УДК 681.3

Анісімов А. В., Глибовець М. М., Кулябко П. П., Марченко О. О., Лиман К. С.

РОЗРОБКА МЕТОДІВ АВТОМАТИЗОВАНОГО РОЗШИРЕННЯ ТА ДОБУДОВИ ОНТОЛОГІЧНИХ БАЗ ЗНАНЬ

Досліджено методи автоматизованої добудови та розширення мереж онтологічних баз знань типу WordNet. Запропоновано метод додавання семантичних відношень-зв'язків між вузлами понять в онтології із використанням евристичних інструкцій, що генеруються системою в автоматизованому режимі.

Вступ

Онтологічні бази знань сьогодні є фундаментальною основою для розробки програм семантичного аналізу текстів природною мовою та багатьох інших сучасних систем штучного інтелекту. Процес побудови онтологічних баз знань як універсальних, так і спеціалізованих є дуже кропіткий та трудомісткий. Він вимагає участі висококваліфікованих фахівців у галузі лінгвістики, логіки, філософії, предметних областей та, звичайно, IT-спеціалістів. Розробка великих онтологій триває десятиліттями і дорого коштує. Тому актуальною стає проблема використання та інтеграції усіх створених онтологічних ресурсів для економії зусиль при розробці власних онтологічних систем штучного інтелекту. Доступні відкриті ресурси, як правило, створювались різними командами під власні цілі. Вони реалізовані з використанням різноманітних, іноді несумісних технологій та мов. Тому інтеграція та адаптація створених онтологій під власні задачі видається серйозною проблемою, яку потрібно розв'язувати з огляду на надмірну складність розробки онтологічних баз знань з нульового циклу.

Найскладнішою проблемою у використанні існуючих онтологій постає невідповідність специфіці конкретної задачі, коли в мережі бази знань не вистачає потрібних даних, наприклад, немає необхідних концептів-понять або зв'язків-відношень між ними. Серед загальнодоступних відкритих ресурсів сьогодні найбільш популярним є онтологічна глобальна база знань WordNet [1], яка містить більше ніж 120 тис. понять та семантичні відношення між ними. Але, звісно, має ряд незручних вад, серед яких, перш за все, мінімальний набір типів семантичних зв'язків між поняттями та недостатність міжкатегоріальних зв'язків (зв'язків між поняттями – різними частинами мови).

Стаття присвячена дослідженню проблеми автоматизації виведення семантичних відношень різного типу між концептами в онтологічних базах знань типу WordNet. В роботі запропонований один з підходів до розробки методів автоматизованої добудови та розширення онтологічної мережі WordNet.

1. Автоматизована добудова та розширення WordNet

Ще в середині 1980-х років Дж. Міллером та його колегами з Лабораторії когнітології Принстонського університету (США) була розроблена модель ментального лексикона людини. Проект, який став першою реалізованою глобальною онтологічною мережею, отримав назву WordNet [2] і з часом став одним з найбільш авторитетних та поширених стандартів, що використовуються для побудови лексико-семантичних баз онтологічного типу.

Популярність і широке розповсюдження WordNet зумовлені, передусім, його істотними змістовними і структурними характеристиками. Принстонський WordNet і всі наступні варіанти для інших мов спрямовані на відображення складу і структури лексичної системи мови в цілому, а не окремих тематичних областей. Теперішня версія WordNet охоплює загальновживану лексику сучасної англійської мови – більш ніж 120 тисяч слів.

Базовою структурною одиницею Принстонського WordNet є синонімічна множина (синсет), що об'єднує слова з тотожним семантичним значенням. Кожен синсет представляє в словникові деяке лексикалізоване поняття даної мови. Для зручності використання словника людиною кожний синсет доповнений дефініцією і прикладами вживання слів у контексті. Синсети в WordNet зв'язані між собою такими семантичними відношеннями, як гіпонімія (родовидове), меронімія (частина-ціле), лексичне виведення (каузація, пресупозиція) та ін.; серед них особливу роль відіграє гіпонімія (IS_A): вона дає змогу організовувати синсети в ієрархічні струк-

тури (дерева). Лексика кожної частини мови представлена у вигляді набору дерев (лісу). Для різних частин мови родовидові відношення можуть мати додаткові характеристики і розрізнятися областю розповсюдження.

