ZESZYTY NAUKOWE WYDZIAŁU ETI POLITECHNIKI GDAŃSKIEJ
Nr 5 Seria: Technologie Informacyjne 2007

Bartosz Kamiński, Olaf Tomczak, Julian Szymański

Katedra Architektury Systemów Komputerowych, Politechnika Gdańska

WONTOUGO[1] – KOOPERACYJNY EDYTOR WORDNETU

Streszczenie

Artykuł zawiera opis systemu pozwalającego na kooperacyjną edycję słownika WordNet. W ramach zrealizowanego projektu wykonano interaktywny interfejs użytkownika w postaci aplikacji opartej na bibliotece TouchGraph umożliwiający wizualną pracę nad semantycznym słownikiem w środowisku rozproszonym. Przedstawiony został sposób przeniesienia słownika z jego źródłowej wersji zorganizowanej na plikach do relacyjnej bazy danych i możliwości, jakie daje opracowana aplikacja w zakresie zespołowej pracy nad słownikiem w środowisku internetowym. Omówiona została również technologia, w której projekt został zrealizowany oraz korzyści płynące z przyjętych rozwiązań ze szczególnym naciskiem na możliwości dalszego rozwoju projektu.

1. WSTĘP

WordNet [1] jest leksykalną bazą danych języka angielskiego inspirowaną psycholingwistycznymi teoriami dotyczącymi organizacji danych językowych przez człowieka. Rozwijany jest od 1985 roku przez Cognitive Science Laboratory na uniwersytecie w Princeton.

System WordNet różni się od budowy tradycyjnego słownika. Organizacja danych w nim nie polega na liście słów wraz z definicjami, lecz opiera się na koncepcji znaczenia reprezentowanego przez synset wraz z powiązaniami leksykalnymi i semantycznymi między nimi. Synset to pojęcie opisujące grupę znaczeniową, do której należą słowa będące synonimami wraz z ich definicją. Zbiór semantycznych zależności pomiędzy koncepcjami ma na celu szczegółowe objaśnienie znaczenia danego i określenie jego kontekstu. Semantyka leksykalna implementowana jest w WordNecie przez odwzorowanie między formą słowa, a jego znaczeniem. Jest to związek wiele do wiele. Słowa mające te same znaczenia są synonimami. Te same słowa wchodzące w skład różnych synsetów są przypadkami polisemii (wieloznaczeniowości).

Tematem niniejszej publikacji jest Projekt Wontougo[2] zrealizowany w katedrze Architektury Systemów Komputerowych, wydziału ETI PG, którego celem jest utworzenie

[1] Nazwa pochodzi od słów „WordNet in TouchGraph”.
[2] www. wontougo.org

kooperacyjnego edytora dla sytemu WordNet z wykorzystaniem bibliotek pozwalających na wizualizować grafy w interakcji z użytkownikiem - TouchGraph [2].

2. CELE PROJEKTU

Wontougo jest aplikacją, której celem jest umożliwienie przeglądania i kooperacyjnej edycji słownika WordNetu przy pomocy graficznego interfejsu. Realizacji projektu przyświecały następujące cele:

1. zaprojektowanie obiektowego modelu danych opartego na bazie danych systemu WordNet oraz metod utrwalania i przechowywania obiektów modelu.
2. zaprojektowanie kontenera do składowania kontekstowych danych i opracowanie warstwy logiki biznesowej umożliwiającej odczyt danych z bazy WordNet i ich konwersję na obiekty modelu,
3. opracowanie graficznego interfejsu użytkownika, umożliwiającego przejrzystą prezentację danych i dokonywanie operacji związanych z edycją słownika,
4. zaadaptowanie aplikacji do warunków i ograniczeń środowiska rozproszonego.

