Struktura systemu tłumaczenia automatycznego z języka polskiego na język angielski

The Structure of a Polish-English MT System

Krzysztof Jassem

Uniwersytet im. Adama Mickiewicza, Poznań
Wydział Matematyki i Informatyki
jassem@math.amu.edu.pl

STRESZCZENIE

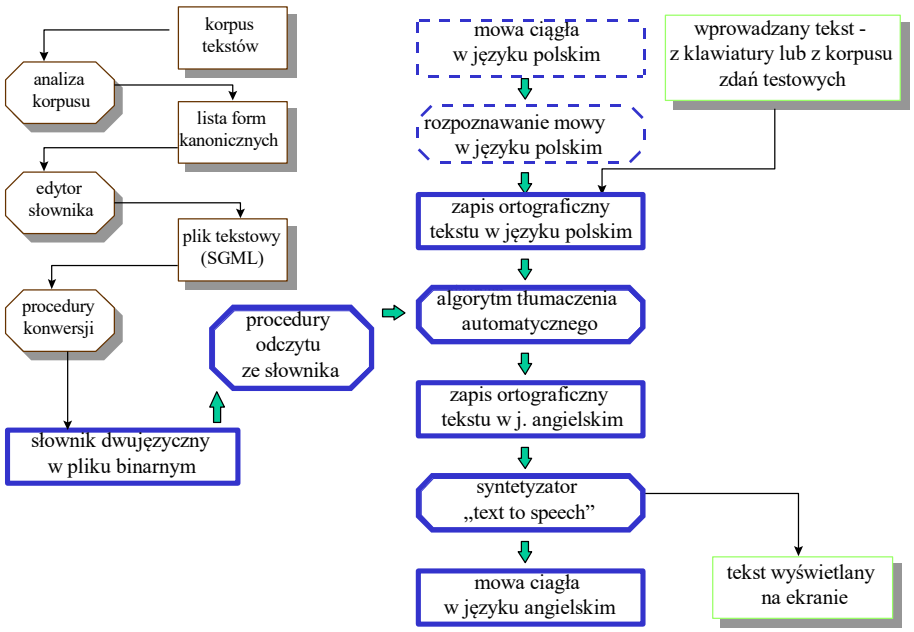
W poniższej pracy przedstawiona jest koncepcja struktury systemu automatycznego tłumaczenia z języka polskiego na język angielski. Koncepcja uwzględnia różnice w stopniu zaawansowania prac poszczególnych modułów systemu. Struktura systemu jest dynamiczna w takim rozumieniu, że umożliwia stworzenie systemu działającego z pewnymi ograniczeniami, a następnie doskonalenie go w miarę postępu prac w poszczególnych modułach.

ABSTRACT

This paper presents the concept of a structure of a Polish-to-English machine translation system. The concept takes into consideration gaps between states of the technological development of individual system modules. The structure of the system is dynamic: it enables the design of a system working under some limitations and an improvement of the system in keeping with the development of research in individual modules.

1. Ogólny schemat struktury systemu automatycznego tłumaczenia

Poniższy rysunek przedstawia strukturę systemu tłumaczenia automatycznego z języka polskiego na język angielski. Strzałkami zaznaczono kierunek przepływu danych. W prostokątach umieszczono dane, w ośmiokątach aplikacje. Grubszymi liniami bez cieni zaznaczono docelową strukturę systemu w momencie, w którym uzna się, że proces doskonalenia systemu został zakończony. Liniami przerywanymi zaznaczono te moduły, dla których nie opracowano jeszcze oprogramowania umożliwiającego połączenie ich z systemem.



2. Krótka charakterystyka poszczególnych modułów systemu

2.1. Korpus tekstów

Zebrano korpus tekstów z dziedziny informatyki, zawierający blisko 200000 wyrazów. Głównym źródłem tekstów były wydruki czasopism komputerowych, znajdujące się na internetowych stronach WWW. Teksty oznaczono w odpowiedni sposób w celu wyeliminowania fragmentów nie związanych z dziedziną informatyki i przygotowania do przetworzenia za pomocą analizatora morfologicznego. Szczegóły dotyczące wstępnego oznaczenia tekstów można znaleźć w pracy [1].

