УДК 004.4'414(045)

ТРАНСФОРМАЦІЯ ТЕКСТОВОЇ ПРИРОДНО-МОВНОЇ ІНФОРМАЦІЇ НА ОСНОВІ ДІЄСЛІВНИХ СЕМАНТИЧНИХ ВУЗЛІВ

М. Ю. Сич

Національний авіаційний університет

m_sych@ukr.net

У статті описано механізм трансформації текстової інформації на природній мові з метою подолання проблеми неспроможності комп'ютерних систем визначити тотожність різних словарних конструкцій з однаковим загальним змістом, але різним змістом окремих слів. Трансформацію проведедено на основі шаблонів дієслівних семантичних вузлів шляхом накладання їх на фактичну природно-мовну текстову інформацію.

Ключові слова: природно-мовна інформація, трансформація, семантична мережа, семантичний вузол.

In the article described mechanism of text information on natural language transformation for purposes of processing by computer systems word constructions with different particular words senses but the same common sens. Transformation is based on the set of semantic nodes patterns, which can be lay on actual natural language text information.

Key words: natural language information, transformation, semantic network, semantic node.

Вступ

Однією з ключових проблем створення автоматизованих систем пошуку текстової інформації на природній мові є те, що такі системи нині не здатні коректно та повно обробити суть глибинний зміст самих констатованих у інформації фактів.

Зміст одного єдиного факту може бути переданий різними словарними конструкціями, які мають різні смисли окремих слів, але однаковий загальний сенс. При обробці вхідного запиту користувача, що також надходить на природній мові, пошукові системи не здатні обробити цю різноманітність у описі подій, і тому навіть наявна в системі інформація знайдена не буде.

Описану проблему назвемо *проблемою констатацій факту*.

Вирішення даної проблеми дозволить значно підвищити ефективність роботи комп'ютерів з інформацією на природній мові не тільки у сфері пошуку інформації, а й багатьох інших.

Модель представлення знань

Як модель знань оберемо семантичну мережу.

Семантична мережа — це по суті орієнтований граф, вершини якого відображають деякі поняття, а дуги — відношення між ними.

Таким чином, семантичні мережі відтворюють семантику предметної області у вигляді понять та відношень.

Семантичні мережі були розроблені Річардом Річенсом у 1956 р. у рамках проекту Кембриджського центру вивчення мови для машинного перекладу.

Для всіх семантичних мереж справедливим є розподіл за арністю і кількістю типів відношень.

За кількістю типів мережі можуть бути *однорідними* і *неоднорідними*.

Однорідні мережі мають лише один тип відношень, *неоднорідні* — більше одного.

За арністю типовими є мережі з бінарними відношеннями, тобто пов'язують рівно два поняття; такі мережі найбільш зручні у використанні. Існують також відношення, що пов'язують більше двох понять — N-арні [1].

У даній роботі використаємо неоднорідну N-арну семантичну мережу. Множина значень вершин — уся сукупність словоформ певної мови. Назвемо цю множину Words.

Множина значень дуг (зв'язків) — сукупність граматичних властивостей словоформ у вигляді питань: хто?, що?, який?, яка?, де?, коли?, скільки? і т.п. Цю множину назвемо Questions. Маємо:

$$Words = \{Word_i\}$$

$$Questions = \{Question_i\}$$

Усі словоформи розділені на кілька груп за морфологічними, граматичними та семантичними властивостями.

У таблиці наведено ці групи з відповідними зв'язками-питаннями.

Перевагами такого способу задання множин значень для вершин та дуг можна вважати їх визначеність та остаточну кількість.

Хоча словоформи і можуть додаватись час від часу у словник, але набір зв'язків між ним залишається стабільним та постійним, лише дуже незначно модифікуючись при доданні у словник словоформ, скажімо, нової мови з її словами.

Механізм автоматичної трансформації текстової інформації у цю модель знань, тобто фактично алгоритм синтаксичного та семантичного аналізу текстової інформації, описаний у [2]. Тут ми для наочності приведемо приклад.

Групи сутностей словоформ

№ з/п	Група	Множина зв'язків-питань
1	Об'єкти	Хто?, що? кого?, чого?, кому?, чому? і т.д.
2	Атрибути об'єктів	Який?, яка?, яке?, які?, якого?, якої?, яких? і т.д.
3	Дії	Що робити?, що роблять?, що робите? і т.д.
4	Атрибути дій	Як?, де?, коли?, куди?, звідки? і т.д.
5	Числа	Скільки?, котрий? і т.д.
6	Частки	Від чого?, звідки?, до чого?, куди? і т.д.

Візьмемо речення «Вузівські підручники мають відрізнятися від шкільних теоретичним рівнем та науковістю».

На рис. 1 графічно зображено модель, що відповідає наведеному реченню.