3. ARCHITEKTURA SYSTEMU

System Wontougo został napisany w języku Java. Architekturę oparto o technologię Java Enterprise Edition w wersji 5.0[3]. Projekt modelu dziedziny (domain model) został zaimplementowany w postaci komponentów encyjnych Enterprise Java Beans w wersji 3.0. Obiekty modelu, są w wyniku mapowania obiektowo-relacyjnego zamieniane na krotki relacyjnej bazy danych SQL, którymi zarządzać może dowolny system zarządzania bazą danych. Istniejąca implementacja oparta została o bazę danych MySQL. Cała logika związana z operacjami odczytu i zapisu obiektów dziedziny znajduje się w pojedynczym komponencie sesyjnym EJB. Komponent ten udostępnia dwa interfejsy: lokalny oraz zdalny, które mogą być wykorzystywane przez inne systemy. Stworzony graficzny interfejs użytkownika jest przykładem wykorzystania takiego zewnętrznego systemu – niezależnej aplikacji operującej poprzez udostępniony interfejs na modelu dziedziny dołączonym do systemu w formie biblioteki.

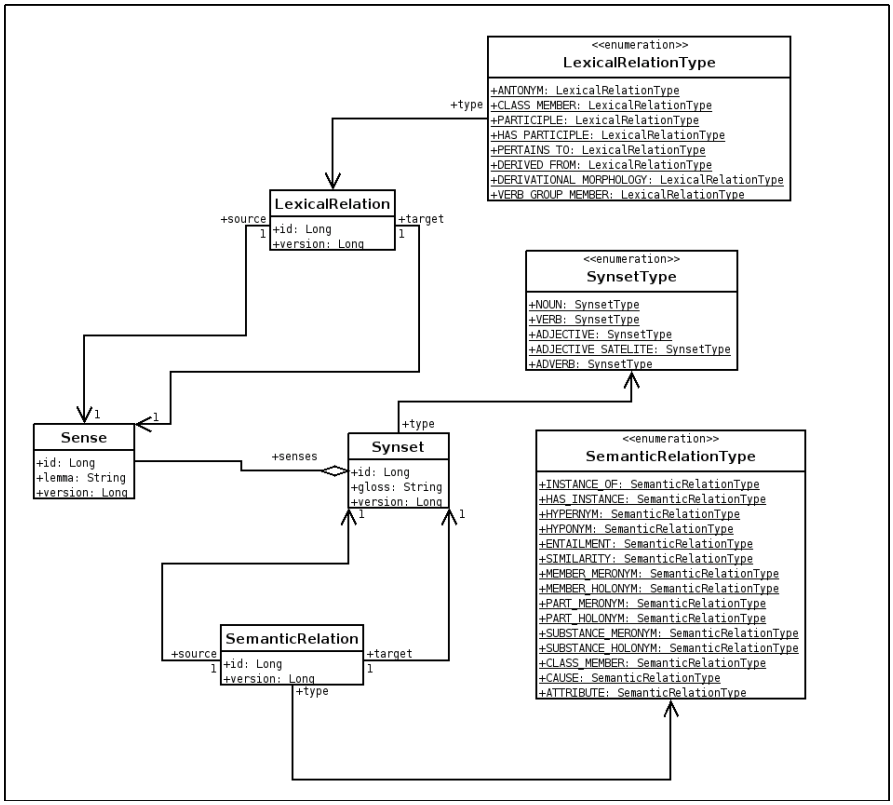
Wykorzystana technologia pozwoliła na implementację modelu dziedziny w postaci „lekkich" komponentów POJO (plain old java objects). Dodatkowo, cała logika związana z „utrwalaniem" obiektów modelu (*persistence*) i zarządzaniem transakcjami (*transaction management*) obsługiwana jest przez kontener, wewnątrz którego uruchamiana jest aplikacja. W przypadku tym może być to dowolny serwer aplikacji zgodny ze standardem JEEE 5.0. Dzięki takiemu podejściu system posiada zwartą i modyfikowalną architekturę.

3.1 Model dziedziny

System WordNet w swojej źródłowej wersji oparty jest na plikach tekstowych. W formie tej nie nadaje się on do kooperacyjnej pracy, dlatego konieczna była jego konwersja na model obiektowy, a następnie wykonany został jego eksport do relacyjnej bazy danych. Przeprowadzona konwersja odbyła się w dwóch etapach:

1. odczytaniu wszystkich krotek z plików tekstowych i stworzeniu na ich podstawie odpowiednich obiektów modelu dziedziny zaprojektowanego dla systemu Wontougo, (rys. 1)

2. wyeksportowaniu utworzonego w ten sposób modelu do schematu SQL, który następnie zapisany został do relacyjnej bazy danych. (rys. 2)

