О.В. Палагін, М.Г. Петренко

Архітектурно-онтологічні принципи розбудови інтелектуальних інформаційних систем

Аннотация: *В работе исследованы и разработаны концептуальные основы построения онтолого-управляемых информационных систем, главными особенностями которых являются метаонтология ЯОКМ и ориентация на аппаратные средства интерпретации информационных структур. При этом последние реализованы на современных программируемых логических интегральных схемах с использованием парадигмы гибкой архитектуры, что обеспечивает, в частности, эффективный механизм обработки индексов, которые идентифицируют лексические единицы ЕЯТ в компьютере.*
Ключевые слова: *онтология, языково-онтологическая картина мира, онтолого-управляемая информационная система.*

Однією з галузей інтелектуальних інформаційних систем (ІІС), що активно розвивається, є онтолого-керовані інформаційні системи (ОКІС), які, в свою чергу, тісно пов'язані з концептуалізацією онтологічних категорій та удосконаленням ієрархічних структур сутностей на всіх рівнях. При цьому онтологічні принципи виступають в ролі об'єднуючого механізму між науковими знаннями конкретної предметної галузі та загальними знаннями, орієнтованими перш за все на вирішення однієї з головних проблем штучного інтелекту - аналіз, синтез та розуміння природної мови комп'ютером.

Базові процедури, що представляють зміст даної проблеми, в більш широкому смислі можна виразити продукційним ланцюгом: “вхідне повідомлення → система знань → реакція”, суть якої допускає мультидисциплінарну системну інтеграцію формального логічного представлення структури та правил виведення, методів і засобів комп’ютерної лінгвістики (в тому числі теорії лексикографічних систем) та віртуальної парадигми, зокрема архітектури комп’ютерної системи з орієнтацією на сучасні електронні компоненти, ефективно підтримуючі технології реконфігурації.

Проектування будь-якої знання-орієнтованої ІС, якою зокрема є і мовно-онтологічна інформаційна система (МОІС), припускає розробку трьох незалежних, але тісно взаємозв’язаних аспектів:

- логічного представлення знань (у даному випадку лексико- і семантико-синтаксичних відношень природної мови (ПМ));
- онтології домену (мовно-онтологічна картина світу (МОКС));
- процесингу (комп'ютерної обробки).

Якщо логіка нам говорить, що що-небудь існує і надає логічні оператори маніпулювання сутностями, то онтологія, по-перше, надає словник цих сутностей, а по-друге є формалізованим представленням всіх видів сутностей - абстрактних і матеріальних, що становлять світ. Судження на ПМ, будучи переведеним у логічне представлення, вже може бути “зрозумілим” комп'ютерові та опрацьовано відповідно до конкретних потреб людини.

На рис.1 показано архітектурно-структурну організацію формальних методів та засобів обробки знань в ОКІС, акцентуючи увагу на онтологічному аспекті. Блоки логічного представлення означають, що використані в даній розробці, представляють відомі методи і засоби, вибрані нами як базові, в тій чи іншій мірі модифіковані для конкретного застосування. Наприклад, концептуальні графи є комбінацією логіки Пірса із семантичними мережами, що використовуються в комп'ютерній лінгвістиці [13]. Формат обміну знаннями (в англійській абревіатурі KIF) служить мовою обміну знаннями між гетерогенними системами баз знань і баз даних [14]. Обидві системи є міжнародними стандартами, розроблялися одночасно, в тому числі й для обробки ПМ, і мають взаємно однозначне представлення.

