*пер. с англ. под ред. В.Л. Иноземцева. – М.: Academia, 1999. – 783 с. 11. The network of everything // Reseach*eu result supplement. – № 11 (January 2009). – P. 30. 12. Ваниянц С. В. Информационный ресурс в экономической сфере [Текст] / С.В.Ваниянц. – М.: Прогресс, 2006. – С.56–59. 13. Василенко В.А. Теорія і практика розробки управлінських рішень [Текст] / Василенко В.А. – К.: ЦУЛ, 2003. – 176 с. 14. Кедровская Л. Г. Номенклатура информационных услуг [Текст] / Л.Г. Кедровская, А. И. Мшвелидзе, Ю. Н. Ухин. – СПб.: ИПКИР, 2005.– 180 с. 15. Каган М.С. Избранные труды в VII томах. Том I. Проблемы методологии. Системное рассмотрение основных способов группировки [Текст]. – Санкт-Петербург: ИД "Петрополис", 2006. – 356 с. 16. Фрэнк Уэбстер. Теории информационного общества [Текст] / Фрэнк Уэбстер; пер. с англ. Е. Л. Вартановой. – М.: Аспект пресс, 2004. – 400 с.*

УДК 004.89

Р.В. Вовнянка[1], Д.Г. Досин[2], В.В. Ковалевич[2]
[1]Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж,
[2]Фізико-механічний інститут ім. Г.В. Карпенка

МЕТОД ВИДОБУВАННЯ ЗНАНЬ З ТЕКСТОВИХ ДОКУМЕНТІВ

Запропоновано метод, алгоритм і засоби для виділення знань з природномовного тексту. Показано, що такий алгоритм має бути багатоетапним і містити ієрархічну кількарівневу процедуру розпізнавання понять, зв'язків, предикатів та правил, які в результаті вносяться до онтології.

Ключові слова: онтологія, навчання онтологій, інтелектуальний агент, база знань, текстовий документ.

In the paper a method an algorithm and tools for selection of knowledge from a text document are suggested. It is shown that this algorithm has to be multistage and involve hierarchical procedure of concepts recognition of relations, predicates and rules which are introduced into the resulting ontology.

Key words: ontology, learning ontologies, intelligent agent, knowledge base, text document.

Вступ. Постановка проблеми у загальному вигляді

Поняття "знання" належить до галузі наукових досліджень методів і засобів прийняття оптимальних рішень. В процесі набуття знань через навчання суб'єкт прийняття рішень використовує доступну йому інформацію для побудови оптимальної стратегії прийняття рішень. Інформація у нашому розумінні набуває статусу знань саме тією мірою, якою вона допомагає носію цієї інформації вирішити його завдання і може бути числово оцінена як виграш від її використання для прийняття рішень у процесі досягнення відповідних цілей. Належно організовану і впорядковану сукупність знань інтелектуального агента називають базою знань. Система впорядкування знань у такій базі знань формально-логічно сформульована у її онтології, визначеній як "експліцитна специфікація концептуалізації" [1], тобто явне формальне означення понять і допустимих семантичних зв'язків між ними.

Суть методу видобування знань з природномовного текстового документа, інакше кажучи, розпізнавання змісту текстового документа, полягає у побудові плану (стратегії) діяльності інтелектуального агента – інформаційної моделі суб'єкта розпізнавання або уточнення такого плану на підставі даних, виділених у текстовому документі, що розпізнається. Тут вважаємо план конкретною реалізацією оптимальної стратегії розв'язання деякої задачі, що стоїть перед інтелектуальним агентом у межах заданої проблемної області.

План будується тією формальною мовою подання знань, якою розроблено інформаційну модель – базу знань інтелектуального агента. Оскільки така база знань вже являє собою певний загальний план функціонування інтелектуального агента, план, збудований на основі розпізнавання змісту природномовного тексту, є субпланом, тобто уточненням (виправленням) і/або деталізацією цього загального плану і ґрунтується на ньому. Цінність інформації, отриманої внаслідок розпізнавання змісту текстового документа, визначається за приростом очікуваної корисності від реалізації уточненого у такий спосіб плану функціонування інтелектуального агента.