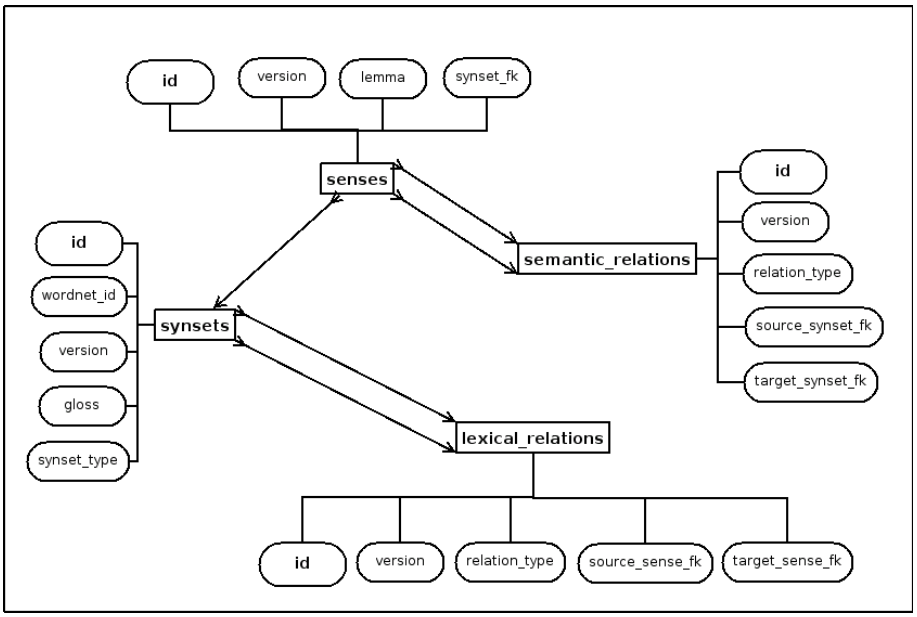


Rys.1. Model obiektowy struktury WordNetu.

Model dziedziny systemu Wontougo zawężony został do podstawowych powiązań odzwierciedlających strukturę bazy WordNet. Składają się na niego cztery podstawowe encje:

1. Synset – znaczenie (zbiór wyrazów bliskoznacznych opisujących pewną koncepcję),
2. Sense – sens (powiązanie między konkretnym wyrazem a jego znaczeniem),
3. LexicalRelation – związek leksykalny (powiązanie między konkretnymi sensami danego słowa),
4. SemanticRelation – związek semantyczny (powiązanie między synsetami).

Model ten w wyniku przekształcenia obiektowo-relacyjnego, mapowany jest na model relacyjnej bazy danych przedstawiony na diagramie ERD (Rys. 2).



Rys.2. Model relacyjnej bazy danych.

3.3 Funkcjonalność systemu Wontougo

System Wontougo udostępnia następujące podstawowe operacje na słowniku:

1. wyszukiwanie sensów danego słowa,
2. wyszukiwanie związków według typu oraz synsetu lub słowa związanego określoną relacją,
3. dodawanie, usuwanie, edycja synsetu,
4. dodawanie, usuwanie, edycja sensu,
5. dodawanie, usuwanie, edycja wybranego powiązania.

Kooperacyjne podejście do edycji leksykalnych danych pociąga za sobą konieczność wykonywania wszystkich działań edycyjnych przy użyciu transakcji w celu zabezpieczenia bazy danych przed utratą integralności. Zrealizowane zostało ono przez tzw. blokowanie optymistyczne oparte na liczniku wersji. Technika ta polega na opatrzeniu każdej encji w bazie danych numerem wersji. Numer ten jest inkrementowany przy każdym udanym zapisie obiektu. Z kolei przed zapisem wykonywany jest test zgodności numeru wersji krotki w bazie danych z aktualnym z numerem wersji zapisywanej instancji obiektu. Jeżeli numery te nie zgadzają się – transakcja zostaje wycofana.

