

„Język zapytań dla leksykalnej bazy danych typu WordNet”

Marek Kubis

Stypendysta projektu pt. „Wsparcie stypendialne dla doktorantów na kierunkach uznanych za strategiczne z punktu widzenia rozwoju Wielkopolski”, Poddziałanie 8.2.2 Programu Operacyjnego Kapitał Ludzki

Celem rozprawy jest opracowanie języka zapytań dla leksykalnych baz danych typu wordnet. Struktura, jaką tworzą dane zgromadzone w wordnetach, jest różna od tej, którą tworzą dane przechowywane w bazach relacyjnych, obiektowych czy semi-strukturalnych. Wordnet jest siecią wzajemnie powiązanych zbiorów wyrażeń synonimicznych (synsetów). Pojedynczy synset reprezentuje wspólne znaczenie wyrażeń wchodzących w jego skład. Synsety są połączone relacjami, które reprezentują związki zachodzące pomiędzy znaczeniami reprezentowanych przez nie wyrażeń. Dla modelu relacyjnego, obiektowego oraz semi-strukturalnego utworzono szereg dedykowanych języków programowania (języków zapytań), które pozwalają konstruować wyrażenia opisujące kryteria, jakie powinny spełniać dane, które użytkownik chce pobrać lub zaktualizować. Dla baz danych typu wordnet nie stworzono dotychczas języka zapytań o możliwościach zbliżonych do tych, które występują w językach przeznaczonych dla innych modeli danych. Brak odpowiedniego języka zapytań jest szczególnie niekorzystny, biorąc pod uwagę zastosowania leksykalnych baz danych w dziedzinie lingwistyki komputerowej i sztucznej inteligencji, do których należą m. in. dezambiguacja znaczeń, analiza semantyczna języka naturalnego, wyszukiwanie i ekstrakcja informacji. W wyniku realizacji celu rozprawy powstanie oprogramowanie WQuery - pierwszy, w pełni funkcjonalny system zarządzania leksykalną bazą danych typu wordnet oparty na dedykowanym języku zapytań.

Wstępna wersja systemu WQuery została wykorzystana do zintegrowania leksykalnej bazy danych PolNet z aplikacją POLINT-112-SMS stanowiącą prototyp systemu z kompetencją językową wspomagającego systemy monitoringu imprez masowych realizowanego w ramach projektu „Technologie przetwarzania tekstu polskiego zorientowane na potrzeby bezpieczeństwa publicznego” Polskiej Platformy Bezpieczeństwa Wewnętrznego.

Pełna wersja systemu WQuery (stanowiąca rezultat rozprawy doktorskiej) będzie mogła zostać wykorzystana w analogiczny sposób przez instytucje naukowe i przedsiębiorstwa z regionu Wielkopolski jako element składowy systemów analizy języka, chatterbotów, systemów ekstrakcji informacji oraz innych aplikacji wymagających dostępu do danych zgromadzonych w leksykalnych bazach danych typu wordnet. Planowane jest m. in. wdrożenie pełnej wersji systemu WQuery w kolejnych fazach projektu POLINT-CITTA-SMS: CITy Tour Assistant przeznaczonego do usprawnienia obsługi turystycznej w Poznaniu.

Przygotowane w ramach rozprawy doktorskiej oprogramowanie będzie można również zastosować jako samodzielne narzędzie do rozwijania danych leksykalnych w obrębie wielkopolskich przedsiębiorstw np. w procesie konstrukcji semantycznych katalogów produktów wspomagających wyszukiwanie informacji na firmowych stronach www. Wdrożenie gotowego systemu zarządzania leksykalnymi bazami danych typu wordnet (analogicznie jak w przypadku systemów zarządzania bazami danych innych typów) wiąże się z poniesieniem niższych nakładów finansowych przez przedsiębiorstwo niż w przypadku budowania systemu tego typu od podstaw.