КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

УДК 681.004

М.Г. ПЕТРЕНКО

ОСОБЛИВОСТІ РОЗРОБКИ ЗНАННЯ-ОРІЄНТОВАНОГО ЛІНГВІСТИЧНОГО ПРОЦЕСОРА

На сучасному етапі одним із перспективних напрямків вдосконалення інтелектуальних інформаційних систем і технологій є побудова знання-орієнтованих, у тому числі онтолого-керованих систем, функціонування яких ґрунтується на автоматизації процесу формалізації змісту природно-мовних текстів (ПМТ) з наступною обробкою формалізованого відображення цього змісту логіко-семантичними методами для розв'язання задач користувачів, орієнтованих на інтелектуальний аналіз. Основною компонентою таких інформаційних систем є підсистема лінгвістичної обробки ПМТ, загальною задачею якої є розпізнавання і добування знань, що містяться в ПМТ, та їх формалізація – з одного боку, і синтез по формалізованому представленню цих знань їх опису природною мовою – з іншого [1].

Таким чином, особливістю лінгвістичної обробки в таких підсистемах є підпорядкування всіх її етапів формуванню елементів формалізованого представлення знань. Традиційно, задача розуміння ПМТ підрозділяється на три етапи: аналіз, інтерпретація (або відображення) і синтез.

У роботі розглядається тільки перший етап – етап аналізу і частково – етап інтерпретації. На етапі аналізу виокремлюється опис сутностей, відбитих у вхідному тексті, виявляються властивості цих сутностей і відношення між ними, представлені у вигляді формальних моделей (наприклад, предикатні моделі).

Існує декілька підходів до моделювання алгоритмів аналізу природно-мовних текстів.

N.G. Petrenko

FEATURES OF DEVELOPMENT KNOWLEDGE-ORIENTED OF THE LINGUISTIC PROCESSOR

The linguistic component knowledge-focused of information systems which functioning is based on automation of process of formalization of sense of naturally-language texts is considered. On the basis of the analysis of known approaches the approach of construction of the linguistic processor which basis are knowledge bases of different levels of the linguistic analysis is offered knowledge-focused, and as a technological basis serves relational control system of a database.

Розглянуто лінгвістичну компоненту знання-орієнтованих інформаційних систем, функціонування якої ґрунтується на автоматизації процесу формалізації змісту природно-мовних текстів. На основі аналізу відомих підходів запропоновано знання – орієнтований підхід до розбудови лінгвістичного процесора, підґрунтям якого є низка баз знань різних рівнів лінгвістичного аналізу, а технологічною основою слугує реляційна СКБД.

Класичними є підходи на основі продукційних та декларативних моделей. В основі першого підходу лежить поняття морфа та його застосування при формуванні відповідних лінгвістичних характеристик моделювання [2, 3]. При другому підході для формування формалізованого опису використовується повна множина лексем та їх словоформ певної природної мови. Пропонуємо комбінований підхід, який використовує переваги вищезгаданих підходів і орієнтується на програмні моделі інтерпретації знань, що описують лінгвістичні складові та відношення між ними, присутні у ПМТ.

Загальні концептуальні положення, на яких ґрунтуються всі підходи (але використовуються по-різному), в тому числі і наш [1], наведені нижче:

- вхідний ПМТ є зв'язний текст, або дискурс;
- зв'язність тексту забезпечується графемними засобами оформлення тексту (наприклад, відношеннями взаємозв'язку між заголовками фрагментів тексту і змістом абзаців), лінгвістичними засобами (граматичними узгодженнями, анафоричними посиланнями та ін.) і екстралінгвістичними (наприклад, часові і причинно-наслідкові зв'язки, існуючі в проблемній області);
- всі ці засоби є інструментом кодування знань про світ;
- як елементи реального чи абстрактного світу виступають його об'єкти, представлені у вигляді понять природної мови;
- членування знань про світ на об'єкти може бути різним і залежить від цілей дослідження. При розгляді загальнонаукових, загальновживаних знань доцільно їх представляти як базу метазнань, структурованих у відповідності з ієрархією мовної картини світу (МКС) [4].

На рисунку показана загальна блок-схема знання-орієнтованого лінгвістичного процесора (ЛП) у розрізі як всіх етапів лінгвістичного аналізу (вертикальні блоки, окреслені пунктирними лініями), так і рівнів вихідних даних, обробки та результатів (горизонтальні блоки, окреслені неперервними лініями).

Особливості знання-орієнтованого ЛП полягають у побудові баз знань графемних, морфологічних, синтаксичних та семантичних одиниць, зв'язаних між собою відповідними відношеннями, та їх інтерпретаціями на всіх етапах обробки вхідного ПМТ. При цьому текст розглядається як об'єкт різного рівня аналізу: як знакова система, як граматична (морфологічна, синтаксична) система, як система знань про світ (лексична семантика). Кожен рівень має свої особливості, свої засоби виразу і припускає наявність специфічних методів обробки. Виходячи з цього, аналіз постає як багаторівневий процес, кожен з етапів якого формує необхідну ознакову інформацію для реалізації наступного етапу. На виході одержуємо семантико-онтологічне відображення вхідного ПМТ в МКС, побудоване реляційною системою керування базами даних (СКБД) і знань.