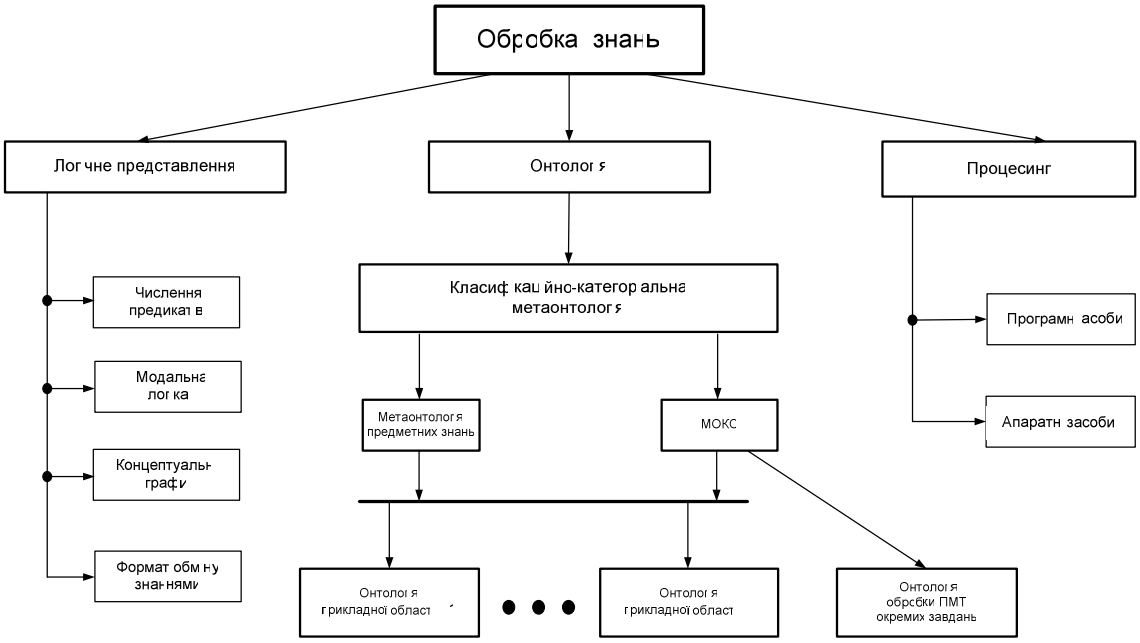

Рис.1. .Архітектурно-структурна організація формальних методів та засобів обробки знань в ОКІС

У гілці процесингу в якості програмних засобів використано реляційну систему керування базою даних (СКБД), що забезпечує зберігання і первинну обробку лексичних одиниць.

Апаратні засоби підтримки зорієнтовано на сучасні ПЛІС-технології та останні досягнення мікроелектроніки відповідно до віртуальної парадигми гнучкої архітектури, архітектури “процесор у пам'яті” та реконфігуровного процесингу. Вибір цих архітектур визначається їхньою взаємодоповнюваністю, що дозволяє досягти поставленої мети найбільш ефективним шляхом.

Основну частину матеріалу присвячено онтологічному аспектові в застосуванні до обробки ПМ.

У багатьох працях [1-10, 15, 16] підкреслюється, що для системи обробки ПМ варто будувати (і використовувати) повну онтологію лексики ПМ, перш за все онтологію лексичних засобів верхнього рівня, наприклад з [16]. “Для обробки природної мови онтологія повинна могти розмістити все, що будь-яка людина могла б сказати. Її концепти повинні покривати повний діапазон змістів слів у мові”.

Вважається, що самим складним у процесі створення загальної онтології є задача класифікації при розробці метаонтології та її найближчих нижніх рівнів. Предмет її дослідження походить від древніх філософів Геракліта та Арістотеля, середньовічних схоластиків (диски Лула) до Куайна, Канта, Лейбніца, Пірса, Хасерла, Уайтхейда та сучасних вчених, як вітчизняних - Соколовской, Широкова, Соловйової, Маторіна, так і закордонних - Гуаріно, Сова та інших.

У багатьох працях визнано, що фундаментальними принципами формування категорій є:

а) проста дихотомія, що була відома ще Геракліту та Арістотелю;

б) тріада або трихотомія, що (у смислі онтології) найбільш повно розробив Пірс і назвав її складові відповідно Первинністю, Вторинністю та Третинністю. Згодом такий розподіл формування категорій одержав назву “принцип Пірса”;

в) математичні теорії, в першу чергу комбінаційний метод Лейбніца генерування решіток, що припускає множення категорій від верхнього рівня до нижнього.

Класифікаційно-категоріальна метаонтологія, сформована на зазначених принципах, розглянута в [9].

