УДК 004.89

В. В. Литвин, Т. М. Черна
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ПРОБЛЕМА АВТОМАТИЗОВАНОЇ РОЗБУДОВИ БАЗОВОЇ ОНТОЛОГІЇ

Досліджено задачу автоматизованої розбудови базової онтології. Запропоновано метод, алгоритм і засоби для виділення знань з природомовного тексту. Показано, що такий алгоритм має бути багатоетапним і містити ієрархічну кількарівневу процедуру розпізнавання понять, зв’язків, предикатів та правил, які в результаті належать до онтології.

Ключові слова: онтологія, навчання онтологій, автоматизована розбудова, база знань, текстовий документ.

In the paper the method of the automatic development of ontology has been developed. A method, algorithm and means for selection of knowledge from the text document is proposed. It is shown that this algorithm has to be multistage and involve hierarchical recognition procedure of concepts, relations, predicates and rules which are included into the resulting ontology.

Keywords: ontology, learning ontologies, automatic development, knowledge base, text document.

Вступ. Постановка проблеми у загальному вигляді

Очевидно, що для того, щоб вручну побудувати повну зв’язану онтологію для певної предметної області (ПО), необхідно затратити достатньо багато часу та ресурсів. Причина таких затрат полягає в тому, що такі онтології повинні містити десятки тисяч елементів, щоб бути придатними для розв’язування широкого кола прикладних задач, які виникають у цих ПО. Отже, ручна побудова онтології людиною-оператором – це тривалий рутинний процес, який, до того ж, вимагає ґрунтовних знань ПО та розуміння принципів побудови онтологій.

Тому побудуємо математичне забезпечення автоматизації побудови онтології, а точніше, її розбудови, оскільки вважаємо, що базові терміни та відношення між ними повинні бути введені людиною-експертом в онтологію вручну. Таку онтологію називатимемо базовою і позначатимемо $O_{base} = \langle C_b, R_b, F_b \rangle$. Тобто побудова онтології починається з моменту, коли в ній вже є якісь дані. Тому такий процес називатимемо розбудовою базової онтології і позначатимемо:

$$\gamma : O_{base} \to O . \qquad (1)$$

Для побудови онтологій, які адекватно описують семантичні моделі ПО, необхідно, насамперед, розв’язати задачі одержання знань із різних джерел для виявлення множини концептів і встановлення ієрархії на цій множині. Оскільки значна частина інформації міститься в природно-мовних текстах (ПМТ), перспективним є одержання знань із текстової інформації, а також інтелектуальне опрацювання спеціально підібраних колекцій ПМТ.

Аналіз останніх досліджень та публікацій

У цій роботі розглянуто методи і моделі інтелектуального опрацювання текстів, результати якого призначені для побудови онтологій ПО, а також використовуються під час онтологічного навчання (Ontology Learning), коли необхідно покращити, розширити, модифікувати існуючу модель онтології або розбудувати онтологію із базової онтології, маючи як джерела знань тільки

текстові колекції. В останньому випадку задача є особливо складною і потребує застосування всього спектра методів інтелектуального аналізу текстової інформації (Text Mining – TM) [1]. У роботі розглядається розв'язування задач ТМ на різних етапах опрацювання ПМТ: одержання інформації (виявлення сутностей – концептів і термінів, їхніх властивостей, фактів, подій, встановлення відношень між сутностями, зокрема асоціативні), категоризація, кластеризація, семантична анотація. Нижче розглянемо засоби автоматизованого аналізу природної мови та програмні продукти, реалізовані на їх основі для наповнення БЗ системи.

Відома низка перспективних лінгвістичних розроблень, серед яких доцільно виділити метод лексико-граматичного аналізу (Part-of-Speeach-tagging), який полягає в автоматизованому розпізнаванні, до якої частини мови належить кожне слово у тексті [2]. Для підвищення точності такого аналізу використовують два типи алгоритмів: ймовірнісно-статистичні та алгоритми на основі продукційних правил, що оперують словами і кодами. Щодо останніх, то вони можуть використовувати правила, автоматизовано зібрані з корпусу текстів [3] або ж підготовлені кваліфікованими лінгвістами [4].

На відміну від лексико-граматичного аналізу метою синтаксичного аналізу (Text Parsing) є автоматична побудова дерева фрази, тобто знаходження взаємозалежностей між різнорівневими елементами речення. Існує велика кількість різних підходів до синтаксичного аналізу текстів, наприклад, Ergo Linguistic Technologies Parser, який розробили Д. Бікертон та Ф. Бралік із університету в Гонолулу [5]. Аналізатор використовує схему нотації, прийняту в Penn Treebank, він орієнтований на впровадження в інтерфейсах типу питання-відповідь і є комерційним продуктом. Іншим вдалим синтаксичним аналізатором є Functional Dependency Grammar, побудований дослідниками з університету в м. Гельсінкі (засновниками фірм Lingsoft та Conexor). В основу аналізатора покладено теорію залежностей, яку вперше запропонував Л. Теснієр, яку реалізовано в межах контекстно-залежної граматики [6]. Також алгоритми використання іменних груп, виділених за допомогою часткових синтаксичних аналізаторів [7], використовують у програмних продуктах TextAnalyst (НВІЦ "МікроСистеми") та Extractor (Інститут інформаційних технологій Національної дослідної ради Канади), зокрема, останній застосовується у пошуковій системі журналу досліджень у галузі штучного інтелекту.