Новим етапом в еволюції Wordnet був проект EuroWordNet, в рамках якого не тільки створено декілька тезаурусів для європейських мов (голландської, іспанської, італійської, німецької, французької), але і вперше була реалізована ідея про об'єднання окремих понятійних мереж у загальну систему. Всі компоненти EuroWordNet були побудовані за єдиною моделлю, що, однак, не означало прямого перекладу англомовного варіанта WordNet.

У рамках проекту EuroWordNet [3] первинна структура словника зазнала деяких змін. Був розширений набір семантичних відношень за рахунок парадигматичних відношень, що зв'язують слова різних частин мови і синтагматичних відношень між дієсловами й актантами-іменниками (наприклад, ROLE_INSTRUMENT: *to light – lamp, torch*).

Однак збагачення онтологічної мережі WordNet додатковими семантичними відношеннями між концептами-синсетами у більшості випадків здійснюється в ручному режимі лінгвістами-фахівцями, що робить подібні проекти дорогими.

Ця робота досліджує проблему розробки засобів автоматизації побудови додаткових семантичних відношень між концептами в онтологічних мережах лексико-семантичних баз знань WordNet. У цьому напрямі найбільш ефективними виглядають два підходи до збагачення онтологій додатковими семантичними відношеннями: знаходження нових зв'язків через існуючі за допомогою спеціальних правил виведення (наприклад, транзитивне виведення через відношення типу *частина-ціле*: листя *частина* гілки, гілка *частина* дерева => листя *частина* дерева; хоча і з деяким послабленням сили відношення *частина*); другим перспективним підходом є частотний аналіз корпусів текстів з пошуком випадків сталої кореляції між парами слів та словосполучень, що значно полегшує виявлення нових неврахованих відношень між синсетами і робить можливою побудову методів автоматизованого поповнення бази новими семантичними відношеннями.

Що стосується першого підходу поповнення бази відношеннями через виведення нових зв'язків, то ключовою ланкою при побудові цих методів є визначення і формування набору правил виведення нових семантичних відношень через існуючі та розробка методики їх застосувань.

У цій роботі розглянуто комбінований підхід. Він ґрунтується на двох гіпотезах.

По-перше, приймається гіпотеза, що слова у тексті, які пов'язані синтаксично, на семантичному рівні представляють сталий, але неявний вид зв'язку між семантичними значеннями цих слів. Наприклад, зв'язок ОБ'ЄКТ – ДІЯ (повний зв'язок – ОБ'ЄКТ – ДІЯ) у тексті може бути представлений таким чином: «Голуб летів»; ОБ'ЄКТ – голуб; ДІЯ – летів.

По-друге, припускається, що неявно існуюче відношення між концептами С1 та С2 в WordNet можна вивести із ланцюжка певної послідовності відношень, який веде від синсету С1 до синсету С2. Тобто якщо знати склад та конфігурацію такого ланцюжка, то можна при знаходженні аналогічних шляхів, що ведуть в мережі онтології від концепту А до концепту В, будувати пряме явне відношення відповідного типу. Постає питання побудови методів визначення таких ланцюжків.

Пропонується за допомогою частотного аналізу великих корпусів текстів виявляти сталі випадки кореляції слів у синтаксичних групах. Якщо знайдена кореляція між деякими лексемами, робиться спроба побудови нового семантичного відношення у мережі WordNet між синсетами, які містять корельовані лексеми. Побудова нового відношення полягає у пошуку найкоротших шляхів між даними синсетами у мережі WordNet. Якщо набір найкоротших шляхів побудовано, то знайдені ланцюжки послідовності відношень аналізуються далі експертами на логічну інтерпретованість, а саме: чи можна цю послідовність відношень замінити прямим відношенням між крайніми вершинами і як можна проінтерпретувати таке пряме відношення. Якщо процес інтерпретації певного ланцюжка послідовності відношень завершився успішно, тоді виконується етап верифікації цього евристичного виводу через статистичні експерименти. У мережі WordNet будуються всі можливі ланцюжки послідовності цього типу між різноманітними синсетами, а потім концепти, що з'єднані цими ланцюжками, зв'язуються прямим відношенням відповідного інтерпретованого типу. Далі експерти визначають відсоток коректно побудованих прямих відношень. Якщо цей показник відповідає критеріям поставленої задачі, послідовність відношень ланцюжка приймається як робоча евристика та використовується для побудови додаткових зв'язків – відношень у мережах онтології.