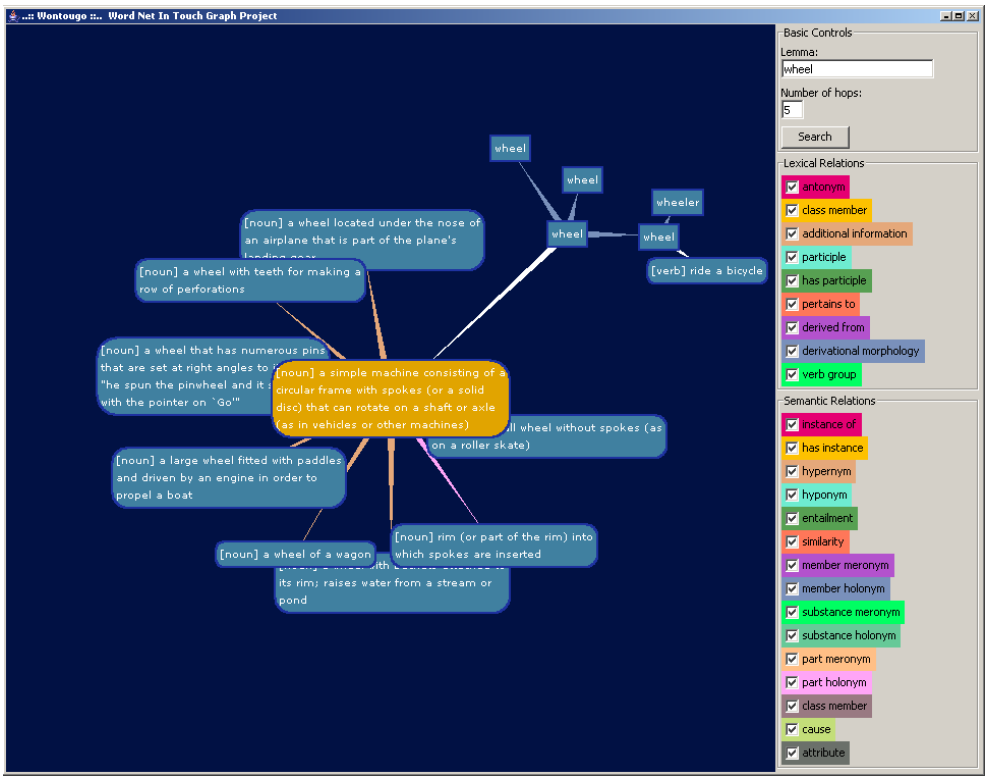
4. INTERAKTYWNA WIZUALIZACJA

Do prezentacji graficznej słownika wykorzystana została biblioteka TouchGraph [7]. Daje ona możliwość zaprezentowania słownika w formie grafu, którego węzłami są słowa i synsety, a krawędziami powiązania leksykalne i semantyczne. Dzięki wykorzystaniu tej

biblioteki użytkownik dostaje możliwość poruszania się po słowniku w przyjemnej i przejrzystej formie.

Przeglądanie słownika rozpoczyna się w momencie podania przez użytkownika interesującego go słowa (lemma). Dla podanej frazy, system odnajduje wszystkie skojarzone z nią znaczenia (synsety) i pozwala określić, od którego z nich ma się rozpocząć przeglądanie słownika.

Użytkownik aplikacji ma możliwość przeglądania kolejnych węzłów grafu. Kliknięcie węzła, pozwala na oglądniecie synsetów i słów powiązanych z wybranym węzłem. Przy rozwijaniu kolejnych węzłów, obiekty leżące w pewnej odległości od aktualnie wybranego węzła są usuwane. Ma to na celu zwiększenie czytelności grafu. Jest to szczególnie istotne w przypadku przeglądania bardzo bogato powiązanych fragmentów słownika. Długość ścieżki (hopów) między wybranym węzłem, a dowolnym innym wizualizowanym jest parametryzowana podczas przeglądania grafu.



Rys.3. Interfejs użytkownika Wontougo wizualizujący koncepcje „wheel”

Drugą funkcjonalnością pozwalającą ograniczyć ilość informacji wyświetlanych na ekranie, jest dobór powiązań (leksykalnych bądź semantycznych), które mają być wyświetlane w prezentowanym grafie.

Dla zwiększenia czytelności, każdy rodzaj powiązania oznaczony został innym kolorem - ułatwia to szybką identyfikację typu relacji. W przypadku niejasności

użytkownik może kliknąć lewym klawiszem myszy na powiązaniu, uzyskując w ten sposób podpowiedź zawierającą typ relacji.