2.2. Analiza morfologiczna korpusu

Analizę morfologiczną korpusu tekstów dokonano w dwóch etapach. W pierwszym etapie korpus poddano działaniu tzw. lematyzatora (analizatora morfologicznego) [6], w wyniku czego otrzymano dane w postaci korpusu “otagowanego”. W korpusie otagowanym każdy wyraz był oznaczony swoim opisem morfologicznym: formą kanoniczną leksemu i cechami gramatycznymi. Lematyzator opierał się na słownictwie zawartym w słowniku [9]. Program nie dokonywał analizy syntaktycznej - dla wyrazów homograficznych różnych leksemów podane zostały formy kanoniczne wszystkich leksemów, do których wyraz mógł należeć - również te, które można by wyeliminować na podstawie “pobieżnej” analizy kontekstu. Zadaniem drugiego etapu analizy - patrz [1] - było ustalenie form kanonicznych tych wyrazów, które w pierwszej fazie zostały odrzucone przez lematyzator jako nie istniejące w słowniku - w przeważającej

części były to techniczne terminy informatyczne i imiona własne - oraz rozstrzygnięcie przy pomocy kompetencji ludzkiej wyboru formy kanonicznej wyrazów homograficznych na podstawie analizy kontekstu.

Uzyskane narzędzia informatyczne są na tyle uniwersalne, że mogą posłużyć do analizy korpusu tekstów z innych dziedzin.

2.3. Lista form kanonicznych

Na podstawie danych wyjściowej drugiego etapu ułożono listę frekwencyjną form kanonicznych wyrazów występujących w opracowanym korpusie. Lista została automatycznie opatrzona znacznikami charakterystycznymi dla formatu dokumentów SGML (patrz [1]). Dzięki temu możliwe będzie zastosowanie listy jako danych wejściowych edytora słownika SGML (patrz punkt 2.4) i utworzenie polsko-angielskiego słownika elektronicznego.

Opracowywany jest algorytm półautomatycznego ekstrahowania form kanonicznych fraz leksykalnych występujących w korpusie.

2.4. Edytor słownika SGML

Edytor jest oprogramowaniem, które ułatwia uzupełnianie wartości pól danych leksykalnych. Znaczenie pól haseł słownika opisane jest w [4]). Ustalenie popularnego formatu dokumentów - SGML - miało na celu uzyskanie uniwersalności słownika, czyli umożliwienie zastosowania go w innych algorytmach tłumaczenia (również typu “human-aided”). Format ten daje ponadto możliwość rozszerzenia (w przyszłości) o inne pola (związane np. z reprezentacją wiedzy, kiedy systemy tłumaczenia dojdą do etapu konsultowania wiedzy “pozaleksykalnej”). Opis edytora podany jest w [8]. Przyjęto, że dane wejściowe i wyjściowe słownika mają ten sam format - dane wyjściowe “z reguły” będą miały wypełnione niektóre z pól, które w danych wejściowych były puste. Takie podejście daje możliwość wielokrotnej modyfikacji opisu każdego hasła, co będzie szczególnie istotne w początkowej fazie uzupełniania opisów haseł w słowniku. Ponadto takie rozwiązanie umożliwia wykorzystanie danych leksykalnych zebranych w dotychczasowej pracy (patrz np. [3]), w której stworzono leksykon około 2000 leksemów (leksykon ten utworzony był heurystycznie - bez systematycznej analizy korpusu tekstów).

2.5. Procedury konwersji i modyfikacji

W celu zoptymalizowania czasu dostępu do słownika, dane leksykalne (w formacie tekstowym) poddawane są konwersji do struktury automatu skończonego i przechowywane w pliku binarnym. Szczegóły na temat implementacji, która dokonuje tej operacji, można znaleźć w [3]. W stosunku do aplikacji tam opisanej należało dokonać poprawek, które uwzględniały format SGML danych leksykalnych oraz możliwość dokonywania różnorodnych modyfikacji słownika przy pomocy edytora opisanego w punkcie 2.4. Przy pomocy edytora dokonywane są zmiany w pliku tekstowym; procedury konwersji muszą zapewniać możliwość dokonania odpowiadających tym zmianom modyfikacji w pliku binarnym. W

pracy [5] autor opisuje, jakie wymagania powinna spełniać struktura danych leksykalnych oraz oprogramowanie obsługi słownika, aby takie modyfikacje były możliwe.

2.6. Słownik w pliku binarnym i procedury odczytu

Przewagą odczytu z pliku binarnego nad odczytem z pliku tekstowego jest wielokrotnie większa szybkość. Ponadto, szybkość odczytu z pliku binarnego nie zależy od liczby jednostek leksykalnych słownika (jest natomiast odwrotnie proporcjonalna do długości odszukiwanego hasła). Pewną niedogodność sprawiają problemy techniczne na styku języków programowania. Odczyt z pliku binarnego słownika trudno jest zrealizować w języku deklaratywnym - takim jak Prolog - w którym zrealizowany jest algorytm tłumaczenia automatycznego. Odczyt zaimplementowany jest w języku proceduralnym *C*. Konsultowanie słownika w procesie tłumaczenia wymaga wywoływania funkcji języka *C* w języku *Prolog,* co stwarza konieczność zbudowania tzw. biblioteki łączonej dynamicznie (DLL). Zagadnienie to poruszane jest również w pracy [5].

