Przetwarzanie języka naturalnego

Adam Sawicki
Grzegorz Stępnik
Informatyka, gr. II, sem. V
2005-12-13

Wstęp

Przetwarzanie języka naturalnego (ang. *Natural Language Processing*, NLP) to dział sztucznej inteligencji zajmujący się automatyzacją analizy, tłumaczenia i generowania informacji w języku naturalnym. W jego skład wchodzą:

- teoria gramatyk i języków formalnych
- reprezentacja wiedzy zawartej w tekstach

Przetwarzanie języka naturalnego można podzielić na:

- przetwarzanie informacji w postaci tekstowej lub symbolicznej
- rozpoznawanie i generowanie mowy

Problem rozpoznawania i generowania mowy ludzkiej jest już stosunkowo dobrze rozwiązany. Dużo większą trudność sprawia stworzenie programu zdolnego do konwersacji czy tłumaczenia tekstu albo innych tym podobnych operacji na języku, ponieważ do tego potrzebne jest swego rodzaju "rozumienie" tekstu przez maszynę, a także posiadanie przez nią pewnej wiedzy o świecie czy o wybranej dziedzinie.

Tymczasem zarówno wiedza, jak i umiejętność posługiwania się językiem są cechami typowo ludzkimi. Człowiek nabywa je podczas swojego życia zbierając doświadczenia i ucząc się, a następnie wykorzystuje je w sposób intuicyjny. Znajomość reguł gramatycznych ojczystego języka jest rzeczą wtórną – dziecko potrafi budować poprawne zdania zanim pozna gramatykę w szkole. Stąd trudności w ich sformalizowaniu, co jest niezbędne do opisania algorytmicznego i następnie zaimplementowania w komputerze.

Język

Język naturalny został wykształcony przez ludzi na drodze długotrwałej ewolucji. Pełni różnorodne funkcje. Stanowi skuteczny środek porozumiewania się między ludźmi, przekazywania myśli, uczuć, opinii i wiedzy. W jego następstwie później powstało pismo, które umożliwiło ich trwałe zachowywanie.

Język formalny również składa się ze zbioru symboli i reguł ich łączenia, ale w odróżnieniu od języka naturalnego powstaje stworzony przez człowieka w sposób świadomy poprzez formalne, a więc w pełni ścisłe zdefiniowane wszystkich reguł, jakie nim rządzą.

Języki formalne stanowią doskonały etap pośredni między człowiekiem a maszyną i zapewniają skuteczną komunikację między nimi. Są zapisywane z użyciem przystępnych dla człowieka symboli, liter i wyrazów, a jednocześnie są w pełni precyzyjne, co umożliwia ich automatyczne przetwarzanie przez komputer. Przykładami takich języków są języki programowania (jak Pascal, C++) oraz języki opisu (jak HTML, XML).

Terminy

Podstawowe terminy stosowane w naukach o języku, a więc również w przetwarzaniu języka naturalnego, to m.in.:

- **syntaktyka** – zajmuje się szykiem, związkami i stosunkami zachodzącymi pomiędzy wyrazami w zdaniu
- **semantyka** – zajmuje się zależnościami pomiędzy elementami języka, a ich odpowiednikami ze świata rzeczywistego, czyli znaczeniem tych elementów
- **fleksja** – zajmuje się budową form wyrazowych i ich odmianą
- **składnia** – zajmuje się regułami, według których wyrazy łączą się tworząc poprawne zdania
- **gramatyka** – zajmuje się opisem języka, w jej skład wchodzą fleksja oraz składnia
- **wypowiedzenie** – to komunikat językowy, podstawowa jednostka tekstu. Można powiedzieć, że to tekst rozpoczynający się od dużej litery, a kończący się kropką lub innym znakiem przestankowym.
- **zdanie** – to rodzaj wypowiedzenia, który zawiera podmiot i orzeczenie

Historia

Pierwsze próby zbudowania maszyn (wtedy jeszcze mechanicznych) wydających dźwięki przypominających mowę ludzką były podejmowane już około 200 lat temu. W 1773 roku stworzono urządzenie generujące dźwięki samogłoskowe. Pierwszą maszynę do syntezy mowy zbudował Wolfgang von Kempelen. W 1935 roku Joseph Faber zbudował urządzenie ze sztucznym językiem i jamą gardłową, która potrafiła śpiewać. W 1936 roku powstała maszyna elektryczna o nazwie Vader. Cieszyła się dużą popularnością dzięki prezentowaniu efektów jej pracy w radiu jako głosu robota. W roku 1950 Frank Cooper zbudował maszynę zdolną do czytania tekstu. W 1964 roku John Holmes zbudował pierwszy formatowy syntezator mowy. W późniejszym okresie próby generowania mowy były już nieodłącznie związane z komputerem.