Очевидно, що деякі послідовності відношень під час аналізу ланцюжків у мережі мають сенс, якщо їх можна проінтерпретувати, то інші позбавлені обґрунтування, тому їх виключають із розгляду. Тобто мають бути сформовані множина P дозволених послідовностей відношень у ланцюжку між С1 та С2 та множина F заборone-

них послідовностей відношень у ланцюжку. Тоді процес знаходження нових відношень між синсетами в онтологічній мережі можна уявити як пошук шляхів-ланцюжків дозволеної послідовності між вершинами цих синсетів. Якщо ланцюжок дозволеної послідовності між С1 та С2 побудовано, то відповідне нове відношення між ними встановлено. Алгоритмічно можна представити процес пошуку та побудови послідовності ланцюжка як роботу автомата, в якому функцію переходів визначають множини P та F. Назвемо ці множини евристиками або інструкціями виведення нових семантичних зв'язків між концептами у мережі WordNet.

Тут аналізувалися відношення типу <іменник>–<прикметник>, котрі представляють зв'язок ОБ'ЄКТ – ЗНАЧЕННЯ АТРИБУТА (ОБ'ЄКТ – АТРИБУТ – ЗНАЧЕННЯ АТРИБУТА). Наприклад: «**Чорний кіт** спав»; ОБ'ЄКТ – кіт; ЗНАЧЕННЯ АТРИБУТА – чорний; АТРИБУТ – колір. Для з'єднання ланцюжком синсетів, сполучених цим зв'язком, треба більше часу, ніж для інших типів відношень, через те, що довжини таких ланцюжків дорівнюють приблизно 10 ланкам-відношень між синсетами. Тож використання побудованих інструкцій пришвидшують пошук в онтологічній мережі під час аналізу.

Для отримання пар <іменник>–<прикметник> використовувався семантико-синтаксичний аналізатор LINK GRAMMAR PARSER [4]. Після обробки корпусу текстів отримано множину пар слів, зв'язаних цим синтаксичним відношенням. Далі в онтологічній мережі будуються найкоротші шляхи між синсетами, які містять слова отриманих пар. Після побудови всіх можливих найкоротших ланцюжків між синсетами, генерується множина прямих відношень. Потім виконується етап вибору тих прямих відношень, що відповідають критерію інтерпретованості через аналіз послідовності відношень побудованого ланцюжка. У вдалих випадках встановлені послідовності приймалися як готові інструкції поповнення відношень онтології. В цій роботі будувалися інструкції дозволених переходів. Інструкції будувалися як дерева, у яких шлях від кореня до листків – дозволені ланцюжки.

Для побудови ланцюжків було використано алгоритм зустрічного пошуку в ширину на графах [5]. Тобто, алгоритм йде від двох вершин мережі назустріч пошуком у ширину, і вершина, де листя дерев пошуку перетинаються, належить ланцюжку від двох початкових вершин.

У WordNet можливі цикли і для того, щоб можна було застосувати пошук в ширину без втрати ланцюжків, треба «розрізати» цикли. При цьому одна вершина в стеку пошуку в ширину враховується двічі, бо ми додаємо до факторів унікальності вершини ще й декілька її батьків.

Розглянемо рисунок 1. Якщо до циклу а) застосувати звичайний алгоритм пошуку в ширину, то отримаємо шлях V1-V2-V4 (або V1-V3-V4). На рисунку 1б) зображено «розрізаний» цикл, який є на 1а), де V4* = V4**. У WordNet в більшості випадків є цикли, як зображені на рис. 1 а), тоді треба враховувати одного батька.

