КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

A.V. Palagin, S.Ju. Svitla, M.G. Petrenko, V.Ju. Velychko

ABOUT ONE APPROACH TO ANALYSIS AND UNDERSTANDING OF THE NATURAL LANGUAGE OBJECTS

Information model of linguistic analysis in LOIS is considered, combined approach to recognition of the syntactic and semantic relations, threefold ambiguity analysis and algorithm of the anafora connections correlation in natural language text are offered.

Рассмотрена информационная модель лингвистического анализа в ЯОИС. Предложен комбинированный подход к распознанию синтактико-семантических отношений, тройной анализ неоднозначностей и алгоритм соотношения анафорических связей в естественноязыковом тексте.

Розглянуто інформаційну модель лінгвістичного аналізу в МОІС. Запропоновано комбінований підхід до розпізнавання синтактико-семантичних відношень, потрійний аналіз неоднозначностей та алгоритм співвіднесення анафоричних зв'язків у природномовному тексті.

УДК 004.318

О.В. ПАЛАГІН, С.Ю. СВІТЛА, М.Г. ПЕТРЕНКО, В.Ю. ВЕЛИЧКО

ПРО ОДИН ПІДХІД ДО АНАЛІЗУ ТА РОЗУМІННЯ ПРИРОДНОМОВНИХ ОБ'ЄКТІВ

Вступ. Ураховуючи випереджаючі (в порівнянні з іншими джерелами) темпи приросту обсягів інформації, що міститься в природномовних об'єктах (ПМО), включаючи мережу Інтернет, актуальність наукових досліджень в області добування знань з ПМО зростає. Зокрема, це стосується сфери знання-орієнтованих інформаційних систем обробки текстової інформації. Одним з підкласів таких систем є онтолого-керовані інформаційні системи (ОКІС) обробки знань, що містяться в ПМО. ОКІС дозволяють як "уявний образ деякої частини реального світу" (сукупність знань про деяку предметну область (ПдО)) використовувати лінгвістичний корпус науково-технічних текстів (ЛКТ), отриманих як із мережі Інтернет, так і з монографій, науково-технічних документів тощо.

У загальному випадку ОКІС обробки знань з ПМО складається з двох інтегрованих ІС: мовно-онтологічної інформаційної системи (МОІС) для "внутрішньомовної" обробки текстової інформації та ОКІС ПдО для "машинної" обробки предметних знань. МОІС обробки текстової інформації на основі мовних знань описана в [1]. Перехід між зазначеними "внутрішньомовною" та "машинною" сферами обробки виконується при реалізації відповідного алгоритму, що використовує базу мовних знань (в основі якої лежить мовно-онтологічна картина світу (МОКС)) і базу знань заданої ПдО. Ланцюжок інформаційних технологій "*Комп'ютерна обробка природномовних текс-*

тів → Представлення знань → Комп'ютерна обробка знань" є реалізація базових процедур аналізу, синтезу та розуміння природної мови комп'ютером, які в більш загальному розумінні можна виразити продукційним ланцюжком *вхідне_повідомлення → система_знань → реакція*.

Суть цього ланцюжка визначається міждисциплінарною системною інтеграцією лінгвістичних та предметних знань, що взагалі, представляє нову інформаційну технологію, яка знаходиться в стадії становлення та інтенсивного розвитку досліджень. У даній роботі розглядається (в основному) задача побудови природномовних лінгвістичних моделей та створення на їх основі ефективних лінгвістичних процесорів (ЛП). Її вирішення лежить не стільки в області побудови повних описів природної мови, скільки в області концептуального осмислення підходу до побудови лінгвістичної моделі як невід'ємної частини системи всіх учасників обробки текстів. Одним з таких підходів може бути чітке базування моделі на прагматиці системи, що об'єднує усіх її учасників навколо цільової обробки ПМО. Під учасниками обробки ПМО мається на увазі всі ресурси та суб'єкти, які залучаються, включаючи ПМО, що обробляється, користувача, нелінгвістичні блоки ІС, проблемну область, контекст і т.п.