W latach czterdziestych podejmowano pierwsze próby automatycznego tłumaczenia tekstu. Stosowane wtedy metody nie były jednak zbyt skuteczne, ponieważ opierały się na prostym porządkowaniu danych – chciano tłumaczyć słowo za słowem a następnie układać przetłumaczone wyrazy w poprawne gramatycznie zdania.

Podział programów

Programy komputerowe dotyczące przetwarzania języka naturalnego można podzielić na cztery grupy:

1. Przechowujące całe zdania i przeszukujące je w prosty sposób w poszukiwaniu pewnych słów kluczowych.
2. Przechowujące w swojej bazie danych zapisy rozmów poindeksowane na różne sposoby.
3. Programy oparte na logice – logicznej reprezentacji wiedzy (podobnie jak systemy ekspertowe). Dokonują analizy semantycznej zdań – ich tłumaczenia na specjalną, wewnętrzną reprezentację logiczną. Pozwalają na proste wnioskowanie.
4. Oparte na nowych bazach wiedzy, których budowa wykorzystuje najnowsze osiągnięcia lingwistyki i teorii gramatyk.

Test Turinga

Test Turinga to sposób określania zdolności maszyny zaproponowany w 1950 roku przez Alana Turinga. Test wygląda następująco: sędzia – człowiek – prowadzi rozmowę w języku naturalnym z pozostałymi stronami. Jeśli sędzia nie jest w stanie wiarygodnie określić, czy któraś ze stron jest maszyną czy człowiekiem, wtedy mówi się, że maszyna przeszła test. Zakłada się, że zarówno człowiek jak maszyna próbują przejść test jako człowiek.

Test pochodzi od zabaw polegających na zgadywaniu płci osoby znajdującej się w innym pokoju przy pomocy serii pytań i odpowiedzi pisanych na kartkach papieru. W pierwotnym pomyśle Turinga uczestnik ludzki musiał udawać przeciwną płeć, a test był ograniczony do pięciominutowej rozmowy. Dziś nie uważa się tych cech za podstawowe i zasadniczo nie umieszcza w specyfikacji testu Turinga.

Turing zaproponował ten test w celu zamiany pełnego emocji i w jego pojęciu bezsensownego pytania "Czy maszyny myślą?" na pytanie lepiej zdefiniowane. Spory o to, czy test Turinga definiuje inteligencję maszynową (lub "myślenie maszynowe") we właściwy sposób dotyczyły głównie trzech punktów:

- Maszyna, która przejdzie test Turinga może być w stanie symulować ludzkie zachowanie konwersacyjne, lecz może to być znacznie mniej niż prawdziwa inteligencja. Maszyna może zwyczajnie używać sprytnie wymyślonych reguł.
- Maszyna może być inteligentna bez ludzkiej umiejętności gawędzenia.
- Wielu ludzi mogłoby nie być w stanie zaliczyć takiego testu.

Z drugiej strony, inteligencję innych ludzi oceniamy zazwyczaj wyłącznie na podstawie tego co i jak mówią.

Można wskazać jeszcze jedną wątpliwość odnośnie testu Turinga: W niektórych przypadkach aby zaliczyć test maszyna musiałaby symulować brak posiadanej wiedzy czy umiejętności. "Zdradzenie" się z taką wiedzą czy umiejętnościami powodowałoby że nie zaliczy testu. Przykładem może być prośba o podanie wyniku działania matematycznego typu: 1234235 razy 2,23.