Онтологія, що представлена блоком “МОКС”, є однією з центральних підсистем ОКІС. Різні аспекти її загальної архітектури викладено у працях [1-3,9]. Обов'язковою умовою її реалізації є формалізована комп'ютерна інтерпретація (як програмними, так і апаратними засобами). Таку онтологію іноді називають наївною картиною світу. Знання про навколишнє середовище в ній вичерпуються системою понять, сформульованих певною мовою на рівні здорового глузду, зв'язаних між собою максимально повною системою відношень, що відбивають навколишній світ з усією множиною його об'єктів та явищ, тобто являють собою лінгвістичну проекцію буття людини, у якій зафіксовано досвід взаємодії з навколишньою дійсністю. МОКС - складова частина прагматичної моделі мовної свідомості, що є ключовим компонентом сучасних інтелектуальних ІС із природномовним представленням, обробкою та актуалізацією знань.

МОКС ми визначаємо як відкриту, експліцитно задану на лексико-смисловому континуумі лексикографічну систему, в якій сукупність категоріальних понять високого рівня формально обґрунтовано та впорядковано в складну ієрархічну структуру за основними типами лексико-семантичних відношень.

Онтологія, як формальний опис загальноприйнятої лексики, представляється стандартною формулою

$$O = <X, P, F>,$$

де X - поняття, характеристики, ролі та атрибути (або контент слова), виражені лексичними засобами ПМ, перш за все такими повнозначними частинами мови, як іменник, дієслово, прикметник та прислівник. В логіці вони представляються, як правило, одномісними предикатами $P(x)$;

P - повна система відношень, така як $P(x,y)$ або $P(x,y,z)$ (за твердженням Пірса, відношення з валентностями чотири і більше можна представити композицією дво- і трьохвалентних відношень);

F - множина функцій інтерпретації, заданих на X і\чи P.

Графічно, МОКС представляє деякий гіперграф, що є результатом склеювання ациклічних орієнтованих графів лексичних одиниць для кожної повнозначної частини мови.

Службові частини мови враховуються на етапі зняття багатозначності та логічного представлення вихідного природномовного тексту (ПМТ).

На рис.2 представлена інформаційна модель ОКІС. Два основних блоки на цьому рисунку становлять у класичному розумінні лінгвістичний процесор.