Постановка задачі. Процес розпізнавання та добування знань з ПМО базується на моделюванні інтелектуальних функцій людини, а саме: на комп'ютерному моделюванні процесу розуміння людиною ПМО. При цьому термін *розуміння* визначається через такі критерії: вміння переказати "своїми" словами, тобто іншими (лексичними, синтаксичними) засобами передати зміст вхідного тексту, вміння відповісти на запитання щодо певного тексту. Процедура розпізнавання базується на засобах формалізації (тобто розробки онтологічних моделей) знань про певну мову та знань про певну ПдО. Оскільки процедури розпізнавання та розуміння є базовими при лінгвістичній обробці ПМО, розглянемо їх більш детально з методологічної точки зору.

В існуючих ІС виокремлюють п'ять основних рівнів розуміння ПМО [2].

Перший рівень характеризується схемою, яка показує, що будь-які відповіді на запитання система формує тільки на основі прямого змісту, виведеного із тексту. В лінгвістичному процесорі виконується морфологічний, синтаксичний та семантичний аналіз тексту і запитань, що належать йому. На виході ЛП отримуємо внутрішнє представлення тексту та запитань, з якими може працювати блок виведення. Використовуючи спеціальні процедури, цей блок формує відповіді. Іншими словами, вже розуміння на першому рівні потребує від ІС певних засобів представлення даних і виведення на цих даних.

Другий рівень. На цьому рівні додаються засоби логічного виведення, засновані на інформації, що міститься в тексті. Це різноманітні логіки тексту (часова, просторова, каузальна та ін.), які здатні породжувати інформацію, явно відсутню в тексті. Архітектура ІС, за допомогою якої може бути реалізований другий рівень розуміння повинна мати додаткову базу знань, в якій зберігаються закономірності, що відносяться до часової структури подій, можливої їх просторової організації, каузальної залежності й т. п. Логічний блок – всі необхідні засоби для роботи з некласичними логіками.

Третій рівень. До засобів другого рівня додаються правила поповнення тексту знаннями системи про середовище. Ці знання в ІС, як правило, носять логічний характер і фіксуються у вигляді сценаріїв або процедур іншого типу. Архітектура ІС, в якій реалізується розуміння третього рівня, зовнішньо не відрізняється від архітектури ІС другого рівня. Однак у логічному блоці мають бути враховані засоби не тільки для чисто дедуктивного виведення, а й для виведення за сценаріями.

Три перераховані рівні розуміння повністю чи частково реалізовані практично у всіх діючих ІС.

Четвертий рівень. На цьому рівні відбувається зміна вмісту бази знань. Вона доповнюється фактами, відомими системі, що вміщуються у тих текстах, які введені в систему. Різні ІС відрізняються одна від одної характером правил породження фактів із знань. Наприклад, в ІС, призначених для експертизи в області фармакології, ці правила спираються на методи індуктивного виведення та розпізнавання образів. Правила можуть бути засновані на принципах ймовірностей, розмитих виведень і т.п. Але у всіх випадках база знань виявляється апріорно неповною і в таких ІС виникають труднощі з пошуком відповідей на запити. Зокрема, в базах знань стає необхідним немонотонне виведення.

П'ятий рівень. На цьому рівні відбувається породження метафоричного знання. Правило породження знань метафоричного рівня, що використовуються для цих цілей, представляють собою спеціальні процедури, що спираються на виведення за аналогією та асоціацією. Відомі в теперішній час схеми виведення за аналогією використовують, як правило, діаграму Лейбниця, яка відображає тільки частковий випадок суджень за аналогією. Іще менш розроблені схеми асоціативних суджень [2].

Існують й інші інтерпретації феномену розуміння. Можливо, наприклад, оцінювати рівень розуміння за здатністю системи до пояснення отриманого результату. Тут можливий не тільки рівень *пояснення*, коли система пояснює, що вона зробила, наприклад, на основі введеного до неї тексту, але і рівень *обґрунтування* (аргументації), коли система обґрунтовує свій результат, показуючи, що він не суперечить тій системі знань і даних, якими вона володіє. На відміну від пояснення обґрунтування завжди пов'язане із сумою фактів і знань, які визначаються теперішнім моментом існування системи. І введений для розуміння текст в одних станах може бути сприйнятий системою як істинний, а в інших – як хибний. Існуючі ІС типу експертних систем, як правило, здатні давати пояснення і лише частково обґрунтування [2].