Переважна частина доступної, сформульованої у певній логічній послідовності й тому зручної для опрацювання інформації зберігається у текстових документах, зокрема на електронних носіях. Достатньо велика частина таких документів доступна on-line, до того ж безоплатно. Серед них є можливість вибрати такі, що написані за достатньо жорстко встановленими правилами побудови і вимогами до змісту так, що, з одного боку, вони залишаються природномовними текстами, а з іншого – максимально формалізовані для їх машинного опрацювання і виділення релевантної інформації, яка може інтерпретуватися інтелектуальною системою розпізнавання змісту як корисні знання.

До такого специфічного класу природномовних текстів можна зарахувати анотації наукових статей. Їх можна знайти через мережу Інтернет, вони, як правило, є у відкритому доступі, не містять графічного матеріалу, побудовані за строго встановленими правилами, написані, окрім інших, також англійською мовою, не містять модальних зворотів, а лише логічно зв’язану послідовність стверджувальних речень. Необхідний для заданої проблемної області (ПО) корпус таких текстів можна вибрати за допомогою інформаційного пошуку за ключовими словами з використанням цілої низки як спеціалізованих пошукових серверів наукових видавництв, так і пошукових серверів загального призначення.

Аналіз останніх досліджень та публікацій

Напрями наукових досліджень та розробки в галузі навчання та наповнення онтологій наведено у табл. 1 [2–4]. Для лаконічності викладення в таблиці використано такі умовні позначення:

- A означає, що вирішення цієї проблеми в проекті ще немає;
- B – для напівавтоматичного методу, реалізованого у проекті;
- C – для реалізацій, в яких участь людини необов’язкова;
- D – для апріорі автоматичних методів, що не передбачають участі людини.

Таблиця 1

Розробки в галузі розроблення і наповнення онтологій

Назва	Метод виділення	Аналіз	Генерування	Апробація	Розвиток
1	2	3	4	5	6
Генерування онтології з бізнес-моделі	Людиною	–	C – Без об’єднання. Пряме перетворення з використанням XSLT	– Людиною, знизу вгору	–
XML2OWL	B – статична таблиця відповідностей	–	C – Без об’єднання. Пряме перетворення з використанням XSLT	– Людиною, знизу вгору	–
UML2OWL	B	–	C – Без об’єднання. Пряме перетворення з використанням XSLT	– Людиною, знизу вгору	–

Продовження табл. 1

1	2	3	4	5	6
Напівавтоматична побудова онтології зі структур DTD	C – автоматичне виділення з джерел у форматі DTD	B – аналіз структури	C – не-стандартне представлення онтології	Людиною	–
Learning OWL ontologies from free texts	C – текстові джерела, методи NLP. WordNet онтологія/словник	–	C – формат OWL	–	–
Побудова онтології для вибору інформації	C	–	C	–	–
TERMINAE	C – текстові джерела, методи NLP.	B – аналіз взаємозв'язків між поняттями	C – нестан-дартне представлення онтології	Людиною	–
SALT	D – текстові джерела, методи NLP. За кількома джерелами	C – аналіз подібності між поняттями	B – Нестандартне представлення онтології	B – обмежене втручання людини	–
Новий метод злиття онтологій за концептами з використанням WordNet	–	B	C – авто-матичне поєднання. Нестандартне представлення онтології	–	–
Розроблення системи автоматичної побудови онтології для заданої ПО	B – головні поняття визначає експерт у цій ПО	–	C	–	–
Наповнення надвеликих онтологій з використанням WWW	C – розширення наявної онтології	–	C	–	–
Видобування знань заданої ПО та їх впорядкування (класифікація з використанням WordNet	C – головні поняття визначає експерт у цій ПО	B – гра-матичний аналіз тексту	C	Людиною	–
A метод напівавтоматичного наповнення онтології з Intranet	C – методи NLP. Багатократне опрацювання джерел	B – аналіз значення понять	B	B – у конфліктних випадках потрібна участь користувача	B – циклічне застосування може забезпечити еволюцію
SymOntoX	-	C – аналіз збігів	B – забез-печення деяких наперед заданих базових понять	Людиною	B – керування версіями, проте за участю людини
Protégé (з використанням відповідних плагінів-додатків)	B – виділення з реляційної бази даних та даних у XML форматі	D – аналіз збігу і відповідності	B – кероване зв'язування. Експорт у різні формати онтології	Людиною	C – розпізнавання еволюції онтології

