

РОЗРОБКА ЕКСПЕРТНОЇ СИСТЕМИ НА БАЗІ МЕТОДУ АВТОМАТИЗОВАНОГО ФОРМУВАННЯ ЛОГІКО- ЛІНГВІСТИЧНОЇ МОДЕЛІ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Інститут комп'ютерних технологій
Національного авіаційного університету

Розглянуто процес розуміння змісту текстової інформації, проаналізовано способи представлення її у формалізованій формі, запропоновано використання експертної системи як програмного засобу реалізації методу автоматизованого формування логіко-лінгвістичної моделі, проведено аналогію між етапами розробки експертної системи та етапами виконання методу, запропоновано концептуальну схему та чіткі правила формування логіко-лінгвістичних моделей тексту

Вступ

Мислення не можливе поза мовою. Всі явища зовнішнього світу знаходять відображення у вигляді категорій представлення. Здійснюючи роботу з мовними структурами, можна змінювати поточний стан відображеного зовнішнього світу. Успіх у формалізації розуміння змісту природної мови досягнутий на шляху використання аналогії між математичними та природними мовами.

Внутрішнє представлення фрагменту реальності представляє собою ситуацію, яка виступає в ролі імітаційної моделі фрагменту реальності і є суб'єктивною. Множина можливих фрагментів реальності відповідає множині внутрішніх представлень – ситуацій. Тому є можливим говорити про існування деякої функції, яка перетворює фрагмент реальності в ситуацію, тобто процес сприйняття. Вхідні аргументи цієї функції мають область визначення у вигляді множини фрагментів реальності і називаються ситуаційними змінними. В якості цих значень змінних виступає інформація, що вилучається з фрагменту дійсності. Вихідні значення функції сприймання мають область визначення у вигляді множини різноманітних ситуацій [1].

Процес розуміння змісту тексту на природній мові також можна розглядати як обчислення деякої функції по перетворенню фрагменту реальності - тексту – у

внутрішню модель – ситуацію, а надалі в зміст тексту.

В алгебраїчній логіці висувається гіпотеза про те, що речення природної мови представляють собою формули деякої алгебри. Математичні тексти теж пишуться у вигляді окремих речень і називаються висловлюваннями. Якщо така гіпотеза підтвердиться, то стане можливим обробляти природну мову так само, як і математичні вирази. В математиці абсолютно точно встановлено, що зміст висловлювання – це предикат, що виражається цим висловлюванням, аргументами такого предикату є змінні, які присутні у даному висловлюванні [2].

Отже, текст на природній мові можна ототожнити з математичною формулою, записаною за певними правилами, а змістом речення природної мови є деякий предикат. Тоді в математиці, до якої належить формула, повинні бути присутні такі елементи, як змінні, операції над змінними, операції над операціями. Послідовність операцій над змінними та операціями буде записана у вигляді формули.

Засобом формалізації текстової інформації, тобто математичною формулою за допомогою якої можна представити речення природної мови і вилучити з неї зміст, являється логіко-лінгвістична модель. Правила побудови, а також алгоритм побудови таких моделей для речень природної мови висвітлені в методі автоматизованого формування логіко-

лінгвістичних моделей [3]. Засобом реалізації цього методу є експертна система.

Постановка задачі

Експертні системи представляють собою комп'ютерні програми, здатні накопичувати знання, що містяться в різних джерелах, і моделювати процес експертизи, тобто рішення спеціалістами тієї чи іншої сфери неформалізованих задач на основі свого професійного досвіду. На відміну від програм, що використовують процедурний аналіз, експертні системи вирішують задачі в певній предметній області на основі дедуктивних досліджень [4]. Такі системи здатні знайти рішення для неструктурованих і погано визначених задач. Відсутність структурованості долається за допомогою евристик (правил, що вказані раніше і сприймаються на віру), це може бути корисним в тих ситуаціях, коли брак необхідних знань або

часу виключає можливість проведення повного аналізу [5].

Прикладом досі не вирішеної і не структурованої задачі є автоматизація процесу побудови логіко-лінгвістичної моделі для будь-якої текстової інформації. Вирішенням цієї проблеми є створення експертної системи на базі методу автоматизованого формування логіко-лінгвістичної моделі текстової інформації.

Експертна система працює в двох режимах: набуття знань та в режимі рішення задач. В режимі набуття знань експерт наповнює систему знаннями (правилами), які дозволяють їй в режимі розв'язку самостійно вирішувати задачі з області експертизи. Режиму набуття знань в традиційному підході до розробки програм відповідають етапи алгоритмізації, програмування та відлагодження програми (рис. 1) [6].