Особливості аналізу ПМО визначаються спрямованістю на формування понятєвої структури, тобто на автоматичне добування знань з текстів та їх прагматичну інтерпретацію у термінах прикладної задачі. При цьому текст розглядається як об'єкт різних рівнів аналізу: як знакова система, як граматична система і як система знань про світ (предметну область). Кожний рівень має свої особливості, свої засоби вираження і, отже, припускає наявність специфічних методів обробки.

На основі виконаного аналізу моделей та загальних принципів комп'ютерної обробки ПМО на рис. 1 синтезовано структурно-логічну схему етапів лінгвістичного аналізу та прийнято наступні скорочення:

- МОКС – мовно-онтологічна картина світу;
- ПМ – природна мова;
- ПМО – природномовний об'єкт.

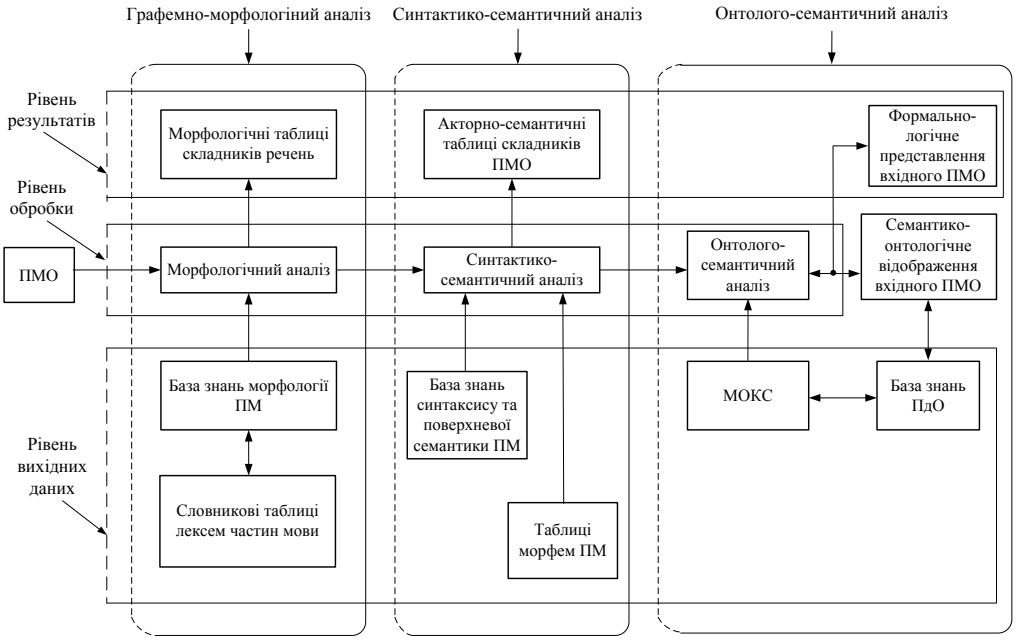

РИС. 1. Структурно-логічна схема етапів лінгвістичного аналізу

На схемі база знань ПМ складається з МОКС та баз знань етапів графемного, морфологічного та синтаксичного аналізу. База знань частини реального світу, до якого належить лінгвістичний корпус ПМО, відображена в блоці "База знань ПдО".

Метою лінгвістичного аналізу є побудова семантико-онтологічної структури та форм логічного представлення ПМО.

Суть комбінованого підходу. При аналізі ПМО одна з основних операцій – розпізнавання синтаксичних та семантичних відношень, що зв'язують слова в тексті. Розпізнавання відношень реалізується на основі їх описів (моделей) [3, 4]. Такого роду моделі обов'язково присутні у всіх методах аналізу, хоча і не завжди явно. Перший критерій для розділення методів аналізу на класи виділяється з урахуванням того, наскільки великі фрагменти реального світу відображені в моделі, що використовується. В якості другого критерію класифікації вибирається мова моделі, що використовується для розпізнавання відношень.