Як початковий етап лінгвістичного аналізу використовується графемний аналіз. При цьому текст розглядається як різновидність знакової системи. Задачею даного етапу аналізу є дослідження властивостей одиниць мови і правил їх поєднання в аспекті знакової природи. В існуючих моделях графемний аналіз реалізовано тільки як засіб розв'язання протиріч аналізу текстових одиниць, що

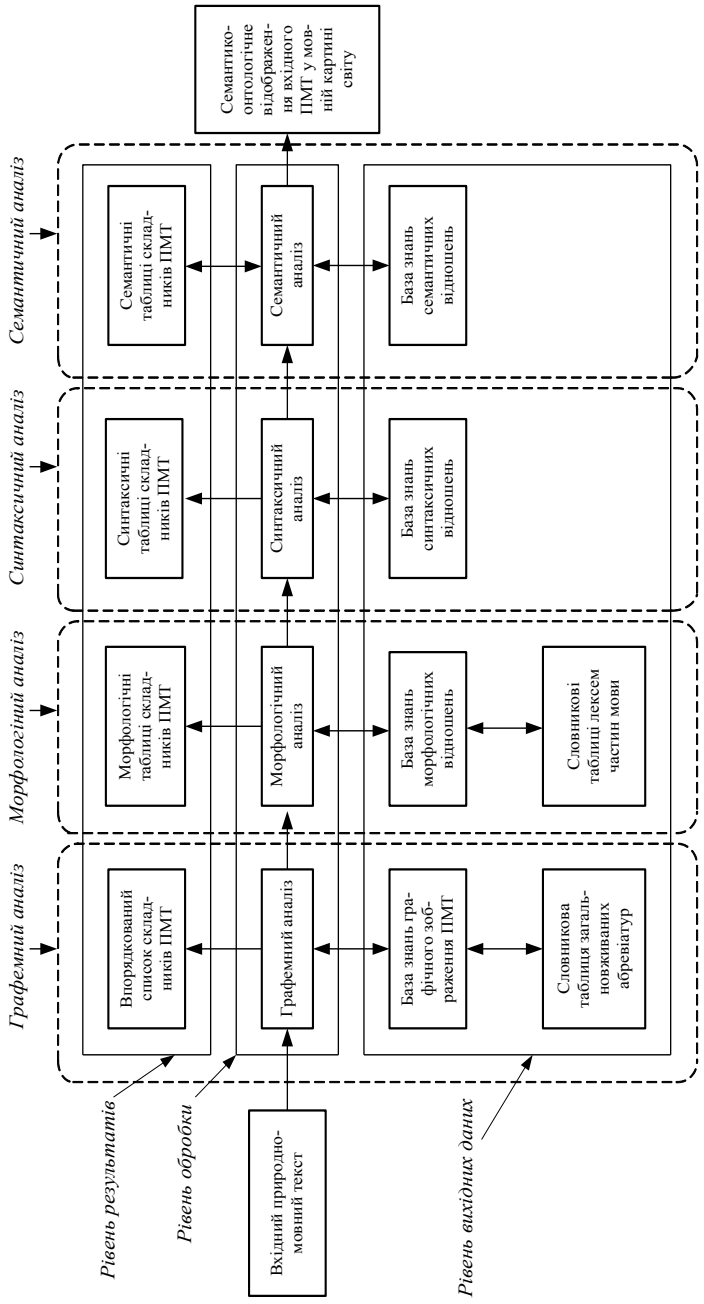

РИСУНОК. Блок-схема знання-орієнтованого ЛП

надходять на обробку морфологічного аналізатора і окреслюється відділенням абревіатур, скорочень та деяких інших класів лексем. Це, в свою чергу, породжує протиріччя на етапі семантичного аналізу тексту, що призводить до помилок семантичного характеру, уникати яких користувачу пропонується самому,

вручну маркірувати текст.

Призначення цього етапу – побудова моделі графемної структури вхідного тексту, в якій виділені і зв'язані відношеннями (де це можливо) такі змістовні одиниці тексту як фрагмент, речення та лексема. Наприклад, для лексем: виділяються класи лексем, які відрізняються своєю графемною структурою і виконують різні функції в тексті. Крім того, аналізуються закономірності сполучуваності деяких лексичних одиниць, які вже на цьому етапі дозволяють об'єднувати декілька лексем в одну тому, що вони виконують в тексті єдину функцію.

Внаслідок графемного аналізу (точніше, інтерпретації бази знань графічного зображення ПМТ з вхідним текстом) текст перетворюється у впорядкований список складників ПМТ. Такий список у спрощеному вигляді представлено нижче.