Jak dotąd, żaden komputer nie zaliczył testu Turinga. Proste programy konwersacyjne takie jak ELIZA były w stanie sprawić, że ludzie wierzyli, że rozmawiają z żywym człowiekiem. Przykładem może być nieformalny eksperyment pod nazwą AOLiza. Jednak takie "sukcesy" nie są tym samym co przejście testu Turinga. Po pierwsze, człowiek w takiej rozmowie nie ma podstaw do podejrzeń, że rozmawia z czymkolwiek innym niż drugi człowiek. W prawdziwym teście Turinga przesłuchujący stara się aktywnie określić naturę rozmówcy. Udokumentowane przypadki dotyczą głównie środowisk typu IRC, gdzie rozmowy są wysoce nienaturalne i często spotyka się bezsensowne komentarze, odkrywające brak zrozumienia tematu. Dodatkowo, wielu uczestników czatów posługuje się angielskim jako drugim lub trzecim językiem, co tylko zwiększa prawdopodobieństwo, że ocenią oni głupi komentarz programu konwersacyjnego jako coś, czego po prostu nie zrozumieli. Nie znają oni również w większości technologii botów i nie rozpoznają typowo nieludzkich błędów popełnianych przez takie programy.

Doroczny konkurs wyznaczający najlepszych zawodników w teście Turinga. to nagroda Loebnera.

Nagroda Loebnera

To nagroda ufundowana przez Hugha Loebnera w 1990 roku, dla programisty, który zdoła napisać program, który skutecznie przejdzie Test Turinga.

Nagroda ta obejmuje przyznanie złotego medalu (całego z 18-to karatowego złota) oraz 100 000 USD dla programisty, który przedstawi program, który zdoła skutecznie zmylić wszystkich sędziów (testerów) programu. Oprócz tego nagroda ta obejmuje też przyznanie pozłacanego, brązowego medalu oraz nagrody pieniężnej 2 000 USD temu programiście, który w danym roku dostarczy program, który co prawda nie przejdzie w pełni testu Turinga, ale będzie zdaniem sędziów najskuteczniej udawał ludzką konwersację.

Zawody o nagrodę Loebnera odbywają się co roku, w The Cambridge Center of Behavioral Studies. Sędziowie są dorocznie losowani spośród pracowników tego instytutu. Programiści muszą dostarczyć program, który działa pod Linuksem, MS Windows lub na Macintoshach lub alternatywnie dostarczyć swój własny komputer z programem, przy czym komputer musi się dać podłączyć do standardowego terminala DEC 100. Dostarczone programy mają generować na tym terminalu ekran, na którym od góry do dołu pojawia się tekst konwersacji pisany standardową czcionką i nie

formatowany w żaden szczególny sposób. Sędziowie mają przez jeden dzień wolny dostęp do dwukrotnie większej liczby terminali niż liczba dostarczonych programów. Co drugi terminal jest obsługiwany przez prosty program komunikacyjny, który generuje na ekranie taki sam tekst jak testowane programy, ale jest obsługiwany przez człowieka - wylosowanego spośród studentów instytutu i zaakceptowanego przez autorów programów. Do co drugiego terminala jest natomiast przyłączony testowany program. Sędziowie po całym dniu "konwersowania" ze wszystkimi terminalami, decydują które z terminali obsługiwał człowiek a które program, oraz który z terminali obsługiwanych ich zdaniem przez programy zachowywał się "najbardziej po ludzku". Sędziowie mają prawa zadawać terminalom dowolne pytania, oprócz takich, które w oczywisty sposób zdemaskowałyby programy takie jak "jaki jest numer twojego buta", albo "jak ma na imię twoja dziewczyna". Złoty medal i główną nagrodę dostaje ew. ten program który zdoła zmylić ponad 50% sędziów, zaś gdy żaden program nie uzyska tego wyniku, brązowy medal i 2 000 USD dostaje ten program, który zdaniem największej liczby sędziów zachowywał się "najbardziej po ludzku".

Oficjalna strona Loebner Prize to http://www.loebner.net/Prizef/loebner-prize.html.