Серед систем, розроблених в Україні, треба зазначити розроблення кафедри математичної інформатики Київського національного університету імені Тараса Шевченка – систему опрацювання текстів природною мовою. Систему створено для розв'язування таких задач, як аналіз та синтез текстів природною мовою, автоматизоване генерування реферату тексту, автоматизована індексація (визначення тематики) тексту. Найвагомішим технічним рішенням у системі є можливість «зважувати» вершини семантичної мережі тексту. При цьому найважливішими вершинами мережі вважаються вершини, котрі мають найбільшу кількість зв'язків з іншими. Цю процедуру можна застосовувати для побудови образу реферату зважуванням вершин і відкиданням найлегших – «маргінальних».

Формування мети

Розробити метод, алгоритм та програмні засоби для автоматизованої розбудови базової онтології.

Аналіз отриманих наукових результатів

1. Структура онтології

Одним з найефективніших підходів до наповнення онтології є її автоматизоване навчання природномовними текстами. Автоматизоване наповнення можна реалізувати за допомогою аналізу текстових документів, застосувавши процесор знань (рис. 1).

У поданій схемі завдання лінгвістичного процесора – виконати його лексичний, лексико-граматичний, синтаксичний та семантичний аналіз. У результаті цього онтологія поповнюється: поняттями, СДО-трійками (суб'єкт – дія – об'єкт) і причинно-наслідковими зв'язками між СДО-трійками. Іншу частину важливих зв'язків між поняттями та їхніми властивостями

встановлює процесор онтологій, котрий будує онтологічну структуру для кожного концепту C_i, отриманого після аналізу тексту. Робота процесора онтологій підтримується відповідною БЗ, основними компонентами якої є, по-перше, множина правил, по-друге, універсальна логічна БД MModWN типу WordNet [8]. Процесор знань застосовують у системі автоматизованого одержання знань із текстових документів, котра, своєю чергою, застосовується для розв'язання задачі семантичного пошуку в повнотекстових БД.

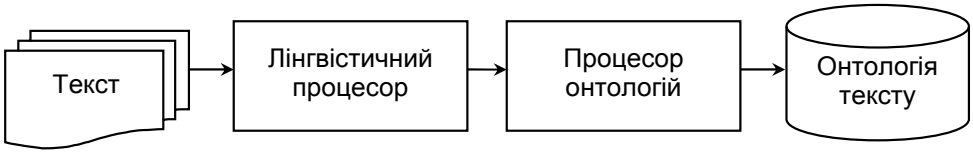

Рис. 1. Структурно-функціональна схема процесора знань

Як зазначено у [9]: онтологія – це мова науки. Мова науки як структуроване наукове знання – багатошарове ієрархічне утворення, в якому виділяють блоки [10]: терміносистема; номенклатура; засоби та правила формування понятійного апарата і термінів.

Отже, з погляду процесу побудови онтології необхідно побудувати її терміносистему O_T та номенклатуру O_N. За нашим підходом базова онтологія повинна точно об'єднувати частину терміносистеми (див. рис. 2), тобто

$$O_B \cap O_T \neq \varnothing .$$

Енциклопедії, термінологічні та тлумачні словники, на основі яких будується терміносистема ПО, як правило, мають чітку структуру і складаються із словникових статей. Тому необхідно дослідити можливі їх структури з метою розпізнавання понять і відношень між ними.

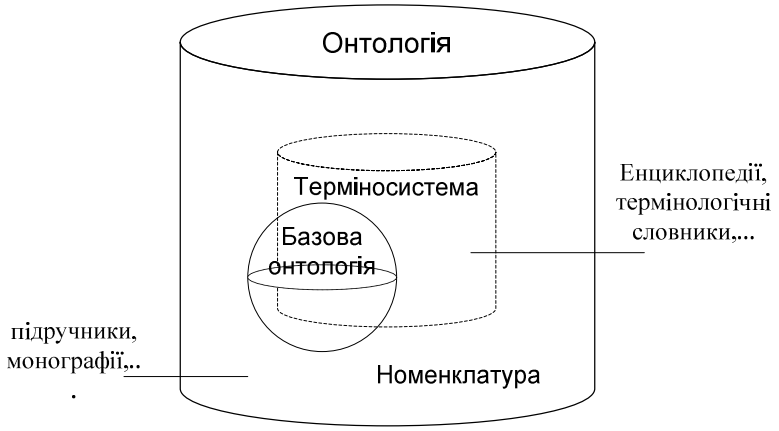

Рис. 2. Архітектура онтології

Побудова номенклатури складніша. Якщо в словниках терміни деяким способом уже виділено, то в наукових текстах (підручники, монографії тощо) їх необхідно виділяти, шукати властивості понять і відношень між поняттями.