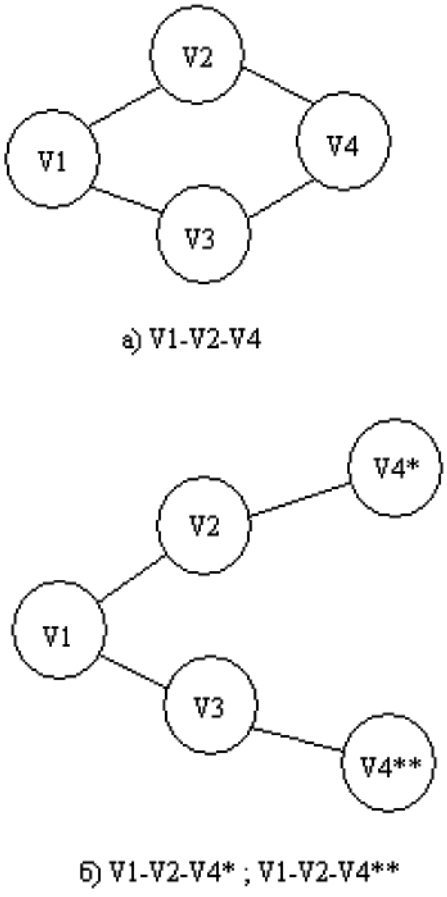

Рис. 1. Процес побудови ланцюжків шляхів у мережі WordNet

2. Результати

Було проведено статистичні експерименти з верифікації інструкцій евристичного виведення нових семантичних відношень між синсетами WordNet. У мережі онтології побудовано множину ланцюжків між різними синсетами згідно з отриманими інструкціями. Синсети, що зв'язувалися цими ланцюжками, сполучалися прямим зв'язком відповідного типу. Експертні оцінювання отриманих семантичних відношень встановили близько 87% коректно побудованих прямих зв'язків, що є достатньо високим показником і вказує на можливість ефективного використання знайдених евристик для автоматизова-

ного поповнення онтологічних баз знань типу WordNet новими семантичними відношеннями.

Треба також зазначити, що виявлення семантичних відношень не завжди є однозначним. Наприклад, пара слів <dog> і <black>, окрім стандартних значень цих слів (тварина собака і чорний колір), отримали під час пошуку найкоротших ланцюжків також інші значення:

<dog> – cad, bounder, blackguard, dog, hound, heel – (someone who is morally reprehensible) – грубіян, хтось, хто заслуговує на осуд;

<black> – black, disgraceful, ignominious, inglorious, opprobrious, shameful – ((used of conduct or character) deserving or bringing disgrace or shame) – ганебний, безчесний, заслуговує на немилість, сором чи ганьбу.

Коректні значення з'являлись під час тестування отриманих інструкцій набагато частіше, однак також потрапляли і результати, подібні до наведеного прикладу. Неприйнятні значення увійшли до результатів і відсіюватимуться у майбутній семантичній постобробці.

1. Miller G. Wordnet : An online lexical database / G. Miller // International Journal of Lexicography. – 1990. – №3 (4).
2. Miller, George A., Christiane Felbaum., J. Kegi, and K. Miller 1988. Wordnet : An electronic lexical reference system based on theories of lexical memory. 17. – P. 181–211.
3. Vossen P., L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, W. Peters. The EuroWordNet Base Concepts and Top Ontology. EuroWordNet (LE 4003) Deliverable D017, D034, D036. University of Amsterdam.
4. Temperley D, Lafferty J., Sleator D. 1995. Link Grammar Parser. – Режим доступу: http://www.link.cs.cmu.edu/link
5. Кормен Т., Алгоритмы : построение и аналіз / Т. Кормен, Р. Ривест. – Минск, 2005.

A. Anisimov, M. Glybovets, P. Kulyabko, O. Marchenko, K. Lyman

DEVELOPING METHODS FOR SEMI-AUTOMATIC EXTENDING AND REPLENISHMENT OF ONTOLOGICAL KNOWLEDGE BASES

Task of replenishment for WordNet type knowledgebases was investigated. Semi-automatic methods for adding correct links and relations between ontology elements were designed. These methods are based on fine heuristics and are quite robust.

УДК 004.4'23

Федорченко В. М.

БАЗА ЗНАНЬ КОМПОНЕНТІВ РЕПОЗИТОРІЯ NRECO

У роботі розглянуто основні аспекти побудови бази знань про програмні компоненти для репозиторія NReco. Визначено онтологію (RDFS/OWL) для опису компонентів різних рівнів абстракції з метою використання цих знань для організації ефективного пошуку в репозиторії. Особливу увагу зосереджено на видобуванні знань з XML-моделей та XSL-трансформацій; запропоновано евристичний механізм видобування знань про метамоделі XML-моделей в умовах відсутності повного формального визначення таких метамоделей.

Вступ

Сучасні підходи до побудови компонентної інфраструктури [1] та методи розробки програмних компонентів дають змогу забезпечити технічну можливість для їх ефективного повторного використання. На практиці ця можливість може бути реалізована лише за сприятливих умов, коли кожен програміст володіє повною інформацією про всі існуючі компоненти та способи їх використання. Сучасні платформи програмування, такі як Java чи Microsoft.NET, налічують тисячі стандартних компонентів, які добре задокументовані та загальновідомі; кількість до-