У більшості методів аналізу процесу розпізнавання відношень передує переклад вихідного природномовного представлення розпізнаних об'єктів (відношень) у мову категорій традиційної граматики (число, рід, відмінок, час тощо). При цьому правила розпізнавання синтаксичних та семантичних відношень оперують граматичними описами слів. Тим часом, перехід до граматичних описів не являється обов'язковою умовою для виконання аналізу ПМО. Інформація, необхідна для розпізнавання синтаксичних та семантичних відношень, міститься безпосередньо в тексті. Тому правомірний інший підхід, заснований на використанні відповідностей між відношеннями та засобами їх вираження в ПМО. В роботі [3] ці два підходи названо відповідно граматичним підходом та підходом безпосереднього розпізнавання.

Цілком зрозуміло, що граматичний підхід і підхід безпосереднього розпізнавання мають свої недоліки та переваги. Зокрема, при першому підході – громіздкість процесів обробки та великий обсяг інформації; при другому підході – трудомісткість розробки словників морфів (переважно лінгвістів) та значний обсяг нерозпізнаних неоднозначностей (при визначенні морфологічних характеристик). Недоліки першого підходу останнім часом нівелюються значними досягненнями в області мікроелектроніки при створенні засобів комп'ютерної техніки (значно збільшено обсяг оперативної та дискової пам'яті) та принциповою можливістю реалізувати трудомісткі процедури аналізу ПМО за допомогою апаратних засобів їх підтримки (що на один-два порядки збільшить потужність обчислень). Недоліки другого підходу усунути принципово неможливо. Тому в роботі вибрано комбінований підхід до аналізу ПМО, який зменшує обсяг нерозпізнаних неоднозначностей до мінімального.

Морфологічний аналіз. Вхідними даними процедури розпізнавання є графемна структура тексту, отримана на попередньому етапі, база знань морфології ПМ та лексикографічна база даних (ЛБД). Остання включає словники лексем, словозмінну і словотвірну моделі вхідної мови, окремі таблиці для всіх частин мови. До кожної лексеми в таблиці приєднується, окрім традиційних морфологічних характеристик, набори синтаксичних і семантичних ознак [5, 6].

Сутність процедури морфологічного аналізу полягає у приписуванні кожній мовній лексемі вхідного ПМО відповідної змістової інформації та їх структуризацію у морфологічній таблиці (МТ). Для тих словоформ, яким в тексті притаманні різного роду неоднозначності, в МТ вказуються всі словоформи омографи з відповідними граматичними характеристиками. Така інформація включає: лексико-граматичні класи та відповідні цим класам граматичні характеристики (наприклад, для іменників – це рід, число, відмінок) та вищезгадані деякі синтаксичні і семантичні ознаки. Вона передається до етапу синтаксичного аналізу.

Синтактико-семантичний аналіз. Кінцевим завданням блоку синтаксичного аналізу є представлення кожного речення заданого природномовного тексту у вигляді синтаксичного дерева (лексеми речення з синтактико-семантичними відношеннями між ними).

Зв'язування слів у реченні відбувається поступово, від словосполучення до групи зв'язаних слів і, зрештою, до об'єднання всіх груп у реченні в одну структуру. Для встановлення зв'язку між окремими словами використовуються природномовні засоби вираження семантичних та синтаксичних відношень. У флективних мовах такими засобами є змінні частини повнозначних слів та службові слова. Такі сегменти словосполучення, які кодують відношення між повнозначними словами, називаються синтаксичними визначниками [4]. Оскільки одному синтаксичному визначнику може відповідати декілька синтаксичних відношень, для однозначності визначення зв'язків між словами вводиться поняття кореляторів [4], які додатково враховують семантичні ознаки слів у словосполученні.

Вихідними даними для блоку синтаксичного аналізу є:

- результат попередніх етапів аналізу ПМО (графемного і морфологічного);
- словник основ (містить основи слів та їх семантичні ознаки);
- список всіх можливих флексій слів;
- база даних з визначниками (містить синтаксичні визначники та списки кореляторів для кожного з них);
- база даних з кореляторами (кожен корелятор складається з назви відношення та списку пар семантичних ознак, між якими це відношення може існувати).

Далі описані основні етапи роботи блоку синтаксичного аналізу природномовного тексту.