Необхідно розробити методи одержання термінів із джерел знань, їхню фіксацію і розподіл за категоріями. Таке видобування пов'язане з природномовним опрацюванням інформації. Такі методи постійно удосконалюються, перебувають у постійному розвитку, тому їхнє подання повинно бути декларативним, оскільки таке подання забезпечує найпростіший спосіб удосконалення. Наукові тексти є монологічними. Отже, робимо висновок, що потрібна технологія, яка дозволила б майже в автоматизованому режимі створювати методи природномовного опрацювання наукового тексту, які вже в автоматизованому режимі дозволили б будувати онтологічні моделі. Так, у роботі [9] для цього використано генетичні алгоритми для генерування моделей рішень та автоматизоване програмування для автоматизованої генерації коду програмного забезпечення. Огляд відомих підходів та проектів автоматизованої побудови онтологій наведено у [11].

2. Особливості автоматизованої побудови онтології

Відправною точкою під час створення будь-якої моделі знань про ПО є вибір його категоріального апарата. Для будь-яких абстрактних систем, які використовують один і той самий тезаурус чи словник, не існує гарантії, що вони зможуть правильно використовувати одну і ту саму інформацію, поки не буде прийнято єдиної концептуалізації. В основу концептуалізації покладено категорію абстракцій, які пов'язані з конструкцією терміна, на якому основується будь-яка онтологія. Обґрунтуємо побудову конструкції терміна як знака семіотичної системи.

Сьогодні не існує єдиного правильного способу моделювання ПО. Однак необхідно виділити деякі фундаментальні правила розроблення онтології:

- ефективний розв'язок завжди залежить від пропонованої програми і очікуваних розширень;
- розроблення онтології – це обов'язково ітеративний процес;
- поняття в онтології повинні бути близькі до об'єктів і відношень в ПО.

У роботі використано категоріальний апарат, виведений на основі робіт лінгвістів, логіків та інформатиків [12, 13]. Визначення категоріального апарата пов'язане, з одного боку, з виявленням концептуальних об'єктів об'єктивної реальності і відношень між ними, з іншого, з їх поданням. Дійсно, одна з інтерпретацій мови наукових текстів пов'язана з розумінням його як знакової системи: мова математики, хімії, тобто з виробленими в різних науках штучними символьними мовами. У них штучні лексика та синтаксис. Ці мови входять в науковий текст, утворюючи тим самим частину мови науки і роблячи його природноштучним утворенням. Насамперед опишемо основні поняття, які використовуватимуться надалі.

Термін – це знак спеціальної семіотичної системи, який має номінативно-дефінітивну функцію. Номінативною – тому, що термін іменує, означає цілий складний змістовний фрагмент із загальної побудованої системи інтенсіоналів (змістів). Дефінітивною – тому, що заміщує дефініцію, яка має експліцитний і/або імпліцитний вигляд із низки висловлювань і розуміє ту дефініцію у своєму вжитку, будучи відносно неї другорядним чинником. Специфіка термінології полягає в усвідомленні змісту знаків мови науки, тобто в можливості того, хто говорить, експлікувати дефініцію використаного терміна. Тобто термін – це знак спеціальної семіотичної системи, який є мінімальним носієм наукового знання, коротка назва поняття, що має дефініцію.

Дефініція – це таке об'єднання форм структурної та субстанціональної дефініцій, за якого із структурної інформації випливає подання про найімовірнішу субстанціональну, а із субстанціональної інформації – найімовірніші структурні взаємодії елементів поля терміносистеми, так що в сукупності ці два аспекти забезпечують подання про її цілісність і функціональну валідність. Це означає, що необхідно мати як субстанціональну дефініцію терміна словесні визначення терміна, а як структурну дефініцію – фрагмент мережі знаків.

Референт – це подання про денотати сутностей реального світу (об'єкт, явище, процес), знання про які описується в знаковій системі.

Концепт – це знання, яке виражається цим поняттям під час концептуального моделювання ПО.

Інтенсіонал – це зміст поняття, який відповідає структурній дефініції і описується як внутрішня форма поняття, що об'єднує його лексис і логос і достатня для задання екстенсіонала.

Екстенсіонал – це об'єм поняття.

Концептуальні об'єкти поділяються так:

- сутність – матеріальні та нематеріальні об'єкти, способи їх розгляду;
- властивість – кількісні, якісні, релятивні (відношення);
- дія – операції, процеси, стани;
- величини – час, простір, …;

Концептуальні відношення:

- квантітативні (збігаються з теоретико-множинними відношеннями тотожності, врахування, вилучення, перетину, об'єднання);
- квалітативні (ієрархічні та функціональні).

