УДК 681.3

Глибовець М. М., Марченко О. О., Никоненко А. О.

ПОБУДОВА УКРАЇНОМОВНОЇ ОНТОЛОГІЇ ЗАСОБАМИ СУБД

Розглянуто методи та засоби побудови лінгвістичних баз знань. Запропоновано підхід до створення асоціативно-семантичної бази знань засобами сучасних СУБД. Розроблено структуру та схему взаємодії об'єктів бази даних на основі запропонованого підходу.

1. Сучасна комп'ютерна лінгвістика

Сьогодні в галузі інтелектуальних систем розвивається багато різних напрямів. Значне місце серед них займають дослідження у сфері комп'ютерної лінгвістики, зокрема систем розпізнання семантики та тематичної направленості тексту [1].

Нині ми спостерігаємо дуже бурхливий розвиток інформаційних технологій та швидке зростання кількості інформації з усіх галузей науки та бізнесу. Особливістю накопичення інформації є її хаотичність та нестандартизованість, а це призводить до того, що більшість даних зберігається у форматі документів. Тому лінгвістична обробка природномовних текстів стала однією з центральних проблем інтелектуалізації інформаційних технологій.

Ще з середини 50-х років минулого століття значні зусилля науковців були спрямовані на розробку математичних алгоритмів та комп'ютерних програм обробки текстів природною мовою. Для автоматизації аналізу та синтезу текстів створювалися різноманітні моделі процесів обробки тексту, а також відповідні алгоритми та структури представлення даних. Більшість досліджень та розробок були присвячені розумінню англійської мови, в той час як головною проблемою для нашого географічного регіону є обробка текстів українською мовою. Поява досить великих електронних словників і лінгвістичних баз знань дала новий поштовх до розробки програмних засобів для семантичного аналізу і обробки текстів.

Одними з найбільш потужних та корисних засобів комп'ютерної лінгвістики сьогодні є онтології. Одним із визначень онтології є таке. Онтологія - це лінгвістична база знань, що містить опис понять трьох типів: об'єкти, властивості, дії. Для запису такого типу інформації застосовують концептуальну схему, що складається з об'єктів та різнотипних зв'язків між ними. Іноді до схеми додають правила або обмеження, що застосовуються до об'єктів. Онтологія може бути загальною (що містить велику кількість семантичних даних про мову та світ у цілому, але мало специфічних термінів та зв'язків, властивих певним предметним областям) чи спеціальною (що містить велику кількість упорядкованих, вузькоспеціалізованих даних та зв'язків між ними, які належать винятково до однієї предметної області).

2. Формати онтологій

Для спрощення роботи з онтологіями створено ряд мов опису онтологій. Метою таких мов є надання можливості задавати додаткову машинно-інтерпретовану семантику ресурсам, зробити машинне представлення даних більш наближеним до реального світу, підвищити можливості концептуального моделювання слабко структурованих Web-даних. Такий підхід поширився й на різноманітні мови опису онтології та на інструментальні засоби, призначені для роботи з ними. Сьогодні виділяють три основні класи мов опису онтологій:

- традиційні мови специфікації онтологій: Ontolingua, CycL та мови, засновані на дескриптивних логіках (такі як LOOM), також мови, засновані на фреймах (OKBC, OCML, Flogic);
- більш пізні мови, засновані на Web-стандартах (XOL, SHOE, UPML);
- спеціальні мови для обміну онтологіями через Web: RDF(S), DAML, OIL, OWL [2].

Коротко охарактеризуємо найбільш поширені та часто вживані мови опису онтологій.

Мова RDF. У рамках проекту семантичної інтерпретації інформаційних ресурсів Інтернету (Semantic Web) був запропонований стандарт опису метаданих документа Resource Description Framework, що використовує Xml-синтаксис.