ТЕКСТ = ЗАГОЛОВОК | {АБЗАЦ1.1, …, АБЗАЦ1.N};

АБЗАЦ1.1 = {РЕЧЕННЯ1.1, …, РЕЧЕННЯ1.M};

……

АБЗАЦ1.N = {РЕЧЕННЯN.1, …, РЕЧЕННЯN.P};

РЕЧЕННЯ1.1 = $\{(СЛ|СФ)_1, …, (СЛ|СФ)_i\}$;

……

РЕЧЕННЯN.P = $\{(СЛ|СФ)_1, …, (СЛ|СФ)_j\}$, де i = 1,I, j = 1,J – кількість спеціальних лексем чи словоформ відповідно в реченнях 1.1 та N.P.

Наступним етапом лінгвістичного аналізу є морфологічний аналіз, мета якого побудувати морфологічні таблиці для кожної словоформи, спеціальних лексем, речень, абзаців і тексту в цілому з урахуванням всіх можливих багатозначностей. Інструментальними засобами морфологічного аналізу є реляційна СКБД, яка як вихідні дані використовує словникові таблиці лексем всіх частин мови. Програмна модель інтерпретації бази знань морфологічних відношень використовує SQL-запити СКБД і вирішує завдання побудови словозміни для змінних частин мови. При цьому морфологічні таблиці включають як граматичні категорії складників вхідного ПМТ, так і їх характеристики граматичної семантики.

Етап синтаксичного аналізу як вихідні дані використовує базу знань синтаксичних відношень, в якій містяться описи синтаксичних конструкцій, впорядковані за правилами синтаксису природної мови. Така структура бази знань подібна до фреймових моделей опису синтаксису ПМТ. Але вона однаково легко піддається інтерпретації як на програмному, так і на апаратному рівні, притому для фреймових структур відомі тільки програмні реалізації. На даному етапі вирішується більшість проблем граматичної багатозначності, так як при інтерпретації відповідної бази знань враховуються правила контекстної сполучуваності граматичних одиниць природної мови. Результатом синтаксичного аналізу є побудовані синтаксичні таблиці для всіх синтаксичних одиниць, що входять у текст, поданий для опрацювання. Якщо на цьому етапі ще не вдалося повністю позбавитись від багатозначності, то відповідна таблиця "помічена" спеціальним маркером, а її рядки містять альтернативні записи можливих синтаксичних конструкцій.

Семантичний аналіз поділяється на два етапи. Задачею першого етапу є усунення синтаксичної багатозначності (у першу чергу омонімії, якщо вона має місце), формування понять, відношень та їх характеристик у межах одного речення. Крім того, на цьому етапі проходить заміщення анафоричних зв'язків (коли замість поняття підставляється займенник), узагальнення понять і відношень, пропусків понять і відношень, що притаманно природній мові. На другому етапі семантичного аналізу об'єктами розпізнавання є всі речення, в тому числі і речення-заголовки. Потім всі фрагменти тексту об'єднуються в єдину логіко-семантичну структуру. При цьому обробка полягає в узагальненні та уніфікації понять, відношень та їх характеристик.

Вхідними даними семантичного аналізу є синтаксичні таблиці, а також бази знань семантичних відношень, які використовують онтологію (складні ієрархічні структури) мовної картини світу [5]. Програмна інтерпретація бази знань в середовищі реляційної СКБД виконує перетворення синтаксичних відношень в семантичні (наприклад, підмет-присудок → об'єкт-дія, ознака підмета → характеристика об'єкта, ознака присудка → характеристика дії і т. ін.). Онтологічна структура МКС використовує такі відношення для категорій верхнього рівня як рід-вид, частина-ціле, елемент-клас; для понять нижніх рівнів – класи семантичних відношень (класифікації, ознакові, кількісні, порівняння, належності, часові, просторові, каузальні, інструментальні, інформаційні, порядкові, модальні, модифікатори, квантифікатори та ін.). Ці відношення містять знання про загальну картину світу на рівні енциклопедичних знань.

Результат лінгвістичного аналізу – побудова семантико-онтологічного відображення вхідного ПМТ в МКС. По суті таке відображення є вхідними даними для семантичного процесора, який на основі формалізованих перетворень конструює деяке індексне поле, комірками якого є семантичні таблиці, поля і записи яких певним чином представляють "динамічну" семантику, або сукупність ситуацій, сценарій, імпліцитно присутній в поданому на опрацювання природномовному тексті.

1. *Замаруева И.В.* Об одном подходе к компьютерному моделированию процесса понимания естественно-языковых текстов // Тр. VI Междунар. конф. "ЗНАНИЕ-ДИАЛОГ-РЕШЕНИЕ", KDS-97, Ялта, 15-20 сентября, 1997. – С. 241–248.
2. *Апресян Ю.Д.* Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992. – 287 с.
3. *Гладун В.П.* Процессы формирования новых знаний. – София: СД "Педагог 6", 1994. – 192 с.
4. *Палагин А.В., Яковлев Ю.С.* Системная интеграция средств компьютерной техники. – Винница: «УНІВЕРСУМ-Вінниця», 2005. – 680 с.
5. *Палагин А.В.* Организация и функции "языковой" картины мира в смысловой интерпретации ЕЯ-сообщений // Inform. Theories and Application. – 2000. – **7**, N 4. – С. 155–163.

Отримано 13.02.2006