У галузі ШІ вважається, що реальний світ складається з об'єктів. Об'єкти можуть складатися із частин. Об'єкти мають властивості, які мають значення. Об'єкти можуть перебувати в різних відношеннях один з одним. Властивості та відношення змінюються в часі. У різні моменти часу виникають події, які активізують процеси, в яких беруть участь об'єкти і які змінюються в часі. Події можуть спричиняти інші події, тобто давати ефект. Світ та його об'єкти можуть перебувати в різних станах.

3. Узагальнена схема опрацювання монологічних текстів

Методи побудови онтологій умовно розділимо на групи. До першої групи ввійдуть традиційні методи природномовного опрацювання тексту, до другої – методи, що стосуються безпосередньо побудови онтології.

Розглянемо технологію аналізу природномовного тексту та побудови на їх основі онтології, запропоновану в роботах [14]. Модифікований вигляд узагальненої схеми наведено на рис. 3.

Методи реалізації перших чотирьох блоків вважають найбільше проробленими. Однак дослідження тривають, тому що задовільного результату їхньої роботи поки не отримано. Так, пропонуємо використовувати онтологію мови для проведення відповідних аналізів.

У функції попереднього опрацювання входять лексичний аналіз, розбивання складних речень на прості, виділення у вихідному тексті розділів, підрозділів, речень, перевірка виконання прийнятих обмежень. На цьому етапі досліджень наукових текстів неприпустимими вважаються складнопідрядні речення, що об'єднують рекурсивно-вкладені означальні речення.

Основне завдання **лексичного аналізу** полягає в розбиванні вхідного тексту документа, що являє собою послідовність одиничних символів, на послідовність лексем. Із цього погляду всі символи вхідної послідовності поділяються на символи, які належать яким-небудь лексемам, і символи, що поділяють лексеми (роздільники). У деяких випадках між лексемами може й не бути роздільників. У результаті лексичного аналізу формується множина лексем $L = \{l_i \mid i=1, \ldots, k,$ k – кількість лексем у тексті$\}$. Кожній лексемі приписується вектор:

$$\rho_i = \left\langle p_i, n_i^l, n_i^s, n_i^p, n_i^d, n_i^c \right\rangle, \qquad (2)$$

де p_i – унікальний номер вектора лексеми; n_i^l – порядковий номер лексеми в реченні; n_i^s – порядковий номер речення в тексті; n_i^p – номер параграфа; n_i^d – номер розділу; n_i^c – номер глави.

Основною функцією пословесного **морфологічного аналізу** є визначення частини мови лексеми l_i і присвоєння їй вектора морфологічної інформації ρ_i. Під час аналізу лексем використовуються словники закінчень, словник флективних класів, словник готових словоформ й словник основ, таблиці сумісності основи флективного класу й вектори морфологічної інформації.

Синтаксичний аналіз. Під час синтаксичного аналізу необхідно однозначно визначити всі синтаксичні одиниці природномовного речення. Синтаксичними одиницями називаються конструкції, в яких їхні елементи (компоненти) об'єднано синтаксичними зв'язками й відношеннями. Синтаксичний зв'язок є вираженням взаємозв'язку елементів у синтаксичній одиниці, тобто слугує для відображення синтаксичних відношень між словами, створює синтаксичну структуру речення й словосполучення, а також умови для реалізації лексичного значення слова. Зазвичай розглядається тільки один вид синтаксичного зв'язку – підпорядкування. Цей вид синтаксичного зв'язку передає відношення між фактами об'єктивного світу у вигляді сполучення двох слів, в якому одне є головним, а друге – залежним. Відношення між лексемами представляються у вигляді лексико-граматичних зв'язків між словами, які являють собою питання від головного слова до залежного (наприклад, система (яка) операційна). Вхідними даними для проведення синтаксичного аналізу є результати морфологічного аналізу, подані у вигляді множини пар $\langle l_i, \rho_i \rangle$, де l_i – лексема, ρ_i – вектор морфологічної інформації лексеми l_i. У результаті проведення синтаксичного аналізу формується граф залежностей G, у вершинах якого містяться лексеми. Вершини з'єднуються дугами, які вказують напрямок зв'язку від головного слова до залежного.

