WNIOSEK O PORTFOLIO:

Opracowanie koncepcji wielojęzycznych generatorów gramatycznych pełniących rolę narzędzi informatycznych typu Machine-Aided Human Translation

Autorzy: Mirosław Gajer, Zbigniew Handzel, Joanna Dybiec-Gajer, Joanna Rokieta-Jaśków

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

Celem planowanych prac badawczych jest przygotowanie wniosku grantowego dotyczącego analizy możliwości budowy i dalszego rozwoju narzędzi informatycznych typu MAHT (ang. Machine-Aided Human Translation).
Realizacja wnioskowanych prac badawczych zakończy się opracowaniem dokumentacji do projektu grantowego dotyczącego wielojęzycznych generatorów gramatycznych wspomagających działalność translatorską oraz proces dydaktyki języków obcych. Rozważana dokumentacja będzie miała formę publikacji naukowej. Ostateczny termin zakończenia związanych z niniejszym wnioskiem działań planowany jest na dzień 30.11.2014. W pracach nad dokumentacją umożliwiającą późniejsze przygotowanie wniosku grantowego docelowo weźmie udział interdyscyplinarny zespół badawczy złożony z czterech osób, w skład którego wejdą zarówno osoby z wykształceniem technicznym (informatyka), jak i filologicznym:

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

1. Opis merytoryczny

Celem wnioskowanego projektu jest dokonanie weryfikacji koncepcji dotyczącej możliwości budowy i praktycznego wykorzystania narzędzi informatycznych pełniących rolę generatorów gramatycznych wspomagających pracę językoznawców, filologów, przekładoznawców i tłumaczy. Rozważane generatory gramatyczne mogą być z dużym prawdopodobieństwem wykorzystywane z powodzeniem także na potrzeby wspomagania procesu dydaktyki języków obcych.
W przypadku wielu współczesnych języków należących do wielkiej rodziny języków indoeuropejskich mamy do czynienia z bardzo rozbudowaną fleksją czasownikową. Ponadto zjawisko odmiany wybranych części mowy znane jest także i w innych rodzinach językowych (afroazjatyckiej, ałtajskiej i uralskiej), gdzie bywa określane mianem alternacji bądź aglutynacji. W przypadku wymienionych rodzin językowych typowy czasownik może posiadać nawet kilkadziesiąt różnych form fleksyjnych, ponieważ zwykle wykazuje niezwykle rozbudowaną odmianę przez osoby, liczby i rodzaje gramatyczne. Ponadto indoeuropejskie czasowniki odmieniają się także poprzez rozbudowany system czasów i trybów gramatycznych. Wszystko to sprawia, że czynne opanowanie w zasadzie dowolnego języka indoeuropejskiego jako języka obcego z reguły sprawia spore trudności, a poprawne tworzenie odpowiednich form fleksyjnych czasowników jest niełatwą do opanowania sztuką. Jako swego rodzaju remedium na wspomniane niedogodności można wskazać tzw. generatory gramatyczne, czyli programy komputerowe pozwalające na automatyczne tworzenie form fleksyjnych dla wybranych części mowy, takich jak na przykład czasowniki, rzeczowniki, przymiotniki bądź zaimki osobowe lub dzierżawcze. W chwili obecnej tego rodzaju proste generatory gramatyczne istnieją już dla wielu języków indoeuropejskich i są ponadto powszechnie dostępne za pośrednictwem odpowiednich stron internetowych. Jednak proponowane przez autorów niniejszego wniosku generatory gramatyczne są pod