4.1 Poruszanie się po grafie

Poruszanie się po grafie zostało zaplanowane w taki sposób, aby przeglądany w danej chwili obszar słownika był jak najbardziej aktualny, zagadnienie to jest kluczowe w kontekście pracy kooperacyjnej, gdzie informacja może zostać zmieniona w dowolnym momencie. Poniżej przedstawiono algorytm zapewniający aktualność informacji wyświetlanych w grafie, w obrębie przeglądanego obszaru. W opisie przyjęto, że wskazany węzeł jest węzłem K. Bieżący węzeł w grafie, ten, który był zaznaczony zaraz przed wyborem, to węzeł B.

1. weź węzeł *K*
2. usuń wszystkie węzły wychodzące „w przód" od węzła *K*, czyli te, dla których wszystkie drogi do węzła *B* prowadzą przez węzeł *K*,
3. z pozostałych węzłów usuń te wszystkie węzły, które są oddalone o więcej niż podana liczba hopów,
4. odśwież węzeł *K* odczytując jego zawartość z bazy danych
    a. jeżeli wybrane słowo lub synset już nie istnieje (wybrany węzeł mógł zostać usunięty równolegle przez innego użytkownika), poinformuj użytkownika, że węzeł już nie istnieje. Tu algorytm rozwijający węzeł kończy się.
    b. jeżeli wybrane słowo lub synset istnieje, pobierz je na nowo z bazy danych, utwórz odpowiedni węzeł i wstaw do grafu. Wstawiony węzeł nazwijmy *K'*.
5. pobierz węzły w otoczeniu węzła *K'*:
    a. dla słowa pobierz słowa w relacji(ach) wybranej(ych) przez użytkownika,
    b. dla synsetu, pobierz synsety w związku(ach) wybranej(ych) przez użytkownika,
6. dołącz pobrane węzły do grafu i połącz je odpowiednimi powiązaniami z węzłem *K'*. Dodatkowo, jeżeli węzeł *K'* jest
    a. słowem – wyświetl dla niego synset w osobnym węźle,
    b. synsetem – wyświetl wszystkie związane z nim słowa w osobnych węzłach,

    Zastrzeżenie: jeżeli dodawany węzeł już istnieje w grafie i nie jest połączony z węzłem *K'* żadną krawędzią, dodaj tylko odpowiednią krawędź między węzłem *K'*, a węzłem istniejącym. Mechanizm ten pozwala na pokazanie zawsze aktualnych powiązań między węzłem *K* (*K'*), a węzłami z jego bezpośredniego otoczenia.

7. sprawdź połączenia węzła *K'* z węzłami, które wcześniej były połączone bezpośrednio z węzłem *K*. Jeżeli żadne połączenie nie istnieje, oznacza to, że wszystkie połączenia zostały usunięte (przez innego użytkownika) i fragment grafu wychodzący od węzłów połączonych wcześniej z węzłem *K* nie powinien być dłużej wyświetlany – usuń ten fragment z grafu,
8. zapamiętaj węzeł *K'* jako bieżący (*B := K'*).

4.2 Optymalizacja rysowania węzłów

Wykorzystana wersja biblioteki TouchGraph (1.22) pozwala na wyświetlanie opisów węzłów w jednej linii. W związku ze znaczną długością opisów większości synsetów, konieczne było wprowadzenie dzielenia linii. Wprowadzona optymalizacja polega na ograniczeniu liczby generowania obrazów węzłów z zawiniętymi liniami. Raz wygenerowany obraz jest buforowany i wykorzystywany do czasu, gdy konieczne jest odświeżenie obrazu węzła. Odświeżenie obrazu węzła staje się konieczne w momencie, gdy użytkownik najedzie wskaźnikiem myszy nad węzeł. W sytuacji takiej w węźle zmienia się kolor tła i węzeł musi zostać przerysowany.

