbiór gramatyczny ma za zadanie „delinearyzcję” zdania, określenie funkcji słów (przy pomocy reguł gramatycznych i innej wiedzy) i stworzenia modelu (np. drzewa) powiązań, np. ten przymiotnik modyfikuje znaczenie tego rzeczownika, który jest podmiotem zdania ... Systemy analizy języka zawsze zawierały jakiś rodzaj rozbioru gramatycznego, choćby bardzo prymitywny. Parser czyli program do rozbioru gramatycznego składa się z części definiującej gramatykę i z metod użycia jej reguł.

Na początku lat 70-tych pojawiły się systemy oparte na bazach wiedzy, które potrafiły sobie radzić z syntaktyką i semantyką zdań w określonej, wąskiej dziedzinie. System LUNAR (W. Woods) potrafił odpowiadać na pytania dotyczące próbek skał przywiezionych z księżyca, w oparciu o dużą bazę danych NASA. LUNAR używał reprezentacji proceduralnej wiedzy (u lingwistów komputerowych nazywa się to semantyką proceduralną), pytania zamieniane były na programy i wykonywane w celu wywnioskowania z bazy danych odpowiedzi. Najbardziej znanym systemem tego rodzaju był SHRDLU (T. Winograd), prowadzący dialog symulowanego robota, który ma za zadanie przestawiać obiekty na symulowanym stole i odpowiadać na pytania z tym związane. Reprezentacja proceduralna odniosła duży sukces tam, gdzie wiedza zawarta w pasywnych deklaracjach struktur danych, interpretowanych przez inne procedury, nie dawała się łatwo wykorzystać.

Popularną deklaratywną reprezentacją wiedzy stała się reprezentacja poprzez sieci semantyczne. Pewne wnioskowania w takich sieciach są łatwe, miano nadzieję, że te, które przychodzą ludziom w sposób naturalny, da się łatwo w nich zawrzeć. Zrobiono znaczny postęp w formalizacji sieci semantycznych. Węzły tych sieci często