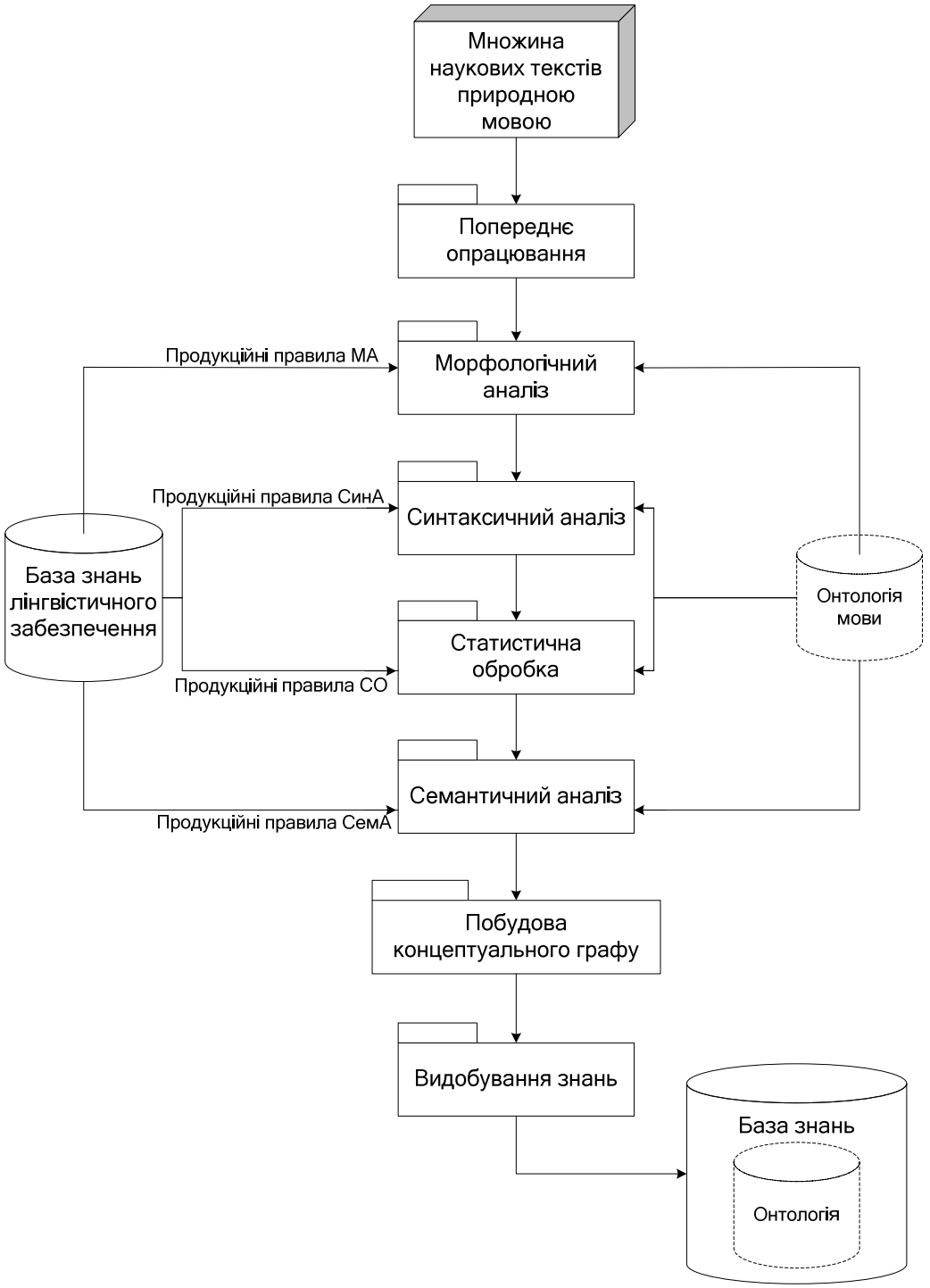

Рис. 3. Загальна схема автоматизованої побудови онтології

Статистичне опрацювання тексту не є обов'язковим для кожної системи природномовного опрацювання тексту. Зазвичай вона є наявна у пошукових системах і системах автоматизованого реферування. Детальніше ці системи розглянуто в розділі. Статистичні методи засновані на частотних характеристиках тексту: частота входження слова в текст, частота спільного входження декількох слів, зважена частота входження тощо. За цими методами відношення між словами не аналізуються з лінгвістичного погляду. Під час статистичного аналізу шукають вхідні послідовності слів і виділяють поняття, під якими розуміють слова й словосполучення, а також визначають їхні частотні характеристики. Особливо важливо знайти субстантивні іменні словосполучення, які задаються схемою: узгоджуване слово + іменник [15].

Пофразний семантичний аналіз. Ціль семантичного аналізу полягає у визначенні для кожного слова й фрази загалом деяких змістовних характеристик. Зміст фрази, як правило, подають у вигляді фрагмента семантичної мережі. Основою для побудови фрагмента є граф залежностей. Результатом семантичного аналізу є перетворення графу залежностей на фрагмент семантичної мережі.

Побудова концептуального графу об'єднує два компоненти: побудова єдиної семантичної мережі й видобування прагматичної інформації, тобто з аналізованого тексту видобувається його прагматичний зміст. Для побудови концептуального графу наукового тексту використано програмний продукт Link Grammar Parser (http://www.link.cs.cmu.edu/link).