Закінчення табл. 1

1	2	3	4	5	6
LOGS	C – аналіз текстових джерел. Механізм NLP. Морфологічний і семантичний аналіз. Машинне навчання правил.	C – подібність ґрунтується на аналізі понять і зв'язків	C – Різні формати. Внутрішня структура онтології матрична	B – перевірка в кінці кожного модуля	–
Навчання онтології	D – виділення з різних форматів (XML, UML, OWL, RDF, text…). NLP, Семантичний і лексичний аналіз. Багатовходовий аналіз джерел	C – бібліотеки для кластеризації, формального аналізу понять та асоціативних правил	C-OWL та RDF/S	B – за підтримки людини	–

Застосовується кілька основних підходів до опрацювання тексту з цією метою – символьний, статистичний та змішаний. Серед найпоширеніших символьних підходів – застосування лексико-семантичних патернів (lexico-semantic pattern – LSP) [2]. У такому підході опрацювання тексту виконується шляхом виявлення певних наперед відомих або встановлених за допомогою машинного навчання реляційних маркерів, які існують у природній мові й дають змогу розпізнати семантичні ролі синтаксичних конструкцій, а у поєднанні з ідентифікацією онтологічних сутностей, які у цьому тексті представляють ці синтаксичні конструкції, виконувати проекцію тексту на онтологію, отримуючи розпізнаний зміст, а згідно з ним – оцінювати новизну, достовірність і корисність отриманих за цим змістом знань. Методи, що базуються лише на статистичних лінгвістичних моделях, здатні тільки поверхнево розпізнавати дискурс, але не можуть виявляти зміст тексту, тобто відображену там логіку семантичного взаємозв'язку між поняттями цієї проблемної області.

Формування цілей

Розробити метод, алгоритм та програмні засоби для виділення знань з природномовного тексту.

Основний матеріал

1. Постановка задачі

Задачу вибору потрібного корпусу текстів розв'язали, реалізуючи в межах цієї науково-дослідної роботи підсистеми інформаційного пошуку програмного пакета CROCUS [5]. На вході підсистеми – множина ключових слів, на виході – множина англомовних анотацій, розміщених у базі даних СУБД MySQL.

Процес видобування знань передбачає здатність як до розпізнавання окремих понять, згаданих у документі, так і до логічної інтерпретації сутності й характеру зв'язків між цими поняттями. Ці дані слугують лише первинною інформацією для ієрархічної, багатоетапної процедури розпізнавання змісту природномовного текстового документа (ПТД). На відміну від традиційних статистичних методів опрацювання ПТД, у яких текст розглядається як множина окремих термінів (слів та словосполучень) без врахування семантичного взаємозв'язку як між термінами, так і між цілими твердженнями, вираженими закінченими реченнями, запропонована і розроблена у цій роботі процедура ґрунтується на розпізнаванні логічних тверджень і тому складається з трьох основних етапів: лінгвістичного, статистично-логічного та планувального. На першому, лінгвістичному етапі засобами морфологічно-синтаксичного аналізу мови, якою цей текст написано, будується послідовність триплетів "суб'єкт зв'язку – семантичний зв'язок – об'єкт зв'язку", кожен елемент яких знаходиться або по ходу аналізу додається до онтології інтелектуального агента. На другому етапі методами машинного навчання на основі отриманої послідовності триплетів розпізнаються твердження у логіці предикатів першого порядку, їх семантичний зміст у термінах

онтології інтелектуального агента та логічний взаємозв'язок між ними. На третьому, заключному етапі на базі прототипу плану або діючого загального плану функціонування інтелектуального агента з отриманої послідовності предикатів будується (доповнюється, коригується) ієрархічна система цілей (задач) і засобів їх досягнення (розв'язання).

По суті, маємо ієрархію розпізнавання: окремі слова, далі – словосполучення, далі зв'язки, далі – твердження, які вже являють собою базовий елемент, цеглини моделі світу інтелектуального агента. Далі можна говорити про розуміння агентом відмінностей між різними моделями світу: своєї і чужої, автора повідомлення, що аналізує цей агент.