Перший етап. Використовуючи словник основ та список флексій, у кожному слові речення виділяється його незмінна складова (основа) та флективна. Проводиться класифікація слів за семантичними ознаками відповідних основ у словнику. При цьому виникає проблема можливої неоднозначності виділення основи слова та визначення його семантичної ознаки. Одним з шляхів її вирішення є врахування характеристик слів, що стоять поруч у реченні, та розширення словника основ додатковими характеристиками. Наприклад, дієслово "мати" омонімічне іменнику "мати", але якщо на етапі морфологічного аналізу визначено що попереду в реченні знаходиться прислівник, тоді "мати" є дієсловом, а якщо попереду знаходиться прикметник, "мати" – іменник.

Незмінна та флективна частини слова та його семантична ознака можуть бути отримані й на попередніх етапах аналізу (етап морфологічного аналізу).

Другий етап. Зв'язування слів у реченні доцільно починати зі словосполучення, що визначає головне відношення (відношення між підметом і присудком) у цьому реченні. Підмет та присудок подалі будемо називати ядром речення. У випадку, коли ядро визначити неможливо, речення аналізується зліва направо, починаючи з перших повнозначних слів.

Для вибраного словосполучення формується синтаксичний визначник, який складається зі службових слів та флективних частин повнозначних слів словосполучення. Наприклад, для виразу «права та свободи» таким визначником буде конструкція типу «-а та -и». Якщо сформований визначник існує у базі даних визначників, йому буде відповідати список кореляторів. Тоді, враховуючи семантичні ознаки слів у словосполученні, в базі даних з кореляторами знаходиться

потрібний корелятор, що встановить зв'язок між словами. Однозначність визначення такого зв'язку забезпечується тим, що для окремого визначника множини пар семантичних ознак для кореляторів з його списку не перетинаються.

Далі до словосполучення поступово приєднуються прилеглі повнозначні слова речення, шляхом встановлення зв'язку між новим словом та одним із слів опрацьованої частини речення. Так створюється група зв'язаних слів. Важливим є вибір слова з групи зв'язаних слів, яке буде пов'язуватись з наступними словами. Це має бути слово з головним відношенням, або останнє слово групи. У випадку, коли неможливо встановити зв'язок між новим словом та словами групи, створюється нова група зв'язаних слів. На завершальному етапі аналізу необхідно спробувати поєднати всі створені групи в одну, яка відображатиме структуру зв'язків між всіма словами речення.

Неможливість встановлення зв'язку між окремими групами в реченні та їх об'єднання свідчить або про складне речення, частини якого не пов'язані (або пов'язані неявно) між собою, або про некоректні зв'язки між словами у групах. Для уникнення проблеми некоректного зв'язування слів необхідно проводити додатковий аналіз можливих зв'язків кожного наступного слова із словами групи та вибирати найбільш вірогідний, або розглядати всі можливі варіанти зв'язків (що недоцільно, враховуючи зростання кількості таких варіантів для кожного наступного слова).

Алгоритм співвіднесення анафоричних зв'язків. Однією з перших необхідно вирішити задачу анафоричних зв'язків, або заміни займенників у тексті на відповідні поняття (іменники). Заміщенню підлягають деякі особові, відносні, вказівні, присвійні та зворотні займенники. Алгоритм заміщення будується на основі аналізу закономірностей (відображених у базі знань синтаксису ПМ) вживання займенників у природній мові й відрізняється в залежності від типу та граматичних характеристик займенника. Із практики відомо, що найчастіше поняттям, якому відповідає займенник є узгоджене з ним за граматичними характеристиками найближче повнозначне слово, яке стоїть попереду, або слово, що входить до ядра даного чи попередніх речень. Тому вихідними даними для проведення заміни займенників є граматичні характеристики слів (отримані в результаті морфологічного аналізу) та синтаксично-семантичні зв'язки між ними (результат роботи синтаксичного блоку). Блок-схема алгоритму заміни більшості займенників показана на рис. 2.

Орієнтація розроблюваної системи на ПМО з правильно побудованими реченнями гарантує те, що завжди можна знайти іменник у реченні, що відповідає займеннику.

Заміщення займенників дозволяє отримати додаткові зв'язки між відповідними поняттями у реченні та загалом у тексті, тому після цього доцільно провести повторний синтаксичний аналіз речень з займенниками.

Структурна схема та загальні принципи функціонування блоку онтолого-семантичного аналізу детально описані в [1, 6, 7].