5. KOOPERACYJNA FUNKCJONALNOŚĆ

Edytor Wontougo pozwala na wykonywanie następujących operacji w rozproszonym środowisku:

1. dla synsetu:
    a. utworzenie nowego – polega na dodaniu do słownika nowego synsetu wraz z opisem, typem synsetu oraz pierwszym słowem należącym do niego,
    b. dopisanie nowego słowa,
    c. dopisanie nowego powiązania semantycznego z innym synsetem – polega na podaniu lemmy[3] (frazy bądź słowa), z którego synsetem ma zostać utworzony związek; dla podanej lemmy odszukiwane są słowa oraz opis ich znaczenia (pobierany z synsetów skojarzonych ze znalezionymi słowami); użytkownik wybiera synset, z którym chce utworzyć powiązanie; na końcu określa typ relacji,
    d. edycja opisu,
    e. edycja typu,
    f. usunięcie,
2. dla słowa:
    a. edycja lemmy,
    b. dopisanie nowego powiązania leksykalnego z innym słowem – polega na podaniu lemmy słowa, z którym ma zostać utworzony związek; dla podanej lemmy odszukiwane są słowa oraz opis ich znaczenia (pobierany z synsetów skojarzonych ze znalezionymi słowami); użytkownik wybiera słowo, z którym chce utworzyć powiązanie; na końcu określa typ relacji,
    c. usunięcie.

Edycja słownika z poziomu grafu odbywa się za pomocą menu kontekstowego dostępnego pod prawym klawiszem myszy. Kliknięcie na słowie, synsecie, bądź pustym obszarze grafu, dostaje odpowiednie opcje do wyboru.

Zmiany wprowadzane w aplikacji użytkownika są od razu utrwalane w bazie danych. Dzięki temu stają się widoczne dla innych użytkowników słownika. O aktualność wyświetlanych danych dba mechanizm odświeżania węzłów, opisany w punkcie 4.1.

[3] Lemma to jedno ze słów używanych do opisu synsetu, przedstawione w postaci kanonicznej - podstawowej formie wyrazu.

W aplikacji klienckiej wykrywane są zmiany wprowadzane przez dwóch użytkowników jednocześnie na tym samym elemencie słownika. W obecnej implementacji systemu, użytkownik zatwierdzający zmianę później od użytkownika równolegle edytującego ten sam element zostaje poinformowany o niemożliwości zatwierdzenia zmiany.

6. DALSZY ROZWÓJ SYSTEMU

W ramach dalszej pracy nad systemem przewidziano wymienione poniżej możliwości rozwoju.

1. Elastyczne łączenie zmian zatwierdzonych przez użytkowników w przypadku równoczesnej edycji jednego elementu słownika.
2. Wprowadzenie rejestrowania użytkowników edytujących bazę słownika oraz połączenie wprowadzanych zmian z kontami ich właścicieli.
3. Umożliwienie korzystania z wielu różnych słowników – dla różnych języków.
4. Wprowadzenie wersji językowych aplikacji klienckiej – zorganizowanie wszystkich napisów w formie zasobu i dodanie tłumaczenia.
5. Kooperacyjna praca wymaga rozwiązywania konfliktów powstających podczas pracy na wspólnych danych, implementacjach chwili obecnej przyjęto regułę, że zatwierdzane są zmiany użytkownika, który pierwszy dokonał edycji. W przyszłych implementacjach planowane jest dopisanie rozszerzenia mechanizmu pozwalające na łączenie wprowadzonych zmian.

BIBLIOGRAFIA

[1] http://wordnet.princeton.edu/
[2] http://sourceforge.net/projects/touchgraph/
[3] http://java.sun.com/javaee/
[4] Fellbaum C., editor.. *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication Series. The MIT Press, Cambridge MA. 1998
[5] Miller G. A. *WordNet: An on-line lexical database*. International Journal of Lexicography, 1990 3(4):235–312. Special Issue.
[6] Jaap Kamps *Visualizing WordNet Structure*, http://staff.science.uva.nl/~kamps/papers/visualize.pdf
[7] http://www.touchgraph.com/

WONTOUGO – COOPERATIVE WORDNET EDITOR

Summary

The article describes system for creating cooperative approach for editing WordNet dictionary in web environment. System provides interactive user interface for visualization semantic relations between dictionary elements. The article presents the way for transferring native WordNet files into relative database structures as well as the editorial capabilities of the client application. The production technology was also described together with profits of chosen solutions underlining future development capabilities.