Загальна схема реалізації методу видобування знань з тексту передбачає такі кроки:

1. Вибираємо прототип онтології як OWL-модель контексту ПО.
2. Перетворюємо аналізований текст на множину речень. Якщо джерелом тексту є анотація наукової публікації у друкованому виданні, першим реченням множини додаємо назву публікації. Останнім – назву друкованого видання.
3. В циклі розбираємо послідовно усі речення множини і будуємо з кожного з них окрему множину пар слів, з'єднаних метасемантичним зв'язком, яка слугуватиме вектором ознак для розпізнавання виду семантичного зв'язку.
4. Окремо з речення виділяємо групу іменника – суб'єкт розпізнаного на попередньому кроці семантичного зв'язку та групу іменника – об'єкт цього зв'язку.
5. До створеного на першому кроці шаблону онтології додаємо поняття, які вдається розпізнати в групах іменників, отриманих на попередньому кроці. Поняття додаються як екземпляри відповідних класів.
6. Якщо онтологія містить і об'єкт і суб'єкт зв'язку, тоді між ними встановлюється виявлений зв'язок. Одночасно до бази знань додається предикат, що відповідає цьому зв'язку.
7. Для визначеної у п.2 множини речень та відповідної їй множини предикатів розпізнаємо логічні залежності між предикатами. Виявлені залежності вносимо до бази знань у формі SWRL-правил.
8. Під час внесення нового правила перевіряємо базу правил на наявність суперечностей. Конфлікти вирішуємо з урахуванням достовірності джерел інформації, за якими внесені предикати, які конфліктують, а також логічної залежності з іншими предикатами бази знань.
9. Отриману систему понять і зв'язків, збудовану на їх основі систему предикатів та функцій, а також збудовану на їх основі систему аксіом і правил використовуємо для побудови плану інтелектуального агента.
10. Задаємо, уточнюємо або визначаємо за виявленими предикатами винагороди за досягнення проміжних цілей плану, імовірність їх досягнення у разі вчинення допустимих дій, а також затрати на виконання цих дій. Розраховуємо оптимальний план, його очікувану корисність.
11. Процес навчання онтології полягає у послідовному (або паралельному) повторенні цієї процедури для всього корпусу навчальних текстів.

Отриманий план інтелектуального агента слугує інформаційною моделлю публікації з погляду цілей і задач її потенційного читача.

2. Вибір прототипу онтології заданої проблемної області

Знання набувають змісту лише в контексті певної проблемної області (ПО), заданої у цьому випадку її онтологією. Набуті з текстового документа нові знання набувають форми змін у первинній онтології, яку слід попередньо сформувати вручну або із застосуванням процедур навчання. Аналіз кожного наступного тексту оснований на застосуванні онтології, доповненої в процесі аналізу попередніх текстів у тій частині, яка стосувалася заданої проблемної області. Для розпізнавання змісту текстів та доповнення онтології ПО ключовим є підхід, за якого першочергово необхідно виявити засоби досягнення мети, рекурсивно призначаючи їх підцілями і шукаючи як у тексті, так і у самій онтології (відповідній цій онтології базі знань) засоби досягнення цих підцілей. Тобто, читаючи, агент будує на основі прочитаного тексту дерево цілей для задачі, розв'язок якої він шукає. У тексті засоби можуть бути формально ідентифіковані як іменникові групи, що стоять за дієсловом 'using', зворотом 'by means' або іншими подібними характерними зворотами,

ідентифікувати які система розпізнавання змісту може навчитися засобами машинного навчання. У зв'язку з цим онтологія інтелектуального агента має містити у своїй основі як на верхньому рівні, так і на рівні прикладних проблем (задач) дерево цілей цієї ПО та відповідне йому дерево рішень.

Для онтології матеріалознавства прототипом дерева цілей можуть слугувати діаграми, які розробила І.В.Федорович у дисертаційній роботі [6].

Ефективність прийняття рішення в будь-якій проблемній області можна визначити як відношення виграшу внаслідок рішення (послідовності рішень) до затрат чи втрат, пов'язаних з прийняттям (неприйняттям) цих рішень. Структура втрат для проблемної області "протикорозійний захист магістральних трубопроводів" подана на рис 1.