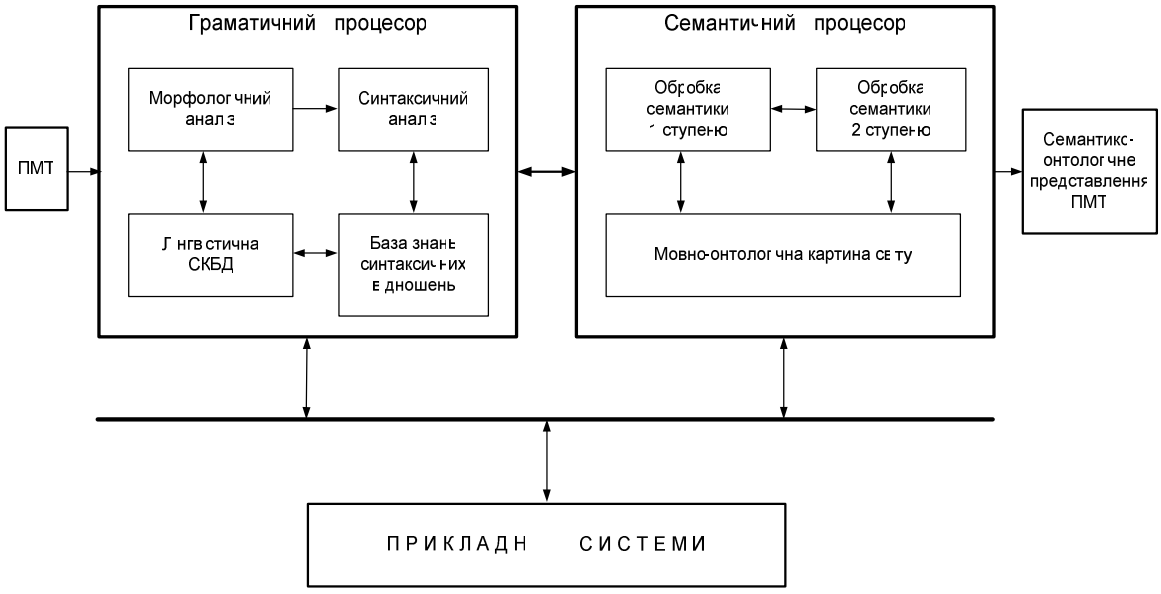

Рис.2. Інформаційна модель мовно-онтологічної онтолого-керованої інформаційної системи

Практична цінність одержуваних результатів при обробці ПМТ, в основному, залежить від повноти інтерпретаційних моделей семантичних структур ПМТ та їхнього формального представлення. Під повнотою ми розуміємо включення в модель як складової семантики першого ступеню (або об'єктової складової), так і складової семантики другого ступеню (або акторної складової). Такий розподіл семантики добре узгоджується як із онтологічною ієрархією концептуальних категорій, так і зі складністю виконання обчислювальних процедур при комп'ютерній обробці ПМТ.

З погляду лінгвістики, семантична складова першого ступеню описується на рівні граматики окремих частин мови, в той час як складова другого ступеню вже описується синтаксичними конструкціями таких одиниць синтаксису, як речення, абзац, параграф, розділ і текст. З погляду математичної логіки, якщо перший ступінь можна описати (досить умовно) численням висловлювань, то другий ступінь повинний описуватися численням предикатів з квантифікованими змінними.

Найбільшої повноти (і відповідно найбільшого ступеню складності) набувають моделі, що описують ПМТ в цілому. Такі моделі описують, зокрема, деякий сценарій (як вищу категорію, що описує явище, взаємовідношення об’єктів, що перебувають у постійному русі), що відображає зміст ПМТ. У свою чергу, як ПМТ ділиться на синтаксичні одиниці, так і загальний сценарій розділяється на окремі сценарії, ситуації та елементарні ситуації.

Описана істотна різниця між об'єктовою та акторною складовими семантики, а також морфолого-синтаксичним аналізом, зокрема в складності їхніх інтерпретаційних моделей, обумовила виділення для моделювання й інтерпретації семантики окремого функціонального модуля - семантичного процесора (СП). Морфологічний і синтаксичний аналіз при цьому виконується граматичним процесором, а точніше окремими його блоками морфологічного та синтаксичного аналізу. Він містить також лінгвістичну СКБД реляційного типу та синтаксичну базу знань.

Лінгвістична СКБД включає окремі таблиці для всіх повнозначних частин мови. До кожної лексеми в таблиці приєднується, крім традиційних морфологічних характеристик, набори синтаксичних і семантичних харак-

теристик [6,7]. Крім того, існує окрема таблиця відмінкових закінчень для формування словоформ лексеми. Всі лексичні одиниці в таблицях відповідним чином проіндексовані та мають однакове інтерпретаційне значення як для граматичного, так і семантичного процесорів. Більш докладно про індексування лексем у МОКС описано в [3]. У синтаксичній базі знань представлено інтерпретаційну модель синтаксичних відношень ПМ, відповідно до якої виконується синтаксичний аналіз вихідного ПМТ. Така складна архітектура граматичного процесора виправдана насамперед тим, щоб ефективно вирішити проблему зняття граматичної та лексичної неоднозначності. Для цього в ньому використано морфологічні, лексичні, синтаксичні та семантичні методи, а остаточне зняття неоднозначностей виконується СП [11].

Основним призначенням СП є побудова формалізованого опису вихідного ПМТ і його відображення в онтологічному дереві МОКС. Інакше кажучи, головним завданням СП є відображення структури тексту на онтологічну структуру МОКС і фіксація семантико-синтаксичної структури окремих речень і текстових фрагментів у вигляді відповідних сукупностей індексів, що зв'язують відношеннями повну множину лексем та їхніх значень, представлених у МОКС.