RDF використовує базову модель даних «об'єкт - атрибут - значення» і здатний відіграти роль універсальної мови опису семантики ресурсів та взаємозв'язків між ними. Ресурси описуються у вигляді орієнтованого розміченого графа. Кожен ресурс може мати властивості, які

у свою чергу також можуть бути ресурсами або їхніми колекціями. Усі словники RDF використовують базову структуру, яка описує класи ресурсів і типи зв'язків між ними. Це дозволяє використовувати різнорідні децентралізовані словники, створені для машинної обробки за різними принципами й методами. Важливою особливістю стандарту є розширюваність: можна задати структуру опису джерела, використовуючи й розширюючи такі вбудовані поняття RDF-схем, як класи, властивості, типи, колекції. Модель схеми RDF включає наслідування класів і властивостей [3].

DAML+OIL - семантична мова розмітки Web-ресурсів, що розширює стандарти RDF і RDF Schema за рахунок більш повних примітивів моделювання. Остання версія DAML+OIL забезпечує багатий набір конструкцій для створення онтології й розмітки інформації таким чином, щоб їх могла читати й розуміти машина [4].

OWL (Web Ontology Language) - мова подання онтологій, що розширює можливості XML, RDF, RDF Schema і DAML+OIL. Цей проект передбачає створення потужного механізму семантичного аналізу. Планується, що в ньому буде усунено обмеження конструкцій DAML+OIL.

Онтології OWL - це послідовності аксіом і фактів, а також посилань на інші онтології. Вони містять компонент для запису авторства та іншої докладної інформації, є документами Web, на них можна посилатися через URI [5].

KIF (Knowledge Interchange Format, або формат обміну знаннями) - заснований на S-виразах синтаксис для логіки. KIF - це спеціальна мова, призначена для використання при обміні знаннями між різними комп'ютерними системами. Мова не призначена для внутрішнього представлення знань усередині комп'ютерних систем або всередині тісно зв'язаних наборів комп'ютерних систем (хоча може бути використана й для цієї мети). Мова була розроблена для опису загального формату представлення знань, незалежного від конкретних систем [6].

CycL (мова опису онтології Cyc) - це гібридна мова, що поєднує в собі властивості фреймів і логіку предикатів. CycL розрізняє такі сутності, як екземпляри, класи, предикати й функції. Синтаксис мови CycL схожий на синтаксис мови Lisp. Словник CycL складається з термів. Множину термів можна розділити на константи, терми (що не є атомами) і змінні. Крім цього, зустрічаються деякі інші типи об'єктів. Терми використовуються для складання значущих виразів CycL, які використовуються для формування суджень, з яких складається база знань [7].

Зважаючи на сказане вище, зрозуміло, що сьогодні не існує ні єдиної, формалізованої та стандартизованої мови для опису онтологій, ні єдиного загальновживаного формату збереження даних в онтологіях. Тому кожен розробник системи для обробки природномовних текстів вимушений розробляти свою онтологію з «нуля», починаючи з формату збереження даних і закінчуючи самим наповненням бази. З'явилися навіть спеціалізовані онтології, які дістали назву «організаційні». Звичайно, така ситуація не є прийнятною й дуже ускладнює, сповільнює та робить більш дорогою розробку нових лінгвістичних систем [8].

Наша розробка - перший крок у напрямі усунення ситуацій, коли для кожного проекту потрібно розробляти нову онтологію. Найближчим часом ми плануємо закінчити проект щодо створення єдиної онтологічної бази для програмних систем, що працюють з українською мовою, а в перспективі - і для російської, англійської та деяких інших європейських мов. Принципи організації онтологічної бази української мови та її структура й будуть описані далі у статті.

3. Структура даних

Для підтримки високої швидкості доступу до даних, цілісності та легкості роботи з даними було розроблено таку схему даних (рис. 1, 2).

Таблиця **Words** призначена для збереження слів української мови та лексичної інформації про них. Тобто фактично ця таблиця - тлумачний словник.

У таблиці **Part_of_speech_dict** зберігається інформація про частини мови, їх скорочення та коротка інформація про функції частин мови.

Таблиця **WordForms** містить інформацію про словоформи української мови у форматі «посилання на нормальну форму слова (що зберігається в таблиці Words), словоформа, лексична інформація про словоформу (рід, число, відмінок і т. д.)».