Рис. 1. Структура втрат для ПО "протикорозійний захист магістральних трубопроводів"

У цій роботі [6] з метою визначення стратегії і тактики дій керівників газотранспортних підприємств у сфері відновлення основних засобів та сприяння пошуку ефективних методів забезпечення стабільності та ефективності внутрішнього механізму відтворення лінійної частини магістральних газопроводів виділено основні чинники впливу. З урахуванням можливості контролю лише внутрішніх чинників їх розглянуто детальніше й наведено на рис. 2.

Сукупність виділених чинників за результатами їх експертної оцінки формує структуру імовірностей досягнення успіху (його максимізації) внаслідок прийняття рішення стосовно відновлення та модернізації газотранспортної системи. Розподіл відмов та аварій на лінійній частині магістральних газопроводів залежно від причин їх виникнення показав, що основною причиною (54 %) є корозія металу труби. За результатами аналітичного оцінювання виявлено, що найбільше впливають такі чинники, як:

- якість виконання робіт з будівництва газопроводів;

Рис. 2. Внутрішні чинники ефективного відтворення нафтогазотранспортної системи

- якість ремонтного обслуговування;
- рівень придатності ізоляційного покриття газопроводу;
- рівень корозійного руйнування газопроводу;
- рівень кваліфікації робітників-ремонтників та інженерно-технічних працівників;
- рівень досконалості прийняття управлінських рішень у процесі відновлення магістральних газопроводів.

Усі ці чинники мають відображатись в онтології, а імовірність їх впливу відображена у дереві рішень цієї проблемної області. В результаті під час аналізу природномовного тексту пошук нових знань передбачає не лише пошук підцілей як проміжних розв'язань головної задачі, а і уточнення впливу окремих чинників на імовірність досягнення відповідних підцілей і, як результат, головної цілі. Так уточнюється загальна інформаційна модель проблемної області, що дає змогу ефективніше використовувати ресурси, приймаючи кращі рішення і, отже, оцінювати цінність нових знань через очікувану економію ресурсів.

Для проблеми модернізації трубопроводу загальна задача виглядає, як зображено на рис. 3. Початковий стан: *Необроблена*. Кінцевий стан (стан мети): *Оброблена*.

Задача ділиться на три підзадачі (підготовка, покриття, захист), перша з яких ділиться ще на чотири підзадачі (розкриття поверхні труби, зняття захисного покриття, знежирення, ґрунтування), як показано на рис. 4. Для розв'язування кожної підзадачі використовуються альтернативні

рішення. Так, для підзадачі “зняття захисного покриття” можна використати одну із трьох альтернатив: механічне, хімічне, термічне. Вся ця інформація зберігається у онтології ПО модернізації нафто- та газопроводів [7].