База знань лінгвістичного забезпечення складається із трьох частин: бази фактів, бази правил і бази знань про ПО. База фактів містить словник готових словоформ, словник закінчень, словник флективних класів і словник основ. База правил складається із продукційних правил лексичного, морфологічного, синтаксичного, статистичного й семантичного аналізів. Повний опис всіх методів наведено у роботах [16, 17].

4. Методи побудови онтологій

Онтологія описує поняття певної ПО і відношення між ними. У цьому сенсі знання стають можливими для повторного використання людьми, базами даних і програмними комплексами. До того ж значно підвищується ефективність як інтелектуальних систем, так і традиційних інформаційних систем [18]. Цим визначається актуальність створення онтологій. Наразі розроблено доволі багато систем, що дають змогу у діалоговому режимі створювати онтології. Однак цей процес характеризується значною трудомісткістю. Тому знання про поняття необхідно одержувати з повнотекстових джерел знань й автоматизовано будувати онтології. Так, наприклад, для створення терміносистеми, що є ядром онтології ПО, знання можна одержувати з термінологічних і тлумачних словників [19]. Проекції терміносистеми на конкретні галузі знань (задача, вид діяльності) називаються номенклатурами [20]. Для побудови номенклатури знання можна видобувати з наукових і навчальних видань.

Побудова терміносистеми предметної області

Можливість автоматизованої побудови онтології ПО забезпечується одержанням знань із якісних термінологічних і/або тлумачних словників. До того ж терміносистема, створена на основі знань, отриманих зі словників, являє собою ядро онтології ПО. Кінцевий варіант онтології повинен бути створений за допомогою об'єднання декількох ядер онтологій, побудованих на основі різних термінологічних словників. Треба зазначити, що термінологічні словники, призначені для створення онтології ПО, повинні відбиратися експертом з галузі знань. Для побудови повної онтології ПО необхідно побудувати номенклатури ПО, які будуються на основі одержання знань із таких наукових текстів, як монографія, навчальний посібник, стаття тощо. Потім необхідно об'єднати терміносистему й номенклатури.

Термінологічні словники як джерела знань. Відомо кілька класифікацій (типологій) словників. Тип будь-якого словника визначається характером лексичного матеріалу й практичним значенням [20]. Так, енциклопедичні (від грецького *enkyklios paideia* – навчання зі всього кола знань) словники містять екстралінгвістичну інформацію про описувані мовні одиниці. Ці словники містять відомості про наукові поняття, терміни, історичні події, персони, географії тощо. В енциклопедичному словнику немає граматичних відомостей про слово, а залежно від обсягу й адресата словника дається більш-менш розгорнута наукова інформація про предмет, який визначається словом. Об'єктом описування лінгвістичних (мовних) словників є мовні одиниці – слова, словоформи, морфеми. У такому словнику слово може бути охарактеризоване з різних аспектів залежно від цілей, обсягу й завдань словника: з боку змісту, словотвору, орфографії, орфоепії, правильності вживання.

Крім того, розрізняють словники з погляду відбору лексики: словники тезаурусного типу й словники, у яких лексика відбирається за певними параметрами. Наприклад, за сферою вживання розрізняють розмовні, просторічні, діалектні, термінологічні словники. За історичним напрямом –

словники архаїзмів, історизмів, неологізмів тощо. З погляду розкриття окремих аспектів (параметрів) слова словники можуть бути етимологічні, граматичні, орфографічні тощо. З погляду розкриття системних відношень між словами виділяють гніздові, словотворчі, омонімічні, паронимічні (план виражень), синонімічні, антонімічні (план змісту) словники.

Розглядатимемо словники з погляду можливості їхнього використання як джерела знань. Для побудови предметної онтології потрібні тільки якісні словники, що містять не тільки визначення терміна, але й опис властивостей, відношень, синонімів та інших елементів знання про термін. Тому як джерела знань краще використовувати словники, у яких дається більш-менш розгорнута наукова інформація про предмет. Така інформація зустрічається в енциклопедичних, тлумачних, термінологічних словниках.

Будь-який словник складається зі словникових статей. Словники відрізняються структурованістю словникових статей.

Основна частина словників не має чіткої структури словникових статей. Як правило, у словниковій статті дається одне або кілька визначень (дефініцій) понять, а потім описується відношення терміна з іншими поняттями, за допомогою яких пояснюється суть терміна. Ці відношення можуть мати родоподібний характер, частина–ціле, задавати метрику терміна, описувати властивості терміна, визначати процеси, які відбуваються з терміном або над терміном тощо.