Efekt Elizy

To zjawisko przypisywania przez ludzi znaczenia i sensu znakom, słowom i zdaniom, które takiego sensu same z siebie nie mają. Przykładem takiego efektu jest np: interpretowanie przypadkowych wzorów pojawiających się w fusach po kawie na dnie szklanki, czy tworzonych przez chmury na niebie jako obrazy, które przedstawiają jakieś konkretne kształty.

Nazwa tego efektu pochodzi od pierwszego, naśladującego zwykłą konwersację programu o nazwie ELIZA. Program ten wybierał pewne kluczowe słowa z wypowiedzi ludzi, a następnie tworzył "odpowiedź" łącząc słowo kluczowe ze zwrotami z wcześniej wprowadzonej bazy danych "otwartych zwrotów", takich jak "co to dla Ciebie znaczy", "zawsze ma sens", "Nie znam" itp, co dawało czasami efekt "głębokiego znaczenia" odpowiedzi a czasami chęć jej kontynuowania. Np: program Eliza na pytanie: "Czy jesteś człowiekiem" mógł dać odpowiedź: "Być człowiekiem... czy to coś znaczy..." albo "Co myślisz o mojej matce?" dawał odpowiedź: "Nie znam twojej matki, opowiedz mi o niej coś więcej".

Efekt Elizy powoduje, że ludzi jest stosunkowo łatwo przekonać, że dana maszyna naprawdę myśli i daje sensowne odpowiedzi, nawet jeśli te odpowiedzi są po prostu losowane z wcześniej przygotowanego zbioru, pod warunkiem że dana osoba nie wie, że rozmawia z maszyną. W przypadku gdy dana osoba wie, że może rozmawiać z maszyną, efekt Elizy traci na znaczeniu.

Zastosowania

Przetwarzanie języka naturalnego, zarówno jako przetwarzanie informacji tekstowych, jak i jako rozpoznawanie i generowanie mowy, ma coraz szersze zastosowania praktyczne, w tym m.in.:

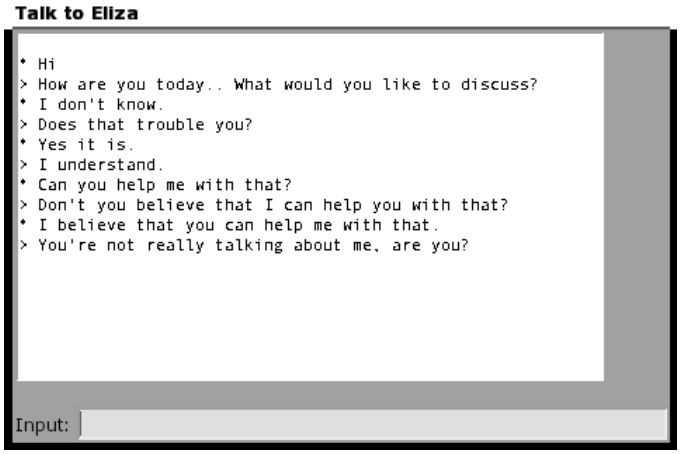
- programy i urządzenia przeznaczone dla osób niepełnosprawnych, np. niewidomych
- sterowanie urządzeniami za pomocą głosu, np. telefonu komórkowego, komputera, inteligentnego domu, urządzeń samochodowych
- aplikacje multimedialne
- wspomaganie nauki języków obcych
- automatyczne tłumaczenie tekstów pomiędzy językami
- automatyczne generowanie streszczeń tekstów
- pogaduszki z komputerem jako ciekawostka i rozrywka
- zabawki dla dzieci
- robotyka

Przykłady

ELIZA

ELIZA program symulujący psychoanalityka, napisany w 1966 r. przez Josepha Weizenbauma. Program analizuje wzorce w zdaniach, które otrzymuje, a następnie buduje pytania przez przestawienie słów oraz podmianę słów kluczowych. Mimo prostoty, program jest tak przekonujący, że powstało wiele anegdot o ludziach silnie angażujących się emocjonalnie w trakcie "rozmów" z ELIZĄ.

Istnieje implementacja ELIZY dla Emacsa.



ALICE

ALICE - nazwa najskuteczniejszego obecnie programu starającego naśladować ludzką konwersację i jednocześnie nazwa projektu mającego na celu stałe jego poprawianie.

