WNIOSEK O PORTFOLIO:

Weryfikacja koncepcji możliwości budowy wielojęzycznego systemu komputerowego przekładu typu Human-Aided Machine Translation opartego na wykorzystaniu idei języka pośredniczącego

Autorzy: dr inż. Mirosław Gajer

Centrum Inteligentnych Systemów Informatycznych Akademia Górniczo-Hutnicza im. Stanisława Staszica al. Mickiewicza 30, 30-059 Kraków
budynek C-2 pokój 426 tel.: 12 617 44 53 www.isi.agh.edu.pl **isi@agh.edu.pl**

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Opis merytoryczny

Celem działań planowanych na najbliższy okres jest przygotowanie wniosku grantowego dotyczącego analizy możliwości budowy prototypu systemu komputerowego przekładu typu Human-Aided Machine Translation opartego na zastosowaniu koncepcji języka pośredniczącego. Realizacja pierwszego etapu działań zakończy się opracowaniem dokumentacji do projektu grantowego dotyczącego systemu komputerowego przekładu typu Human-Aided Machine Translation. Rozważana dokumentacja będzie miała formę publikacji naukowej. Ostateczny termin zakończenia związanych z tym działań planowany jest na dzień 30.06.2014. W pracach nad dokumentacją umożliwiającą późniejsze przygotowanie wniosku grantowego docelowo weźmie udział interdyscyplinarny zespół złożony z trzech osób:

- dr inż. Mirosław Gajer (AGH, Katedra Informatyki Stosowanej),
- dr Joanna Dybiec-Gajer (Uniwersytet Pedagogiczny w Krakowie),
- dr inż. Zbigniew Handzel (Uniwersytet Jagielloński).

1. Opis

Celem wnioskowanego projektu jest dokonanie weryfikacji koncepcji dotyczącej możliwości budowy i praktycznego wykorzystania wspomaganego przez użytkownika systemu komputerowego przekładu typu Human-Aided Machine Translation opartego na zastosowaniu koncepcji języka pośredniczącego.

Wnioskowany system ma być z założenia systemem wielojęzycznym i ma umożliwić przekład w dowolnym kierunku pomiędzy dowolnie wybraną parą języków. W początkowym etapie weryfikacji koncepcji Human-Aided Machine Translation system testowany będzie głównie dla języków polskiego i angielskiego. W etapie kolejnym zostaną dodane również inne języki, przede wszystkim z germańskiej grupy językowej (niemiecki, niderlandzki, afrikaans, szwedzki, norweski, duński, islandzki) oraz z grupy języków romańskich (francuski, portugalski – w wersji europejskiej i brazylijskiej, hiszpański, włoski, kataloński, rumuński). Dalsza weryfikacja koncepcji będzie miała na celu zbadanie możliwości rozszerzenie funkcjonalności systemu o wybrane języki słowiańskie (rosyjski, ukraiński, czeski, słowacki, słoweński, chorwacki, serbski, macedoński, bułgarski) oraz inne języki należące do rodziny indoeuropejskiej (grecki, albański, ormiański, walijski, perski, hindi, urdu, pendżabski, bengali). Docelowo wnioskowany system uwzględniał będzie także wybrane języki nie należące do

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

indoeuropejskiej rodziny językowej, które są określane mianem języków nostratycznych (arabski, hebrajski, turecki, węgierski, fiński, estoński, koreański, japoński). Planowane jest także podjęcie próby rozszerzenia funkcjonalności systemu na inne wybrane języki orientalne, takie jak chiński (mandaryński i kantoński), tajski, wietnamski, indonezyjski, malajski i tagalski.

Na wstępie należy zaznaczyć, że przeznaczenie planowanego systemu jest nieco odmienne, niż ma to miejsce w przypadku w pełni automatycznych systemów komputerowego przekładu, gdzie na wejście systemu zawsze zadawany jest gotowy już tekst w języku wyjściowym, który następnie tłumaczony jest przez komputer, a na wyjściu systemu otrzymujemy gotowy produkt w postaci przekładu rozważanego tekstu na wybrany język docelowy. Tymczasem w przypadku systemu typu Human-Aided Machine Translation tłumaczenie tekstu ma miejsce bezpośrednio na etapie jego powstawania, przy czym tekst podlegający przekładowi tworzony jest przez użytkownika w tzw. języku kontrolowanym, który dopuszcza do użytku tylko wybrane konstrukcje składniowe oraz ma precyzyjnie zdefiniowaną warstwę leksykalną. Nałożenie tego rodzaju więzów na tekst tworzony w języku wyjściowym stanowi gwarancję odpowiednio wysokiej jakości uzyskiwanych przekładów, ponieważ w dużym stopniu wyeliminowane zostaje wówczas zjawisko wieloznaczności języka naturalnego, zarówno na poziomie leksyki, jak i morfologii oraz składni.

Opracowywany system ma być w zamierzeniu systemem otwartym zarówno pod względem leksykalnym, jak i składniowym. Otwarcie leksykalne polega na tym, że dopuszczalna jest rozbudowa zasobów słownictwa języka poprzez dodawanie do lingwistycznych baz danych nowych rekordów – jednostek leksykalnych. Z kolei otwarcie składniowe polega na dopuszczeniu możliwości rozbudowy zestawu dozwolonych struktur składniowych o nowe typy związków wyrazowych.

Jeżeli chodzi o zasoby leksykalne tworzonego systemu, to oprócz podstawowego słownictwa należącego do sfery języka ogólnego, planowane jest rozwijanie zasobów leksykalnych, głównie pod kątem przyszłych zastosowań systemu w obszarze nauk ścisłych, (matematycznych, technicznych, ekonomicznych), a także lingwistycznych, przyrodniczych, biologicznych i medycznych.