Словникова стаття будь-якого словника розпочинається із заголовного слова, яке є іменем терміна або термінологічного словосполучення. Заголовне слово може бути набране прописними буквами, виділене жирним шрифтом або іншим способом. За заголовним словом йде текст, який пояснює заголовну одиницю в словнику й описує її основні характеристики. За ступенем структурованості тексту можна виділити словники, що мають строгу структуру словникових статей. Так чітку структуру словникової статті має термінологічний словник з основ інформатики та обчислювальної техніки [21], словникова стаття якого має чотири частини:

- заголовна частина, яка містить заголовок словникової статті й дефініцію терміна;
- частина, в якій розкрито зв'язки терміна або термінологічного словосполучення з іншими словами в реченні або тексті;
- ілюстративна частина демонструє реальне вживання терміна або термінологічного словосполучення;
- довідкова частина розкриває походження терміна або термінологічного словосполучення.

У кожній частині словникової статті можна виділити структурні елементи, які мають певний порядок слідування. Так, заголовна частина починається із заголовка словникової статті. Для термінологічного словосполучення за заголовком вказується його скорочення: наприклад, *інформаційні технології (ІТ)*. У заголовній частині слів-термінів за заголовком наводять граматичну характеристику: наводяться закінчення родового відмінка однини, повністю форма множини, закінчення родового відмінка множини й вказівка на рід іменника. Потім розгорнуте тлумачення, яке розкриває структуру поняття терміна та його складових частин, його англійський й німецький відповідники. Окремі значення багатозначних термінів відзначаються порядковим номером, за яким йдуть їхні тлумачення. Отже, термінологічні словники містять термінологію однієї або декількох спеціальних галузей знань або діяльності, яка використовується у сучасному світі. У них дано основні поняття, без яких важко обійтися в конкретній діяльності, і доволі детальні пояснення. Структура словникових статей термінологічних словників різна й розробляється укладачами для кожного словника. Ступінь структурованості змісту словникової статті, порядок слідування її структурних елементів, повнота викладення залежать від цільового призначення словника, специфіки галузі знань.

Побудова семантичної мережі знаків-фреймів як моделі подання терміносистеми

Інтерпретація знака "поняття" t є центральною ланкою моделі подання знань й ототожнюється з елементарним фрагментом Φ семантичної мережі *SF* ПО: $t \xrightarrow{def} \Phi$, де Φ – знак-фрейм семантичної мережі.

Оскільки кожна вершина такого знака-фрейму є вектор або множина, то вона розкривається пучком компонентів вектора або множини, які, своєю чергою, можуть теж розкриватися пучком компонентів. Отже, буде побудована єдина семантична мережа *SF* знаків-фреймів.

Побудова семантичної мережі знаків-фреймів, аналіз побудованої мережі, об'єднання мереж здійснюється за допомогою методів, які задаються у вигляді систем продукцій. Спочатку активізуються продукції, які виявляють множину заголовних термінів $T = \{t_i\}$. Заголовними термінами t_i заповнюються фрейми-прототипи концептуального об'єкта "Поняття", у результаті формується множина екзофреймів $\{\Phi_{Vi} \mid i = 1,...|T|\}$.

Отже, на початковому етапі матимемо множину ізольованих фреймів $\{\Phi_{Vi}\}$, які в міру заповнення слотів фрейму-прототипу об'єднюються в єдину мережу:

$$SF = \bigcup_{i=1}^{n} \Phi_{Vi},$$

де *n* – кількість термінів ПО; Φ_{Vi} – фрейми, які описують всі концептуальні об'єкти ПО.

Отже, основною процедурою побудови мережі є послідовний аналіз кожної словникової статті термінологічного словника, що складається з таких процесів розпізнавання: дефініцій; квантитативних відношень; кваліітативних відношень. Розпізнавання множини дефініцій терміна. Під час видобування дефініції виникають такі ситуації:

1) словникова стаття може мати одну або кілька дефініцій;

2) якщо в статті є одна дефініція, то вона починається після першого символу '–', що зустрівся в словниковій статті;

3) якщо в статті є кілька дефініцій, то їх можна нумерувати або арабськими цифрами, або буквами латинського алфавіту;

4) у разі нумерації дефініцій після номера може стояти або символ '.', або символ ')'.

Отже, ознаками початку дефініції є символ '–' або символ вигляду "#.", "#)". Тут символ '#' позначає арабську цифру або латинську букву. Крім того, дефініція завжди розташована в першому реченні словникової статті. Це означає, що для виявлення дефініції потрібно визначити, чи містить перше речення словникової статті зазначені ознаки.

Зазначимо, що напрям досліджень автоматизованої по(роз)будови онтологій, БЗ за допомогою природомовних текстів та систем на їхній основі активно розвивається. Зокрема щорічна Європейська конференція зі штучного інтелекту проводить засідання окремої секції з навчання онтологій, на якій розглядає досягнення в галузі їх автоматизованого формування.