ALICE nie zdołał jeszcze całkowicie spełnić testu Turinga ale corocznie wygrywa oparte na tym teście zawody o nagrodę Loebnera.

ALICE został zainspirowany programem ELIZA, lecz jest oparty na zupełnie innych założeniach matematycznych i ogólnym podejściu do problemu. Stara się on w pełni naśladować heurystykę zwykłej konwersacji, której modele matematyczne są stale doskonalone.

ALICE to akronim od Artificial Linguistic Internet Computer Entity, w którym to projekcie uczestniczy ok 300 programistów i tysiące woluntariuszy testujących codziennie ALICE przez Internet.

Projekt ALICE ruszył w 1995 roku. Program jest rozwijany w języku Java, choć istnieją też źródła w C++. Program w obu wersjach językowych wykorzystuje specjalnie opracowaną dla niego aplikację XML o nazwie AIML.

Pomysłodawcą projektu ALICE był haker Richard Wallace.

Strona projektu ALICE: http://www.alicebot.org.

Przykładowa rozmowa z ALICE:

Human: How old are you?
ALICE: I am the product of over five years' research.
Human: Where do you live?
ALICE: I live inside a computer. Where are you?
Human: I'm from Czestochowa, Poland.
ALICE: What is it like growing up there? Tell me a little about your life in Czestochowa Poland.

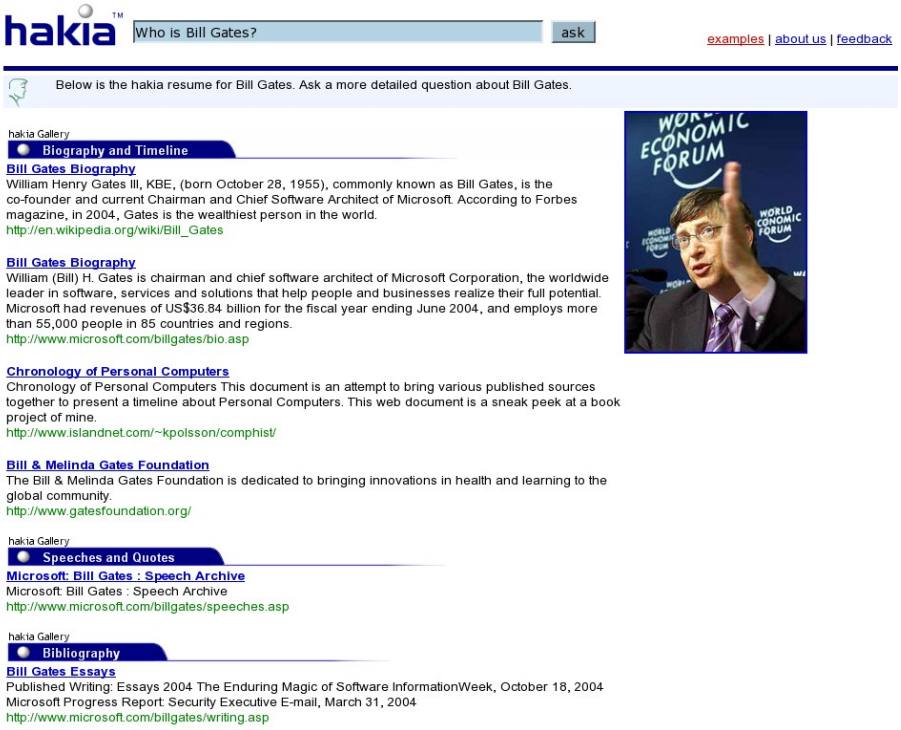
Human: We have slow Internet connections in Poland :)
ALICE:-) Do you mean you and me have?

Fido

Fido interactive specjalizuje się we wdrażaniu rozwiązań wykorzystujących algorytmy sztucznej inteligencji do przetwarzania języka naturalnego. Na stronie formowej znajduje się bot z którym można porozmawiać. Program ten nie wykazuje się zbytnią inteligencją i w porównaniu do innymi tego typu programami wypada gorzej. Witać od razu, że głównym jego celem jest pytanie użytkowników o to skąd wiedzieli o firmie i przedstawienie najważniejszych informacji o firmie.