Шаблон на дієслівному семантичному вузлі

Частково вирішити проблему констатацій фактів можливо, якщо створити набір шаблонів словарних конструкцій, що будуть містити в собі еквіваленти різних формулювань фактів.

У даному випадку ми будемо говорити про шаблони з основою на словах, що позначають дії.

Дієслівним семантичним вузлом будемо називати вершину семантичної мережі, яка містить смисл слова, що є певною дією, тобто фактично частина мови слова у вершині є дієслово.

Під шаблоном будемо розуміти граф семантичної мережі, що відображає характер взаємодії слів, пов'язаних певним чином.

Наприклад, якщо «хтось народився в певному місці», то «певне місце є батьківщиною когось».

Таким чином, створюється шаблон з основою «народитися», із вказаними взаємозв'язками з об'єктом дії та місцем дії, у якому вказано, що місце дії є батьківщиною носія дії.

Частина вершин шаблону містить смисли слів — «народитися» та «батьківщина», а частина лише характерні ознаки сутностей та характер їх взаємодії, таких як: об'єкт дії є певною сутністю, а місце дії — певним смислом, що позначає місто, країну, тощо. Таким чином створюється образ певного положення речей у моделі навколишнього світу. Графічно шаблон у розкритому вище сенсі можна подати як показано на рис. 2.

Частину шаблону, об'єднану певним змістовним навантаженням, будемо називати *гілкою шаблону*. Таким чином, маємо в шаблоні на рис. 2 дві гілки. Вершини 1, 2, 3, 4 належать до однієї частини шаблону і являють собою гілку шаблону 1, вершини 5, 6, 7, 8 — гілку шаблону 2. На рис. 2 для зручності перша гілка позначена суцільними лініями, друга — пунктиром.

Як уже було сказано, частина вершин шаблону має чіткі смисли, це вершини 2, 3, 7, 8.

Інша група вершин містить у собі лише ознаки сутності, що має бути приписана певній вершині у процесі накладання шаблону на фактичну інформацію.

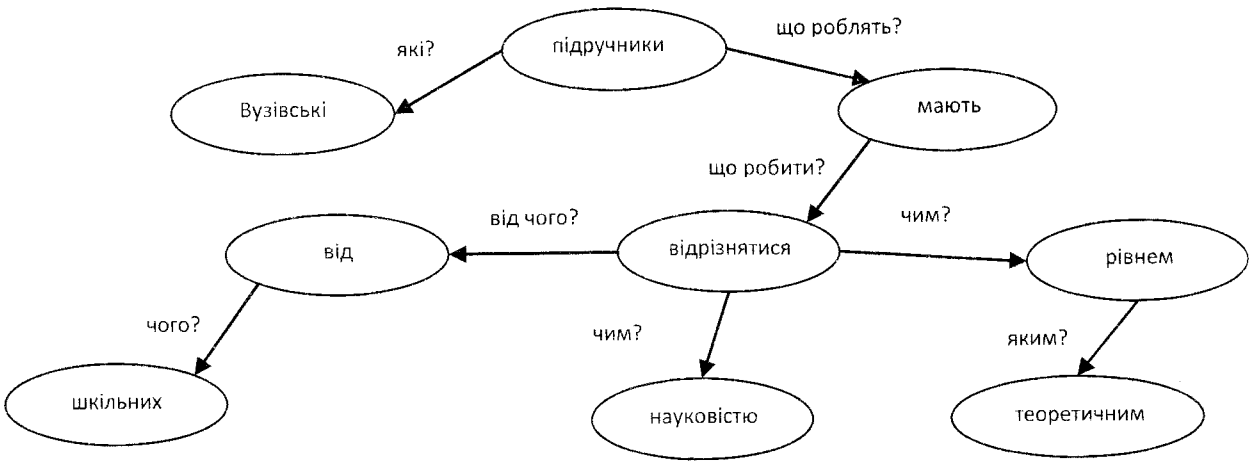

Рис. 1. Граф семантичної мережі для речення «Вузівські підручники мають відрізнятися від шкільних теоретичним рівнем та науковістю»