Пакет **Morphology_Processing** містить набір процедур та функцій для маніпуляції з даними між морфологічними об'єктами всередині бази даних. Пакет також є об'єктом бази даних, що значно поліпшує часові характеристики під час роботи з даними БД.

Таблиця **Users** зберігає інформацію про користувачів бази даних, надані їм права та привілеї на об'єкти БД, термін дії цих прав та іншу службову інформацію. Ця таблиця є зовнішньою для всіх об'єктів, що зберігають або обробляють інформацію про мову.

Таблиця **Expressions** призначена для збереження словосполучень української мови та лексичної інформації про них.

У таблиці **Words_in_Expressions** зберігається інформація про те, які слова та в яких поєднаннях входять до словосполучень.

Частина бази даних, що відповідає за роботу з морфологією:

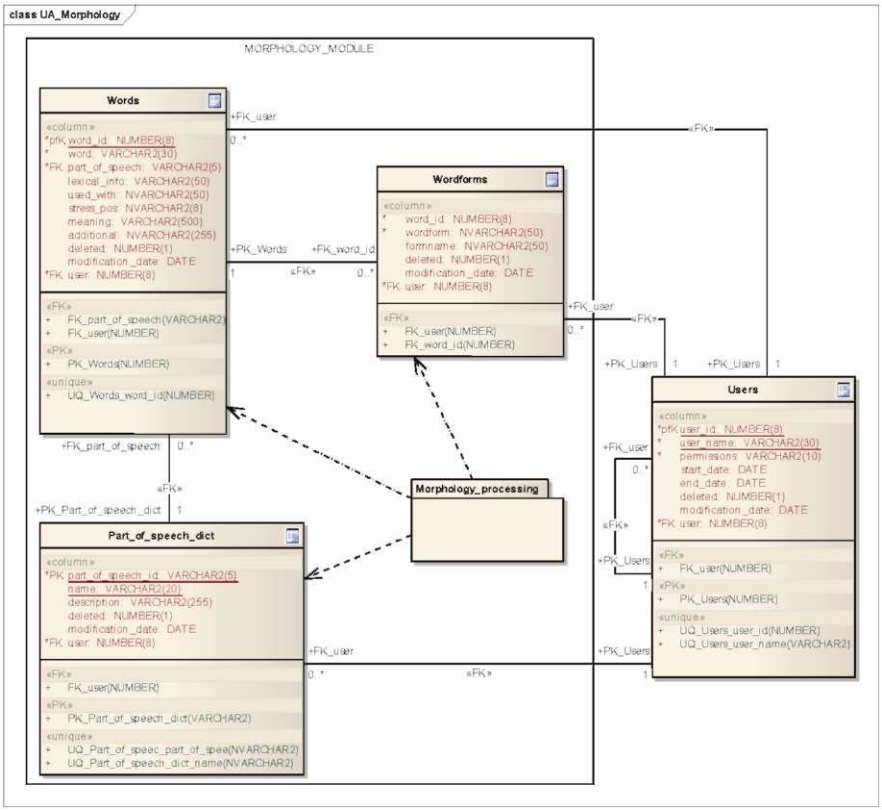

Рис. 1. Діаграма класів для даних про морфологію

Частина бази даних, що відповідає за _роботу з онтологією:

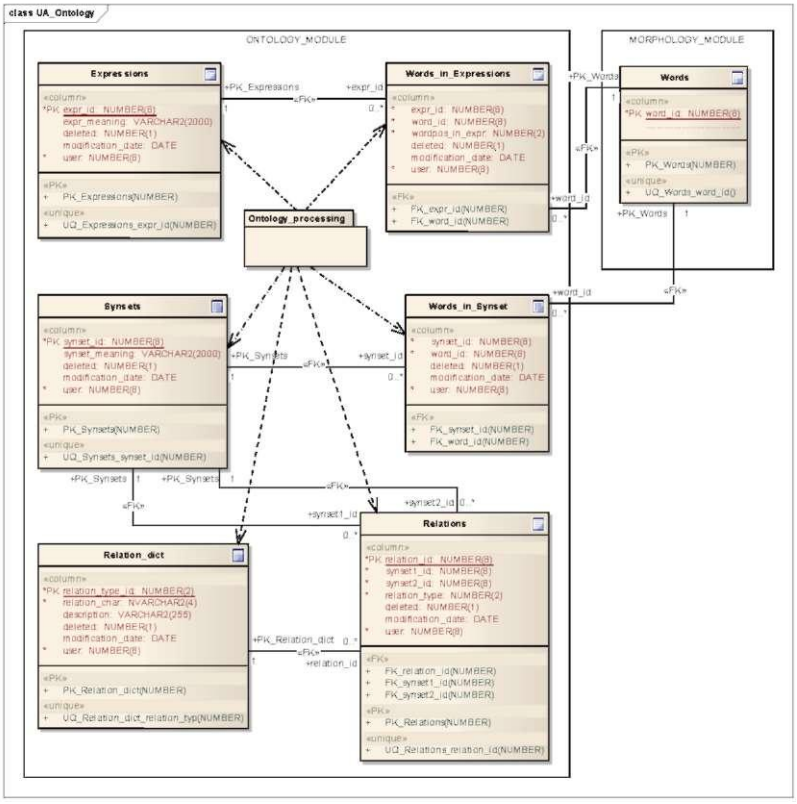

Рис. 2. Діаграма класів для даних про семантику

Таблиця **Synsets** зберігає інформацію про значення синсетів української мови.

У таблиці **Words_in_Synsets** зберігається інформація про те, які слова до яких синсетів належать.

Таблиця **Relation_dict** зберігає інформацію про типи асоціативно-семантичних зв'язків, їх скорочені назви та суть кожного типу зв'язку.

У таблиці **Relations** зберігається інформація про зв' язки між синсетами.

Пакет **Ontology_Processing,** як і пакет **Morphology_Processing,** містить набір процедур та функцій для маніпуляції з даними всередині бази даних, але між онтологічними об'єктами. Пакет також є об'єктом бази даних.

Таблицю **Users** не показано на цій діаграмі для полегшення розуміння, але в онтологічній частині бази вона теж використовується, як це помітно зі структури таблиць.

4. Перспективи використання технології

Застосування запропонованого підходу надає значні переваги при розробці онтологічних баз знань у плані підтримки високої швидкості доступу до даних, цілісності та зручності роботи з даними. Це дає можливість побудови потужних лінгвістичних онтологічних баз знань, які у свою чергу є фундаментом для створення ряду інтелектуальних прикладних лінгвістичних систем, що працюють на рівні семантики. Лінгвістичні процесори на базі семантичного аналізу гратимуть роль функціонального ядра в системах автоматичного перекладу, реферування, індексації, програм підтримки діалогу природною мовою, у природномовних інтерфейсах та в багатьох інших системах штучного інтелекту.

1. *Глибовець М. М., Остапенко О. Ю.* Алгоритм спрощення синтаксичних структур тексту природною мовою до стандартних речень // Проблеми програмування. - 2008. - № 2-3. Спец. вип. - С. 476-483.
2. *Гладун А. Я., Рогушина Ю. В.* Онтологии в корпоративных системах. Ч. II // Корпоративные системы. - 2006. - № 1.
3. http://www.w3.org/RDF/
4. http://www.daml.org/
5. http://www.w3.org/TR/owl-features/
6. http://ksl.stanford.edu/knowledge-sharing/kif/
7. http://www.cyc.com
8. *Глибовець М. М.* Семантичний пошук із поширенням активації // Наук. праці Миколаїв. держ. гуманітарного ун-ту ім. Петра Могили. Сер.: Комп'ютерні технології. - Миколаїв, 2008. - Вип. 77. - С. 197-205.

M. Glybovets, O. Marchenko, A. Nykonenko

DEVELOPING UKRAINIAN ONTOLOGY WITH ASSETS OF DBMS

Modern methods and means for building linguistic knowledge bases were considered. The way of designing associative-semantic knowledge base by means of existing DBMSs was offered. The structure and the scheme of database objects' interaction were developed on the base of the suggested approach.