2.7. Rozpoznawanie mowy w języku polskim

Stan zawansowania prac w tym zakresie nie umożliwia jeszcze włączenia modułu rozpoznawania mowy do systemu tłumaczenia. Na Politechnice Wrocławskiej dokonywane są badania, które w niedalekiej przyszłości powinny umożliwić włączenie tego modułu - w zakresie około 100 wyrazów - do systemu. Obecnie dane (zdania i frazy języka polskiego) wprowadzane są do systemu w zapisie ortograficznym z klawiatury lub pliku tekstowego, zawierającego kilkaset przykładowych zdań i fraz testowych.

2.8. Algorytm tłumaczenia automatycznego

Szczegóły dotyczące algorytmu można znaleźć w [2]. W stosunku do opisywanej tam wersji systemu, modyfikacji uległy niektóre szczegółowe rozwiązania. Zmieniono środowisko pracy algorytmu. Algorytm, zrealizowany wcześniej w języku Arity Prolog w środowisku DOS, został przepisany w języku SWI-Prolog pracującym w środowisku Windows. W ten sposób zapewniono zgodność środowiska pracy wszystkich obecnie zaimplementowanych modułów systemu. Dokonano teoretycznego opracowania trudnego zagadnienia kwantyfikacji i określoności polskich grup rzeczownikowych pod kątem późniejszego zaimplementowania w systemie tłumaczenia [7]. W najbliższej przyszłości planuje się takie poszerzenie algorytmu, aby efektywnie wykorzystana była informacja semantyczna i kontekstowa zawarta w słowniku opartym na zanalizowanym korpusie tekstów.

2.9. Dane wyjściowe systemu tłumaczenia automatycznego

Pierwotnie danymi wyjściowymi systemu były zdania i frazy języka angielskiego (wyświetlane na ekranie) oraz ich drzewa struktury frazowej. W ostatnim czasie podjęto próby dołączenia do systemu syntetyzatora mowy języka angielskiego w celu zapewnienia wyjścia fonicznego. Dokonano przeglądu

syntetyzatorów reklamowanych w sieci Internet. W pracy [10] scharakteryzowano kilkanaście takich syntetyzatorów. Dokonano porównania jakości syntezy oraz przeanalizowano syntetyzatory pod kątem możliwości ich implementacji w systemie tłumaczenia. Brano pod uwagę aspekt możliwości zastosowana wyjścia algorytmu tłumaczenia (zdania języka angielskiego) jako wejścia do syntezatora. Do dalszych doświadczeń wybrano syntezator Infovox. Umożliwia on bezpośrednie wyjście foniczne dobrej jakości.

BIBLIOGRAFIA

[1] Graliński F., (1998) *Realizacja pół-automatycznej ekstrakcji leksemów występujących w korpusie polskich tekstów informatycznych*, w : "Speech and Language Technology Vol. 2", Poznań.

[2] Jassem K., (1997) *POLENG - a Machine Translation System Based on an Electronic Dictionary*, w: "Speech and Language Technology Vol.1", Poznań.

[3] Jassem K., Lison M., Mączynski R., (1997) *The implementation of a bilingual dictionary amenable to on-line modification*, w: "Speech and Language Technology. Vol.1", Poznań.

[4] Jassem K., (1998) *Format wejścia w polsko-angielskim słowniku elektronicznym konstruowanym dla potrzeb tłumaczenia automatycznego*.

[5] Lison M., (1998) *Model i realizacja struktury danych leksykalnych słownika elektronicznego skonstruowanego dla potrzeb automatycznego tłumaczenia*, w: "Speech and Language Technology. Vol.2" Poznań.

[6] Obrębski T., (1998) *Wykorzystanie lematyzatora słownika POLEX do oznaczania form wyrazowych w korpusie tekstów informatycznych*, w: "Speech and Language Technology. Vol.2", Poznań.

[7] Piasecki M., (1998) *Wybrane aspekty reprezentacji semantycznej określników języka polskiego*, w: "Speech and Language Technology. Vol. 2", Poznań.

[8] Rutkowski B., (1998) *Edytor polsko-angielskiego słownika w formacie SGML*, w: "Speech and Language Technology. Vol. 2", Poznań.

[9] Szymczak M. (ed.) (1982) *Słownik języka polskiego*. Warszawa. PWN.

[10] Wypych M., (1998) *Przegląd rynku syntetyzatorów mowy*, w: "Speech and Language Technology. Vol.2", Poznań.