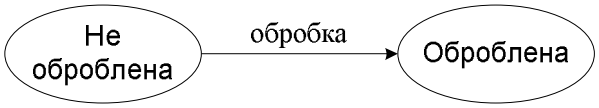

Рис. 3. Загальна задача модернізації трубопроводу

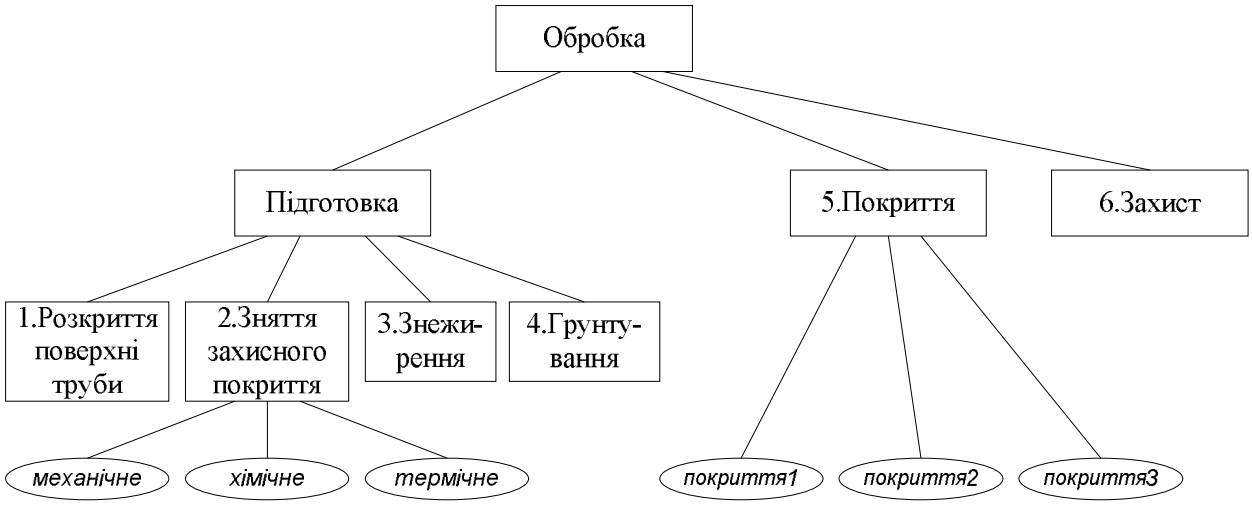

Рис. 4. Декомпозиція задачі “Обробка”

Отже, загалом необхідно послідовно розв’язати шість підзадач $P_1, P_2, ..., P_6$. Для кожної задачі необхідно вибрати метод розв’язання (альтернативу). Якщо G – наявний ресурс, r_e – бажаний термін експлуатації трубопроводу, то раціональність прийняття рішень полягатиме в:

$$\begin{cases} U = \sum_{i=0}^{N-1} U\left(a_{ij}^k\right) \to \max, \\ r \geq r_e, \\ \sum_{i=0}^{N-1} g_{ij}^k \leq G. \end{cases}$$

Детальніше цю математичну модель описано у [5].

3. Виділення формальних ознак семантичних зв’язків між поняттями у реченні

Для формального подання природномовного речення у термінах онтології та описової (дескриптивної) логіки предикатів першого порядку необхідно визначити тип предиката. Розпізнавання типу можна виконати за дієслівною групою цього речення та службовими словами, які до дієслівної групи можуть не входити. Для цього необхідно застосувати метод машинного навчання системи розпізнавання, вхідними даними для якої будуть результати розбору природномовного речення спеціальним синтаксично-семантичним парсером. Такий парсер розбиває речення на пари слів, пов’язані деяким метасемантичним зв’язком. У результаті кожне речення парсер подає множиною триплетів, що складаються з суб’єкта такого зв’язку, об’єкта зв’язку і самого метасемантичного зв’язку певного виду. Ці триплети можуть бути використані як ознаки наявності в реченні того чи іншого семантичного зв’язку, на основі якого має бути збудований предикат як логічне формальне представлення цього речення.

У роботі ми використали Link Grammar Parser (далі – LGP) [8]. Цей програмний засіб є ’open source’-продуктом, має відкриту ліцензію типу GPL, добре документований, а тому доцільність його застосування для цієї задачі не викликає сумнівів. Приклад вікна з довідковою інформацією та

ілюстрацією результатів розбору простого речення під час роботи з програмою командного рядка наведено на рис. 5.

3.7. COMMANDS AND VARIABLES. It is possible to modify the running of the parser in various ways, while running it, by typing in certain commands. The basic commands can be seen by typing "!help". Others are listed under "!variables". Many of these are self-explanatory. For example, "!width" changes the width of the parser display. Other commands relate to speed and robstness features; see section 7.

A few commands deserve special mention. One useful command is "!![word]". This queries the parser for information about a particular word. The parser will output list any entries of the word, with their word subscripts, the word-files in which they appear, if any, and the number of disjuncts on each word. (A disjunct is a combination of connectors which constitutes a legal use of the word.) Multiple entries of a word will be listed with their word subscripts.

The "!verbosity" command controls the amount of information that is displayed. With "!verbosity=1" (the default), information such as the following is shown:

```
linkparser> the quick brown fox jumped over the lazy dog
++++Time                                          0.04 seconds (0.04
total)
Found 2 linkages (2 had no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=18)
 +---------Ds---------+             +-------Js------+
 |     +-------A------+             |   +-----Ds----+
 |     |       +---A--+---Ss--+--MVp-+   |     +--A--+
 |     |       |      |       |      |   |     |     |
the quick.a brown.a fox.n jumped.v over the lazy.a dog.n
Press RETURN for the next linkage.
linkparser>
```

With "verbosity=0", no information is shown except for the graphic linkage display. With verbosity set at 2 or 3, information is shown about the individual stages of parsing the sentence. (Infomation is also shown about the constituent derivation process, if this is being done.) If one wants to suppress the graphic display as well, this can be done with the command "!graphics". (This can be useful it one wants to have only the constituent bracketing as output; in that case, type "!verbosity=0", "!graphics", and "!constituents=1 (or 2)".