Висновки і перспективи подальших наукових розвідок

Проаналізовано стан досліджень та розробок у галузі видобування знань з природомовних текстів. Запропоновано загальний алгоритм, необхідні методи і засоби для виділення нових знань з природномовного тексту, показано, що такий алгоритм має бути багатоетапним і включати в себе ієрархічну кількарівневу процедуру розпізнавання понять, зв'язків, предикатів та правил, які в результаті вносяться до онтології з метою виконання перерахунку очікуваної корисності.

Побудовано метод розбудови онтології, який ґрунтується на використанні вже існуючої онтології під час аналізу текстових документів, які використовуються під час побудови номенклатури та терміносистеми онтології. Здійснено класифікацію відношень та розроблено відповідні шаблони для їх пошуку у природномовних текстах. Все це дало змогу автоматизувати процес розроблення онтології, а отже, суттєво зменшити на це витрати.

1. Feldman R. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data / R. Feldman, J. Sanger – Cambridge University Press, 2007. – 410 p. 2. Анисимов А. В. Система обработки текстов на естественном языке / А. В. Анисимов, А. А. Марченко // Научн.-теорет. журн. "Искуственный интелект", ІПШІ "Наука і освіта". – 2002. – Вип. 4. – С. 157–163. 3. Гладун А.Я. Формирование тезауруса предметной области как средства моделирования информационных

потребностей пользователя при поиске в Интернете / А. Я. Гладун, Ю. В. Рогушина // Вестник компьютер. и информ. технологий. – № 1. – 2007. – С. 26–33. 4. Shanjian L. A composite approach to language [Електронний ресурс] / L.Shanjian, M.Katsuhiko // encoding detection. – Режим доступу: http://www. mozilla.org/ projects/intl/UniversalCharsetDetection.html. 5. Малащук Е. В. Обзор существующих алгоритмов Data Mining для глубинного анализа текстов и методов извлечения знаний / Е. В. Малащук, Д. В. Бабин, С. М. Вороной // "Искусственный интеллект", ІПШІ "Наука і освіта". – 2005. – Вип. 4. – С. 618–626. 6. Хомский Н. Аспекты теории синтаксиса. / Н. Хомский. – М., 1972. – 259 с. 7. Фаин В.С. Машинное понимание естественного языка в рамках концепции реагирования / В.С. Фаин // Интеллектуальные процессы и их моделирование. – М.: Наука, 1987. – С. 375–391. 8. Miller G.A. WORDNET: A lexical database for English / G.A. Miller // Communications of ACM (11). – 1995. – P. 39–41. 9. Найханова Л. В. Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования / Л. В. Найханова. – Улан-Удэ: Изд-во БНЦ СО РАН, 2008. – 244 с. 10. Никитина С. Е. Семантический анализ языка науки / С. Е. Никитина. – М.: Наука, 1987. – 143 с. 11. Литвин В. В. Бази знань інтелектуальних систем підтримки прийняття рішень / В. В. Литвин. – Львів: Видавництво Львівської політехніки, 2011. – 240 с. 12. Дальберг И. Организация знаний: ее сфера и возможности / И. Дальберг // Организация знаний: проблемы и тенденции: Программа и тез. докл. конф. – М., 1993. – 14 с. 13. Мельников Г. П. Основы терминоведения / Г. П. Мельников. – М.: Изд-во ун-та дружбы народов, 1991. – 116 с. 14. Башмаков А.И. Интеллектуальные информационные технологии / А. И. Башмаков, И. А. Башмаков. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2005. – 304 с. 15. Беловольская Л. А. Синтаксис словосочетания и простого предложения [Електронний ресурс]. – Режим доступу: ttp://www.philology.ru/ linguistics2/ belovolskaya-01.htm. 16. Гладун В. П. Формирование тематических знаний на основе анализа ЕЯ текстов сети Интернет / В. П. Гладун и др. // Труды международной конференции Диалог'2003. – М.: Наука. – 2003. – С. 190–192. 17. Совпель И. В. Система автоматического извлечения знаний из текста и ее приложения / И. В. Совпель // Науч.-теорет. журнал "Искуственный интелект", ІПШІ "Наука і освіта". – 2004. – Вип. 3. – С. 668–677. 18. Гаврилова Т. А. Формирование прикладных онтологий / Т. А. Гаврилова // Труды XX нац. конф. по ИИ – КИИ-2006. – М.: Физматлит, 2006. – Т. 2. 19. Buchheit M. Decidable Reasoning in Terminological Knowledge Representation Systems / M. Buchheit, F. M. Donini, A. Schaerf // Journal of Artificial Intelligence Research. – 1993. – Vol. 1. – P. 109–138. 20. Мельников Г. П. Основы терминоведения / Г. П. Мельников. – М.: Изд-во ун-та дружбы народов, 1991. – 116 с. 21. Ершов А. П. Терминологический словарь по основам информатики и вычислительной технике / А. П. Ершов, Н. М. Шанский, А. П. Окунева, Н. В. Баско. – М.: Просвещение, 1991. – 159 с.