Структура інформаційних зв'язків між процесорами та прикладною системою (рис.2) універсальна, що дозволяє передавати інформацію як “знизу-вверх”, так і “зверху-вниз”.

В якості прикладної системи може слугувати розроблювальна нами онтолого-керована пошукова система [12]. Одним з її призначень є пошук документів, їхня обробка з урахуванням “фонових” знань і часткова класифікація. При цьому МОКС виконує функції зняття неоднозначностей у документах та їх остаточна класифікація. Така система по своїм функціональним характеристикам близька до російської технології класифікації Rubryx і проекту “Інтелектуальна пошукова машина” [10].

Близькими за призначенням є системи обслуговування множинного потоку документів, зокрема реферування із використанням процедури узагальнення. Вона є невід'ємною частиною ОКІС більш широкого призначення.

На закінчення можна привести приклади закордонних знання-орієнтованих ІС різного призначення, що використовують онтологію лексичних засобів верхнього рівня для англійської мови, аналогом якої є МОКС для української мови: Pangloss, Mikrokosmos, Revised Upper Model, Cyc та інші.

У роботі досліджено та розроблено концептуальні основи побудови онтолого-керованих інформаційних систем, головними особливостями яких є метаонтологія МОКС та апаратні засоби інтерпретації інформаційних структур. При цьому останні реалізовані на сучасних програмовних інтегральних логічних схемах з використанням парадигми гнучкої архітектури, що забезпечує, зокрема, ефективний механізм обробки індексів, що ідентифікують лексичні одиниці ПМТ у комп'ютері.

Література

1. Палагин А.В. Организация и функции "языковой" картины мира в смысловой интерпретации ЕЯ - сообщений. //Information Theories and Application. – 2000. – Vol. 7, №4. С.155-163.
2. Палагин А.В., Яковлев Ю.С. Системная интеграция средств компьютерной техники. – Винница: «УНІВЕРСУМ-Вінниця», 2005. – 680 с.
3. Палагин А.В. Архитектура онтологоуправляемых компьютерных систем. - Кибернетика и системный анализ. - №2, 2006. – С.111-124.
4. Широков В.А. Феноменологія лексикографічних систем. - К.: Наукова думка, 2004. - 327с.
5. Маторин С.И. Системологическое исследование структуры системы категорий. //НТИ. Сер.2. 1997. №3. С3-7.

6. Замаруева И.В. Об одном подходе к компьютерному моделированию процесса понимания естественно-языковых текстов. – Труды VI Межд. конф. "ЗНАНИЕ-ДИАЛОГ-РЕШЕНИЕ", KDS-97, Ялта, 15-20 сентября, 1997г., С.241-248.
7. Апресян Ю.Д. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. – 287 с.
8. Соколовская Ж.П. «Картина мира» в значения слов. – Симферополь. «Таврия». 1993.
9. Палагін О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу. - Математичні машини й системи, 2006, №3. - С.91-104.
10. Поляков В.Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации. - http://virtualcjglab.cs.msu.su/html/polyak.html
11. Петренко М. Г. Особливості розробки знання-орієнтованого лінгвістичного процесора. – Комп'ютерні засоби, мережі та системи. - 2006, №5. – С.18-22.
12. Севрук О.О., Петренко М.Г. Знання-орієнтована пошукова система на основі мовно-онтологічної картини світу //Тези доповідей XIII міжнародної конференції з автоматичного управління "Автоматика-2006". – м. Вінниця, 25-28 вересня. – 2006. – С.413.
13. NCITS T2 (1998) *Conceptual Graphs, A Presentation Language for Knowledge in Conceptual Schemas,* Working draft of proposed American national standard, Document No. X3T2/96-008.
14. NCITS T2 (1998) *Knowledge Interchange Format,* Working draft of proposed American national standard, document (or available at http://logic.stanford.edu/kif/dpans.html).
15. Guarino N. Some Ontological Principles for Disigning Upper Level Lexical Resourses. //Proceedings of First International Conference on Language Resources and Evaluation, Granada, Spain, 28-30 May.
16. John F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000.