Рис. 5. Довідка ***Link Grammar Parser*** *з прикладом результатів розбору простого речення*

Передумовою виявлення семантичного зв'язку із застосуванням LGP є наявність (і виявлення) дієслівної групи. Її наявність визначається відповідними дієслівній групі великими і малими буквами з достатньо великого переліку. Наприклад, усі зв'язки, розташовані праворуч від зв'язку 'S*' вказують на дієслівну групу.

Слова, що супроводжують (на які вказують) ці букви-символи зв'язку, визначають вид семантичного зв'язку (ім'я предиката). Ці слова є, як правило, дієсловами: "знає", "має", "належить", "відноситься", або дієслівними словосполученнями " належить до", "складається з" тощо. Прикметники інтерпретуються як властивості й також можуть бути розпізнані у реченнях через семантичний зв'язок "має властивість".

Щоб розпізнати семантичний зв'язок, необхідно виконати такі дії:

- розібрати речення за допомогою LGP;
- знайти дієслівну групу через символи зв'язку праворуч від "Ss";
- знайти дієслово, на яке вказують ці символи зв'язку;
- знайти суб'єкт дії (підмет у реченні), на який вказує символ "Ss";
- знайти об'єкт дії (очевидно, означення у реченні), тобто предмет, на який спрямована дія;
- перевірити в онтології наявність цього виду семантичного зв'язку і у разі відсутності, створити його;
- перевірити наявність в онтології сутностей, що означають об'єкт та суб'єкт дії. Тут можливі різні варіанти (див. табл. 2):

У випадку (4), очевидно, речення ігнорується. У випадку (3) також нічого пов'язувати: так, наприклад, якщо використовується відомий онтології зв'язок IS-A, який пов'язує два не відомі онтології терміни: X та Y, таке твердження також доведеться ігнорувати. У випадках (2) і (5) з'являється можливість внести поняття до онтології через відомий зв'язок, наприклад, у реченні "Іван читає X" X вноситься до онтології як сутність, яку можна курити. При цьому слід розрізняти онтологію як множину допустимих семантичних зв'язків між поняттями і базу знань як множину фактів про цю модель дійсності. Онтологія описує зв'язки між класами, а база знань – зв'язки між екземплярами цих класів.

Таблиця 2

Варіанти опрацювання подій системою CROCUS

№	Є зв'язок	Є суб'єкт	Є об'єкт	Можлива дія
1	+	+	+	Внести до бази знань
2	+	+	–	Додати невідоме поняття до онтології
3	+	–	–	Ігнорувати
4	–	–	–	Ігнорувати
5	+	–	+	Додати невідоме поняття до онтології

Зв'язки можуть бути безумовними та умовними. Умовні зв'язки записуються як правила. Безумовні зв'язки є частковим випадком умовних і записуються як факти, у вигляді предикатів.

Отже, для навчання системи навичкам розпізнавання нових типів семантичних зв'язків у реченнях потрібний модуль індуктивного навчання за семантичними ознаками. Речення-приклад дає послідовність семантичних зв'язків між словами. Кілька таких однотипних речень підряд з вказанням назви зв'язку дають системі можливість виявити підмножину спільних ознак і створити ознакову функцію: {*Vj*}=>Linkx, де *Vj* – *j*-та ознака у вигляді:

```
organ->S->is
is->O->part
a->D->part
part->M->of
of->J->organism
an->D->organism
```

Вхідними даними для модуля індуктивного навчання слугують змінні – необмежена множина слів і константи – обмежена множина символів граматичних зв'язків {S, D, O, J, M, ...}. Маємо також результат роботи LGP – пари слів, поєднані метасемантичними зв'язками у певній послідовності, маємо множину дієслів, кожне з якої може стати початком координат в реченні в разі виявлення.