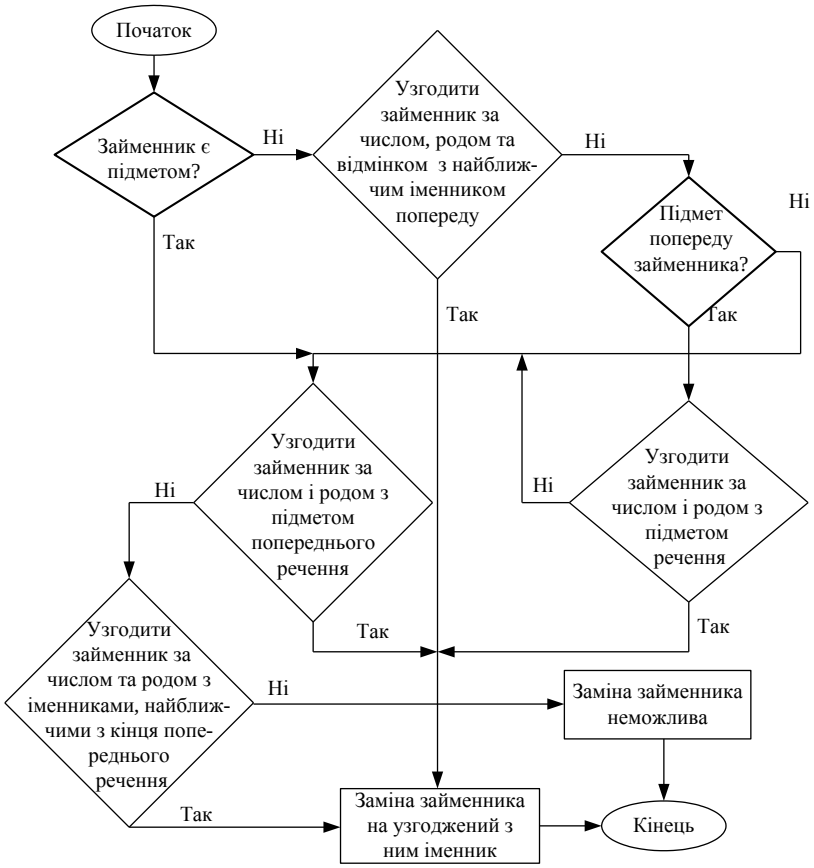

РИС. 2. Блок-схема алгоритму заміни займенників

Інформаційна модель лінгвістичного аналізу. Блок-схема узагальненого алгоритму лінгвістичного аналізу та формалізації деякого ПМО (рис. 3) являє собою послідовність етапів графемно-морфологічного, синтаксичного, об'єктово-семантичного, акторно-семантичного, онтологічного, онтолого-семантичного та формально-логічного аналізу. Деякі аспекти безпосереднього лінгвістичного аналізу вищерозглянуто та описано в [5–8].

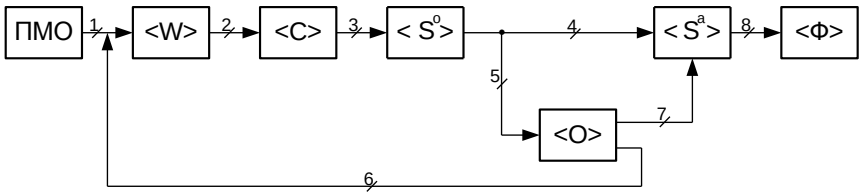

РИС. 3. Блок-схема лінгвістичного аналізу

На рис. 3 прийнято наступні скорочення:

- ПМО – природномовний об’єкт;
- W – послідовність словоформ, що описують ПМО;
- C – послідовність синтаксичних структур речень, що входять в ПМО;
- S^O – послідовність об’єктово-семантичних структур речень;
- S^a – послідовність акторно-семантичних структур речень;
- O – послідовність онтографів;
- Ф – формально-логічне представлення ПМО;
- 1 – графемно-морфологічний аналіз;
- 2 – синтаксичний аналіз;
- 3 – об’єктово-семантичний аналіз;
- 4 – акторно-семантичний аналіз;
- 5 – онтологічний аналіз;
- 6 – аналіз неоднозначностей;
- 7 – онтолого-семантичний аналіз;
- 8 – формально-логічний аналіз.