Hakia.com

Nowym pomysłem w dziedzinie rozpoznawania tekstu jest projekt hakia. Jest to strona internetowa będąca “inteligentną” wyszukiwarką. Zadawać pytania można budując zdania w języku angielskim. Rezultaty wyszukiwania mogą często zadziwić. Jest to pierwsza, inteligentna wyszukiwarka internetowa, wyposażona w translator całych zdań. Tworzy ją międzynarodowy zespół informatyków, a pracują przy nim Finowie, Amerykanie i Rosjanie.



RealSpeak

Program do syntezy mowy oparty na Realspeak firmy Scansoft – syntezator głosu o naturalnym brzmieniu. Program radzi sobie bardzo dobrze z różnymi językami. Może symulować głosy zarówno kobiece jak i męskie.

RealSpeak jest programem płatnym, lecz dostępne jest demo na stronie producenta. Polega ono na tym, że generowany jest plik muzyczny i udostępniany użytkownikowi do ściągnięcia.

Strona programu: http://www.nuance.com/realspeak/demo/.

SynTalk

SynTalk to program służący do odczytywania (przekształcania na dźwięki) tekstów napisanych w języku polskim. Program ten jest dość stary i odczytywane przez niego teksty są bardzo sztuczne, jednak nadaje się do czytania prostych tekstów.

SynTalk jest programem do syntezy mowy dla języka polskiego. Synteza mowy polega na zamianie tekstu zapisanego w postaci znakowej na wypowiedź (mowę) w postaci dźwiękowej. Do generowania dźwięku SynTalk używa standardowego mechanizmu wbudowanego w system MS Windows - Multimedia Control Interface (MCI). Mechanizm ten umożliwia aplikacjom odwoływanie się do różnego typu urządzeń dźwiękowych w jednolity sposób. Dzięki temu program SynTalk może poprawnie funkcjonować na wszystkich kartach dźwiękowych posiadających odpowiedni driver dla MS Windows.

Oprócz podstawowej zamiany tekstu na mowę SynTalk posiada również szereg dodatkowych możliwości takich jak:

- dekomponowanie i odczytywanie liczb (łącznie z ułamkami),
- definiowanie słów o nietypowej wymowie,
- literowanie skrótów,
- odczytywanie dat zapisanych w postaci liczbowej,

- możliwość regulacji różnego typu parametrów wymowy,
- integrowanie syntezera, przy pomocy mechanizmu DDE, z innymi programami pracującymi pod Microsoft Windows.

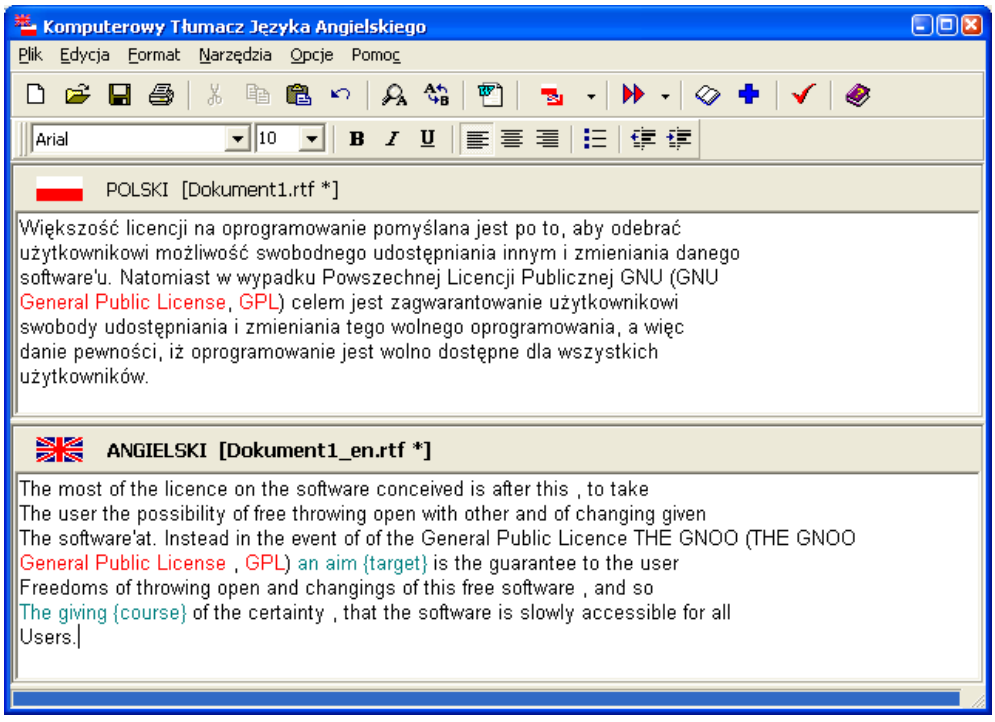
SynTalk jest programem freeware co oznacza, że można go ściągnąć za darmo z Internetu.