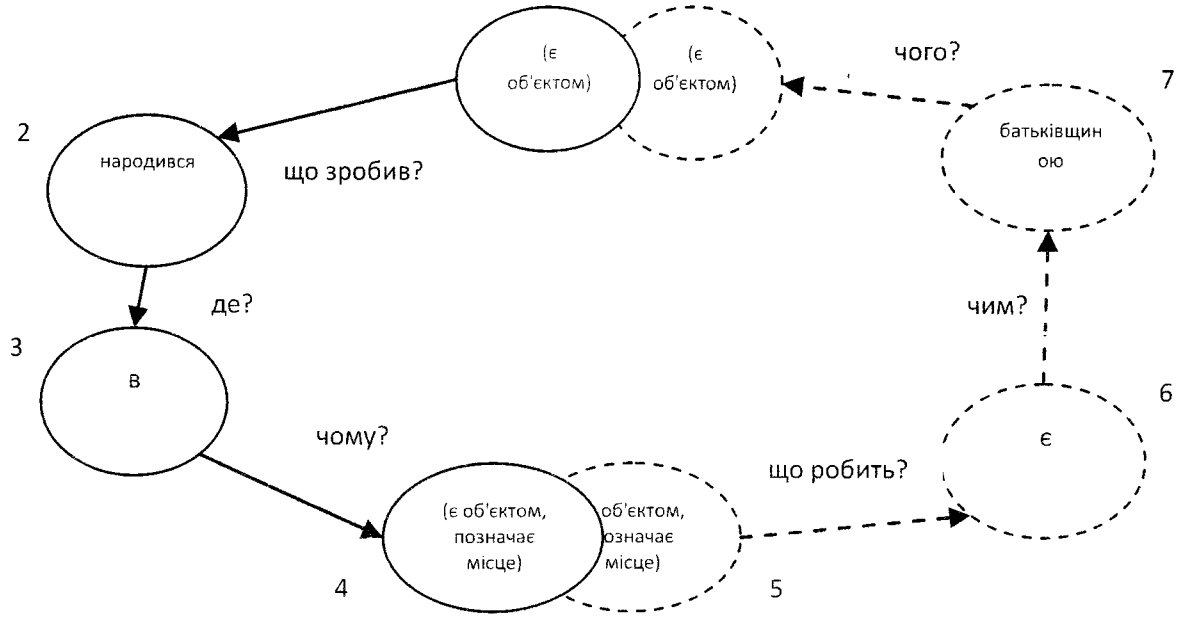

Рис. 2. Шаблон на дієслівному семантичному вузлі «народитися»

Вершини з чіткими смислами слів будемо називати *змістовними вершинами*, вершини з ознаками — *змінними*.

Саме наявність двох типів вершин дає можливість використання шаблону. За рахунок змінних вершин відбувається абстрагування від конкретних сутностей реального світу, описаних у текстовій інформації, і перехід до їх основних властивостей, а змістовні вершини дають можливість підбирати конкретні ситуації, де шаблон доцільний для підстановки.

Шаблоном можна подати дуже значну кількість інформації про об'єкти та взаємозв'язки між ними. Оскільки шаблон по суті є графом семантичної мережі з певними словоформами у вершинах і характером зв'язків на дугах, то представити ними можна і досить складні констатації фактів.

У найпростішому випадку у вершинах може міститися мінімальна інформація про саму словоформу, а точніше про сутність, що стоїть за нею — її текстове представлення. Проте у вершині можна розмістити і додаткову інформацію. Ця інформація може носити певний морфологічний, синтаксичний, або семантичний характер.

Так, з попереднього прикладу видно, що вершина 1 повинна при накладанні шаблону на інформацію обов'язково отримати зміст, що буде відображати певну сутність-об'єкт, вершина 4 також має містити сутність-об'єкт, який до того ж має означати певне місце.

Накладання шаблону на інформацію

Пропонований механізм синтезу фактів являє собою процес послідовних операцій: синтаксичний аналіз вхідної текстової інформації з метою встановлення точних смислів слів та характеру взаємозв'язків між ними, представлення інформації у вигляді семантичної мережі, пошук дієслівних семантичних вузлів та відбір зі списку шаблонів, а також накладання шаблону на вхідну інформацію і отримання нових констатацій фактів на основі шаблону інформації.

Графічно послідовність етапів обробки показано на рис. 3.

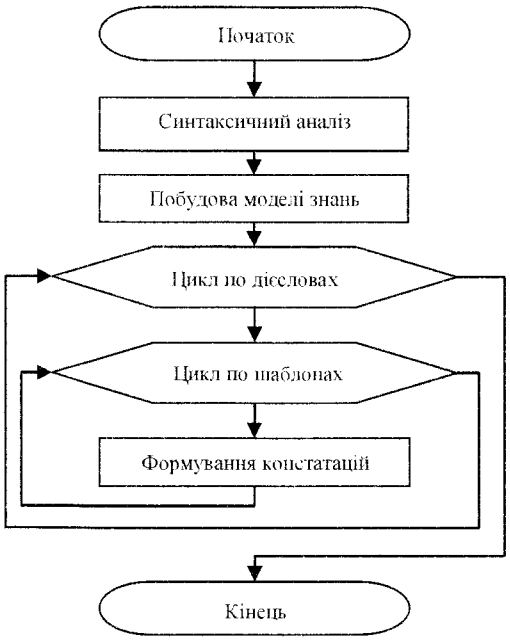

Рис. 3. Послідовність операцій отримання нових констатацій фактів

Синтаксичний аналіз

Відбувається поділ текстової інформації на речення, речень — на слова.