Далі акцентуємо увагу на аналізі неоднозначностей, притаманних будь-якому ПМО. Зазначений аналіз виконується ітеративно завдяки зворотному зв’язку від блоку побудови онтографів O.

Однією із суттєвих складових пропонованого підходу є потрійний аналіз неоднозначностей, два з яких вищеописано, а третій, найбільш складний, опишемо далі.

На перших двох етапах виконуються спрощення для неоднозначностей морфологічного та синтактико-семантичного типу. При цьому відповідні записи у морфологічній таблиці ПМО видаляються. Перехід до третього етапу аналізу неоднозначностей (що мають контекстно-семантичні витоки) виконується тоді, коли для будь-якої словоформи морфологічній таблиці ПМО залишилось два (чи більше) записи.

Спочатку виконується спроба побудувати акторно-семантичне відображення деякого речення ПМО для перших записів словоформ (із морфологічної таблиці), що входять у речення. При цьому інтерпретатор аналізує відповідний онтограф речення згідно правил бази знань природної мови. Якщо результат інтерпретації – “істина”, то формується акторно-семантична структура речення, а якщо – “хибність”, то активується зворотній зв'язок від онтологічного блоку до блоку морфологічного аналізу (на рис. 3 позначено цифрою 6). Однією із умов при формуванні значення істинності може бути умова зв’язності онтографа.

Наступний крок – формування складників речення з послідуючих записів морфологічної таблиці, побудова синтаксичної та об’єктово-семантичної структури речення та спроба побудувати акторно-семантичну структуру. Ітераційний процес продовжується доти, доки не буде побудовано повністю акторно-семантичну структуру речення та відповідні онтографи. На завершення виконується формування формально-логічного представлення ПМО.

Висновки. Розглянуто комбіноване розпізнавання синтаксичних та семантичних відношень, що зв'язують слова в тексті, яке являє собою підхід безпосереднього розпізнавання з елементами граматичного аналізу. Запропонований підхід дозволяє на етапі синтактико-семантичного аналізу зменшити обсяг нерозпізнаних неоднозначностей до мінімального. Використання наведеного алгоритму заміщення займенників дає можливість отримати додаткові зв'язки між відповідними поняттями у реченні та у всьому тексті. Потрійний аналіз неоднозначностей з використанням зворотного зв'язку від онтологічного блоку до блоку морфологічного аналізу дозволяє побудувати більш адекватну акторно-семантичну структуру речення. У подальших дослідженнях у цьому напрямку необхідно конкретизувати побудову онтологічного представлення ПМО в цілому, розробити та обґрунтувати механізм формального переходу від мовного до машинного представлення текстової інформації.

1. *Палагин А.В., Петренко Н.Г.* К проектированию онтологоуправляемой информационной системы с обработкой естественно-языковых объектов // Математичні машини і системи. – 2008. – № 2. – С. 14–23.
2. *Рыков В.В.* Управление знаниями. – http://ryk-kypc2.narod.ru/part2.doc.
3. *Гладун В.П.* Процессы формирования новых знаний. – София: Педагог, 1994. – 190 с.
4. *Гладун В.П., Величко В.Ю.* Конспектирование естественно-языковых текстов // Proceedings of the XI International Conf. "Knowledge-Dialogue-Solution"(KDS'2005). – Varna, Bulgaria. – 2005. – Vol. 2. – P. 344–347.
5. *Палагін О.В., Петренко М.Г.* Модель категоріального рівня мовно-онтологічної картини світу // Математичні машини і системи. – 2006. – № 3. – С. 91–104.
6. *Палагін О.В., Петренко М.Г.* Архітектурно-онтологічні принципи розбудови інтелектуальних інформаційних систем // Математичні машини і системи. – 2006. – № 4. – С. 15–20.
7. *Palagin A., Gladun V., Petrenko N., Velychko V., Sevruk A., Mikhailyuk A.* Informational model of natural language processing // International J. "Information Technologies and Knowledge" . – 2008. – Vol. 2 – P. 5–6.
8. *Палагин А.В., Петренко Н.Г.* К вопросу системно-онтологической интеграции знаний предметной области // Математичні машини і системи. – 2007. – № 3–4. – С. 63–75.

Отримано 23.10.2008