Виявлення семантичних зв'язків у лінгвістичній підсистемі CROCUS побудовано на застосуванні байєсівського розпізнавання множини ознак збережених в онтології патернів відомих семантичних зв'язків. Вивчення d ознак j-го семантичного зв'язку:

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p\left(X_k \mid C_j\right). \quad (1)$$

Розпізнавання *j*-го семантичного зв'язку за *d* виявленими ознаками:

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p\left(X_k \mid C_j\right). \quad (2)$$

Як ознаки (дескриптори) використано результат розбору речення природномовного тексту на пари слів, пов'язаних синтаксично-метасемантичними зв'язками за допомогою LGP. Для простого тестового речення:

```
[(a)(test.n)(is.v)(an)(example.n)]
```

результат розбору:

```
[[0 1 0 (Ds)][1 2 0 (Ss)][2 4 0 (Ost)][3 4 0 (Ds)]]
```

результат розпізнавання типу семантичного зв'язку за (2):

```
1) cause: 1.0882684165532656E-4;
2) caused-by: 0.013810506200916856;
3) is-a: 0.024124901979118252;
4) is-about: 0.0;
5) part-of: 0.0022765542079946285;
6) same-as: 0.0;
7) similar-to: 1.0261341731138478E-6;
```

Тестування розроблених програмних засобів, що реалізують описаний вище алгоритм, підтверджує коректність його роботи.

Висновки

Отже, проаналізовано стан досліджень та розробок у галузі видобування знань з природномовних текстів та машинного навчання онтології інтелектуального агента. Обґрунтовано необхідність покласти в основу структури онтології план оптимального функціонування такого агента у заданій проблемній області. На цій основі запропоновано оцінювати цінність нових знань, виділених з природномовного тексту, за змінами такого плану, які необхідно вносити у план, щоб зберегти стратегію його виконання оптимальною з урахуванням цих знань. Для цього необхідно обчислювати поточну очікувану корисність від реалізації оптимальної стратегії до і після внесення до плану нових знань. На прикладі проблеми модернізації газотранспортної системи показано схему побудови такого плану для закладення загальної структури понять та зв'язків відповідної онтології цієї проблемної області. Запропоновано загальний алгоритм, необхідні методи і засоби для виділення нових знань з природномовного тексту, показано, що такий алгоритм має бути багатоетапним і містити ієрархічну кількарівневу процедуру розпізнавання понять, зв'язків, предикатів та правил, які в результаті вносяться до онтології з метою виконання перерахунку очікуваної корисності. Сформована у такий спосіб онтологія нижнього рівня може слугувати точною моделлю інформаційних потреб користувача системи інформаційного пошуку, необхідною для автономного пошуку чи моніторингу.

1. Gruber T. A translation approach to portable ontologies / T.Gruber // Knowledge Acquisition. – 1993. – № 5 (2). – P. 199–220. 2. Інтелектуальні системи, базовані на онтологіях // Д.Г. Досин, В.В. Литвин, Ю.В. Нікольський, В.В. Пасічник. – Львів: Цивілізація, 2009. – 414 с. 3. Agirre E. (2000). Enriching very large ontologies using the WWW / E.Agirre, O.Ansa, E.Hovy, D.Martinez // In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00). – 2000. – P. 347–349. 4. Alfonseca E. Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures / E.Alfonseca, S.Manandhar // EKAW-2002, Siguenza, Spain. Published in Lecture Notes in Artificial Intelligence. – 2002. – P. 2473 (Springer Verlag). 5. Литвин В.В. Бази знань інтелектуальних систем підтримки прийняття рішень / В.В.Литвин. – Львів: Видавництво Львівської політехніки, 2011. – 240 с. 6. Федорович І.В., Організаційно-економічне забезпечення процесу відтворення лінійної частини магістральних газопроводів: автореф. дис. … канд. екон. наук, Івано-Франківський національний технічний університет нафти і газу, Івано-Франківськ, 2011. 7. Досин Д.Г. Архітектура інтелектуальної системи інформаційного пошуку в мережі Інтернет / Д.Г. Досин, В.М. Ковалевич // Штучний інтелект. – 2012. – № 3. – С. 241–252. 8. Daniel Sleator and Davy Temperley. 1991. Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.