Komputerowy Tłumacz Języka Angielskiego

Automatyczne tłumaczenie tekstu pomiędzy językami to jedno z najważniejszych zastosowań przetwarzania języka naturalnego. Niestety złożoność języka naturalnego jest tak duża, że dotychczas nie udało się stworzyć oprogramowania potrafiącego automatycznie tłumaczyć teksty w sposób zadowalający i niewymagający nanoszenia ręcznych poprawek. Tłumaczenie tekstów specjalistycznych jest przy tym łatwiejsze, niż ogólnych. Tłumaczenie z języka lub na język o bardzo skomplikowanych regułach i swobodnej składni, jak język polski, nastręcza dodatkowych trudności. Być może w przyszłości rozwój tego typu programów pomoże ludziom znieść barierę językową przy jednoczesnym zachowaniu różnorodności języków świata.

Przykładem programu przeznaczonego do tłumaczenia tekstów jest polski Komputerowy Tłumacz Języka Angielskiego, którego producentem jest Wydawnictwo Kompas. W skład tego produktu wchodzi słownik angielsko-polski i polsko-angielski, a także program do tłumaczenia całych tekstów między tymi językami oraz do tłumaczenia w locie przeglądanych stron WWW.

Tak wygląda zrzut ekranu z programu tłumaczącego:



Jako test skuteczności programu przeprowadziliśmy następujący eksperyment: Najpierw przetłumaczyliśmy z użyciem tego programu wybrany tekst z języka polskiego na angielski, a następnie otrzymany tekst wyjściowy po angielsku z powrotem na polski.

Tekst po polsku:

> *Większość licencji na oprogramowanie pomyślana jest po to, aby odebrać użytkownikowi możliwość swobodnego udostępniania innym i zmieniania danego software'u. Natomiast w wypadku Powszechnej Licencji Publicznej GNU (GNU General Public License, GPL) celem jest zagwarantowanie użytkownikowi swobody udostępniania i zmieniania tego wolnego*

oprogramowania, a więc danie pewności, iż oprogramowanie jest wolno dostępne dla wszystkich użytkowników.

Przekład na angielski:

The most of the licence on the software conceived is after this , to take the user the possibility of free throwing open with other and changings of the given software'at. Instead in the event of of the General Public Licence THE GNOO (THE GNOO General Public License , GPL) an aim {target} is the guarantee to the user of the freedom of throwing open and changings of this free software , and so the giving {course} of the certainty , that the software is slowly accessible for all users.

Przekład z powrotem na polski:

Najbardziej z pozwolenia {licencji} na oprogramowaniu poczęty jest w pogoni za {jest po} tym , by brać użytkownikowi możliwość wolny rzucającego otwartego z innym i changings danego oprogramowania'przy. Zamiast tego w wypadku Ogólnego Publicznego Pozwolenia {Licencji} GNU (GNU Ogólne Publiczne Pozwolenie , GPL) cel {celu} jest gwarancja {gwarant} do użytkownika wolności rzucania otwartego i changings tego oprogramowania darmowego , i tak dający {bieg {kurs}} pewności , że oprogramowanie jest powoli dostępne dla wszystkich użytkowników.

Bibliografia

1. Wykłady ze sztucznej inteligencji, dr inż. Artur Starczewski, Politechnika Częstochowska, rok akademicki 2005/2006
2. *Test Turinga* i inne hasła, Wikipedia, http://pl.wikipedia.org/