Zaproponowane podejście oparte na zastosowaniu języków kontrolowanych jako języków wyjściowych przekładu powinno istotnie podnieść jakość tłumaczeń otrzymywanych przy użyciu systemu typu Human-Aided Machine Translation.

Wnioskowany system może być prawdopodobnie z powodzeniem stosowany w roli interaktywnych rozmówek turystycznych, z możliwością dalszej rozbudowy w kierunku dodania modułów rozpoznawania i syntezy mowy.

Ponieważ działanie systemu oparte jest na idei wykorzystania języka pośredniczącego przekładu, ewentualna rozbudowa systemu o dalsze języki będzie stosunkowo prosta, gdyż polegać będzie na dodaniu modułu

tłumaczącego z danego języka do języka pośredniczącego i z języka pośredniczącego na dany język.

Reasumując, wnioskowany system ma być z założenia systemem otwartym zarówno semantycznie, jak i składniowo z możliwością systematycznej rozbudowy o kolejne języki. Tłumaczenie z języków kontrolowanych jest gwarantem wysokiej jakości uzyskiwanych przekładów.

W realizacje projektu zaangażowany będzie docelowo kilkunastoosobowy zespół składający się zarówno z informatyków (programowanie systemów sieciowych i urządzeń mobilnych), jak i osób z wykształceniem językoznawczym lub filologicznym.

2. Charakterystyka i typ potencjalnych nabywców

Rozwijany system ma w założeniu umożliwić sporządzanie wielu różnych wersji językowych dokumentów tekstowych. System umożliwi ich równoczesne wytworzenie dzięki koncepcji zastosowania języka pośredniczącego przekładu. W ten sposób można będzie na przykład jednocześnie tworzyć wiele wersji językowych wiadomości SMS lub tekstów poczty elektronicznej.

Ponadto proponowany system może z powodzeniem zastąpić tradycyjne rozmówki turystyczne, zapewniając tym samym efektywną wzajemną komunikację osób nie znających żadnego wspólnego im języka.

Co istotne, ponieważ wnioskowany system ma być z założenia systemem wielojęzycznym, dlatego może znaleźć nabywców nie tylko w Polsce, ale również i w innych krajach, których języki narodowe będą w systemie uwzględnione.

3. Opis istniejących materiałów promocyjnych

Istnieje prezentacja przedstawiająca podstawowe idee związane z zasadami funkcjonowania systemu typu Human-Aided Machin Translation.

Ponadto najbliższym czasie ukażą się w recenzowanych czasopismach naukowo-technicznych artykuły poświęcone zagadnieniom związanym z realizacją koncepcji wnioskowanego systemu.

4. Potencjalni rozmówcy

Dr inż. Mirosław Gajer – Katedra Informatyki Stosowanej AGH
Dr Joanna Dybiec-Gajer – Katedra Przekładoznawstwa UP (tłumacz przysięgły języków angielskiego i niemieckiego)

5. Kierunki potencjalnego zastosowania projektu

Jednoczesne tworzenie wielu wersji językowych tekstów poczty elektronicznej i wiadomości SMS.

Mobilne wielojęzyczne rozmówki turystyczne z dodatkową opcją rozpoznawania i syntezy mowy.

6. Silne i słabe strony projektu

Do silnych stron projektu można zaliczyć:

- Nowatorskie podejście do zagadnienia przekładu komputerowego z wykorzystaniem języków kontrolowanych (system typu typu Human-Aided Machine Translation);
- Oparcie działania systemu na wykorzystaniu języka pośredniczącego przekładu – łatwość rozszerzania systemu o nowe języki;
- Otwartość leksykalna systemu – możliwość poszerzania zasobów słownikowych o nowe jednostki;
- Otwartość składniowa systemu – możliwość wprowadzania nowych konstrukcji gramatycznych;
- Wielojęzyczność systemu – możliwość tłumaczenia z dowolnego języka uwzględnionego w systemie na dowolny inny język w nim występujący.

Za słabą stronę projektu można uznać stosunkowo duży nakład pracy związany z akwizycją wzorców translacyjnych z wielu różnych języków potrzebnych do sprawnego funkcjonowania systemu.

Ponadto pewnym problemem może być fakt, że do korzystania z systemu typu Human-Aided Machine Translation konieczne jest posiadanie przez jego użytkownika elementarnej wiedzy o języku (wiadomości typu, co to jest podmiot, orzeczenie i dopełnienie zdania, co to są zaimki osobowe, wskazujące i dzierżawcze, czym różni się strona czynna od biernej itp.), co może stanowić w praktyce swego rodzaju barierę utrudniająca jego powszechne wykorzystanie. Być może krótkie szkolenie dotyczące zasad korzystania z systemu będzie w stanie skutecznie rozwiązać ten problem.

7. Czynniki ryzyka

Ponieważ podejście do komputerowego tłumaczenia typu Human-Aided Machine Translation oparte na językach kontrolowanych i języku pośredniczącym przekładu jest podejściem nowatorskim i zgodnie z wiedzą wnioskodawcy nie istnieją jeszcze tego typu wielojęzyczne systemy, jest zapewne sprawą dyskusyjną, czy takie nowe rozwiązanie może kiedyś w przyszłości zdobyć większą popularność i stać się dużym sukcesem rynkowym oraz czy potencjalni użytkownicy do pracy z tego typu systemem będą się w stanie przekonać i przyzwyczaić (jest przy tym rzeczą niezmiernie ważną, aby obsługa systemu i współpraca z nim były możliwie jak najbardziej intuicyjne oraz aby obsługa systemu wymagała od jego użytkownika jedynie absolutnie niezbędnego minimum wiedzy o języku).