Кожному слову приписується набір граматичних властивостей, і подальше встановлення зв'язків між словами з множини *Questions*.

Побудова моделі знань

Перетворення графу з результатами синтаксичного аналізу в граф семантичної мережі, що є більш зручною для використання. Відбувається уточнення смислів слів, характеру зв'язків між словами, усунення невизначеностей.

Цикл по дієсловах

Для кожного дієслівного семантичного вузла інформації відбувається підбір відповідних шаблонів.

Цикл по шаблонах

Для всіх відповідних шаблонів проводиться накладання на інформацію з метою отримання нових констатацій.

Формування констатацій

На основі шаблону виводяться нові констатації фактів.

Після отримання текстової природно-мовної інформації у вигляді описаної вище семантичної мережі необхідно виділити в ній вершини, що містять у собі дієслова, тобто позначають дії.

Після цього для кожного дієслова з усієї сукупності шаблонів обираються ті, основа яких рівна за смислом до поточного дієслова.

По кожному шаблону потрібно провести суміщення його вершин з вершинами інформації.

Для змістовних вершин це суміщення проводиться на основі рівності смислів слів, для змінних вершин — на основі рівності визначених для вершини шаблону ознак.

Якщо всі вершини певної гілки шаблону знайдені, то інші його гілки можна використати для отримання нових констатацій фактів.

Схематично цей процес показано на рис. 4.

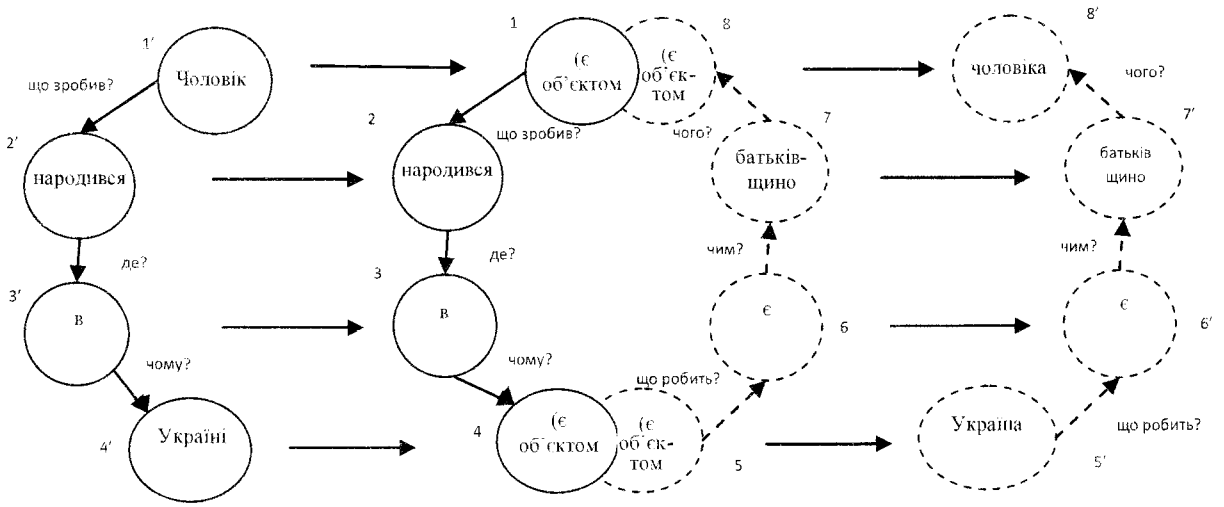

Рис. 4. Схема трансформації факту на дієслівному семантичному вузлі «народитися»

Висновки

Значною проблемою обробки текстової природно-мовної інформації є багатоманітність можливих констатацій одного факту різними словарними конструкціями з різними смислами окремими слів, та єдиним смислом висловлювання.

Вирішити цю проблему можна створивши набори шаблонів, що будуть пов'язувати сутності реального світу в єдині образи або конструкції.

Як моделі представлення знань використовується семантична мережа з фіксованим набором зв'язків між вершинами.

Суть механізму зводиться до накладання графу семантичної мережі шаблону на граф семантичної мережі фактичної інформації, з подальшим використанням шаблону для отримання нових констатацій фактів.

У даній статті запропоновано будувати шаблони на основі дієслівних семантичних вузлів, що позначають певні дії.

ЛІТЕРАТУРА

1. *Частиков А. П.* Проектирование экспертных систем / А. П. Частиков, Т. А. Гаврилова, Д. Л. Белов. — СПб. : БХВ-Петербург, 2003. — 393 с.

2. *Сич М. Ю.* Алгоритм семантичної обробки текстової інформації. // Проблеми інформатизації та управління: зб. наук. праць. — К. : НАУ, 2009. — Вип. 1(25). — С. 159–164.

Стаття надійшла до редакції 22.03.2012