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

wieloma względami rozwiązaniami unikatowymi, nie posiadającymi dotychczas żadnych znanych powszechnie dostępnych odpowiedników.
Przede wszystkim głównym założeniem wnioskowanego projektu jest, że projektowane w jego ramach generatory gramatyczne mają być dostępne w postaci systemu wielojęzycznego otwartego na możliwość systematycznego rozbudowywania poprzez dodawanie do niego nowych języków. Tego rodzaju elastyczność zostanie uzyskana przez zastosowanie koncepcji języka pośredniczącego przekładu (tzw. interlingua). W roli języka pośredniczącego przekładu planowane jest wykorzystanie sztucznego języka *ido*, który stanowi zmodernizowaną, pozbawioną pewnych istotnych mankamentów i znacznie ulepszoną w stosunku do swego pierwowzoru wersję języka *esperanto*.
Podstawowym zadaniem wnioskowanego wielojęzycznego generatora gramatycznego jest dokonywanie odmiany przez osoby, liczby, rodzaje gramatyczne, a także różne czasy i tryby wybranych przez użytkownika fraz czasownikowych VP (ang. verb phrase), które zbudowane są z odpowiednio wybranego przez użytkownika czasownika V (ang. verb) i występującej po nim frazy rzeczownikowej NP (ang. noun phrase), zgodnie z następującą regułą bezkontekstowej gramatyki transformacyjno-generatywnej: VP → V + NP.
Tego rodzaju frazy czasownikowe posiadają w zdecydowanej większości przypadków precyzyjnie zdefiniowane znaczenie, podczas gdy występujący w izolacji czasownik jest przeważnie wieloznaczny. Na przykład bez wątpienia wieloznacznym jest polski czasownik „kopać", podczas gdy frazy czasownikowe „kopać piłkę", „kopać rów" czy też „kopać pod kimś dołki" posiadają już tylko jedno precyzyjnie określone znaczenie, w związku z czym ich automatyczny przekład na dowolnie wybrany język obcy nie powinien już nastręczać większych trudności.
Jak już uprzednio wspomniano, zadaniem wnioskowanego wielojęzycznego generatora form fleksyjnych fraz czasownikowych jest dokonywanie ich automatycznej odmiany, a także ich automatyczne tłumaczenie na wybrany przez użytkownika język obcy. Proces tłumaczenia maszynowego odmienianych fraz czasownikowych realizowany będzie w dwóch odrębnych etapach. W etapie pierwszym wybrana przez użytkownika fraza czasownikowa tłumaczona jest język pośredniczący przekładu, którym jest, jak już wspomniano, sztuczny język *ido*, pełniący rolę języka komunikacji międzynarodowej. W etapie kolejnym odpowiednia fraza czasownikowa należąca do języka *ido* tłumaczona jest na uprzednio wybrany przez użytkownika język docelowy.
Proponowany przez autorów system może być z powodzeniem wykorzystywany przez profesjonalnych tłumaczy jako narzędzie informatyczne typu MAHT (ang. Machine-Aided Human Translation). W tym wypadku dużą zaletą proponowanego systemu jest nie tylko podpowiadanie tłumaczowi poprawnej formy fleksyjnej wybranej przez niego frazy czasownikowej, ale także to, że system stanowił będzie swego rodzaju leksykon związków frazeologicznych, bez znajomości których nie jest możliwy poprawny przekład.
Innym potencjalnym obszarem zastosowań wnioskowanego systemu jest dydaktyka języków obcych. W opinii autorów projektowany system może stanowić nieodzowną pomoc w nauce języków obcych, ponieważ może posłużyć jako narzędzie informatyczne umożliwiające nie tylko opanowanie fleksji i reguł składniowych języków obcych, ale także zapoznanie się z bogactwem frazeologii danego języka.

2. Charakterystyka i typ potencjalnych nabywców

Wnioskowane narzędzia informatyczne typu MAHT powinny znaleźć nabywców wśród językoznawców, filologów, tłumaczy oraz dydaktyków przekładu i osób uczących się języków obcych.

Niezmiernie istotną cechą jest fakt, że wnioskowany system ma być z założenia systemem wielojęzycznym i z tego powodu 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 generatorów gramatycznych.
Ponadto w najbliższym czasie ukażą się w recenzowanych czasopismach naukowo-technicznych artykuły poświęcone zagadnieniom związanym z realizacją podstawowych 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)
Dr hab. Anna Turula – Wydział Neofilologii, Uniwersytet Pedagogiczny w Krakowie

5. Kierunki potencjalnego zastosowania projektu

Wykorzystanie wnioskowanego systemu typu HAMT przez tłumaczy jako narzędzi informatycznych wspomagających proces przekładu.
Pełnienie przez wnioskowany system roli pomocy dydaktycznych wspierających proces nauki języków obcych.

6. Silne i słabe strony projektu

Do silnych stron projektu można zaliczyć:

- Nowatorskie podejście do zagadnienia budowy wielojęzycznych generatorów gramatycznych;
- Oparcie działania systemu na wykorzystaniu języka pośredniczącego przekładu – łatwość rozszerzania systemu o nowe języki;
- Otwartość leksykalna systemu – możliwość poszerzania zasobów słownikowych o nowe jednostki;

- Wielojęzyczność systemu – możliwość tłumaczenia z dowolnego języka uwzględnionego w systemie na dowolny inny język w nim występujący.

Za słabą stronę projektu można uznać fakt, że proponowane narzędzia informatyczne typu MAHT mają unikalny charakter i zgodnie z wiedzą wnioskodawców na świecie nie istnieją jeszcze tego rodzaju programy komputerowe. W związku z tym trudno jest przewidzieć, czy taki nowy pomysł znajdzie szersze zastosowanie, gwarantujące tym samym sukces rynkowy.

7. Czynniki ryzyka

Wykorzystanie do wspomagania procesu przekładu automatycznych generatorów gramatycznych 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 gwarantować tym samym duży sukces rynkowy. Trudno jest także na bieżącym etapie wyrokować, czy potencjalni użytkownicy do pracy z tego typu systemem będą w stanie się w jakimś stopniu przekonać, tak aby mógł być w praktyce stosowany na szerszą skalę.