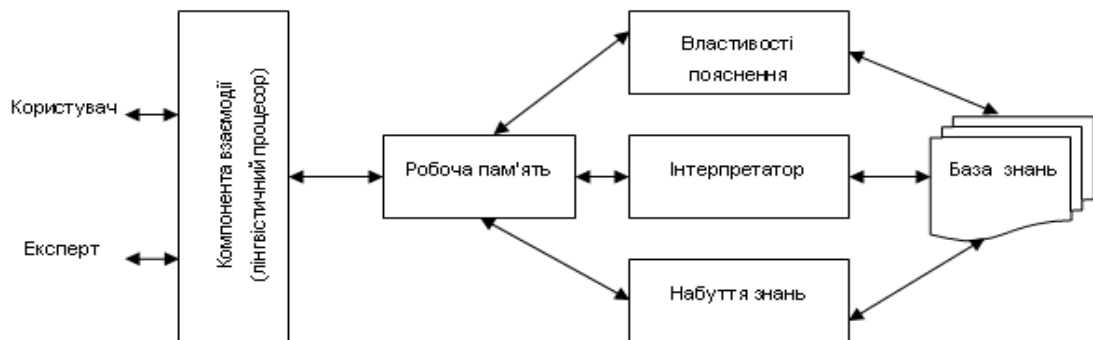


Рис. 1. Узагальнена схема експертної системи

В режимі вирішення задач після обробки їх лінгвістичним процесором запити надходять до робочої пам'яті системи. Лінгвістичний процесор перетворює вхідні дані на природній мові в дані, представлені на внутрішній мові системи; перетворює повідомлення системи з внутрішньої мови на природну. Інтерпретатор на основі вхідних даних, продукційних правил та загальних фактів про проблемну область формує рішення задачі.

Метод автоматизованого формування логіко-лінгвістичних моделей полягає

в автоматизації процесу перетворення синтаксичних конструкцій (речень) природної мови (ПМ) в формули логіки предикатів першого порядку. Метод дозволяє вилучити з вхідної текстової інформації знання, формалізовані у відповідності з певними структурними правилами, і які комп'ютер зможе автономно використовувати при вирішенні задач за наперед вибраними алгоритмами, такими, як логічний вивід [3].

Механізмом для здійснення процесу перетворення речення в формули логіки предикатів служить автоматизований

синтаксичний аналізатор, на основі результатів якого формується логіко-лінгвістична модель.

Для кожного етапу методу характерні конкретні етапи розробки експертної системи (рис. 2) [6]:



Рис. 2. Етапи розробки експертної системи

Подача текстової інформації на вхід системи

Етап подачі текстової інформації на вхід системи в розробці експертної системи відповідає етапу ідентифікації вхідного тексту.

Для вилучення знань з текстової інформації необхідно застосувати до вхідного тексту метод, який дозволив би розглянути кожне введене слово речення систему взаємопов'язаних елементів з метою дослідження зв'язків між ними. Таким методом є аналіз (від грецького - «розкладання») – в логіці – понятійне, тобто таке, що робить можливим утворення понять, розклад єдиного на множини, цілого – на частини, складного – на його компоненти, події – на окремі ступені, зміст – його елементи, поняття – на його ознаки.

Розбиття вхідного тексту на словоформи відбувається шляхом застосування логічного прийому розкладання за ознаками на складові частини, щоб дослідити текстову інформацію (речення) та зв'язки між складовими словоформами, а також вилучити знання, які на виході системи будуть перетворені в логіко-лінгвістичну модель.

Нехай речення представляє собою систему, що складається з елементів (сло-

воформ): простих (слів) та складних (словосполучень, які автоматично включають в себе слова). Тоді об'єктом управління є складна організаційна система, що складається з n простих взаємодіючих елементів $S_i, i = \overline{1, n}$ та з m складних взаємодіючих елементів $S_j^i, i = \overline{1, n}; j = \overline{1, m}$. Таким чином ідентифікація вхідного тексту є тотожністю «розбиття вхідного тексту на словоформи» для того, щоб ідентифікувати слова речення з простими елементами формальної системи.

Етап концептуалізації

Етап концептуалізації є тотожним етапу визначення характеристик елементів формальної системи в методі автоматизованого формування логіко-лінгвістичних моделей. На даному етапі експлікуються ключові поняття, відношення і зв'язки між елементами, про які йшлося на етапі ідентифікації, а також визначаються характеристики цих елементів, необхідні для опису подальшого процесу розв'язку поставленої задачі. Концептуалізація даних представляє собою процедуру введення онтологічних представлень в накопичений масив емпіричних даних; первинну теоретичну форму, що забезпечує теоретичну організацію матеріалу; схему зв'язку понять, що до-

зволяє продукувати гіпотези про природу та характер взаємозв'язків. На цьому етапі визначається спосіб організації подальшої роботи, що дозволить рухатися від первинних теоретичних концептів до більш абстрактних конструктів.

Грамматика української мови вивчає закономірності розвитку та функціонування граматичних одиниць, формує систему правил про те, як необхідно ними користуватися і як з морфем формувати слова та змінювати їх, шукати необхідні форми та об'єднувати їх в словосполучення, як знаходити для слів та словосполучень єдино правильний порядок та формувати з них перечення для висловлення думки. Мовні одиниці характеризуються в граматиці тим, що мають граматичне значення. Граматичне значення слів виражається в їх граматичних формах, які будуть називатися граматичними харак-

теристиками формальної системи $z_i = (z_{k1}(s_i); k1 = \overline{1, m1})$, де $m1$ – кількість граматичних характеристик i -го елемента системи, $i = \overline{1, n}$. Величину $m1$ визначено як сталу для всіх елементів формальної системи, так як кожне слово в українській мові має ряд морфологічних ознак, характерних для інших слів, наприклад, частина мови, відмінок, рід, число і т.д.

Тобто визначення характеристик елементів системи дає первинне концептуальне пояснення, що носить інтерпретаційний (розпізнавальний) характер.

Процес визначення характеристик кожного елемента формальної системи представляє собою концептуальну схему, що задає множину можливих способів роботи на теоретичному рівні, припущень про природу і властивості елементів, які досліджуються (рис. 3):

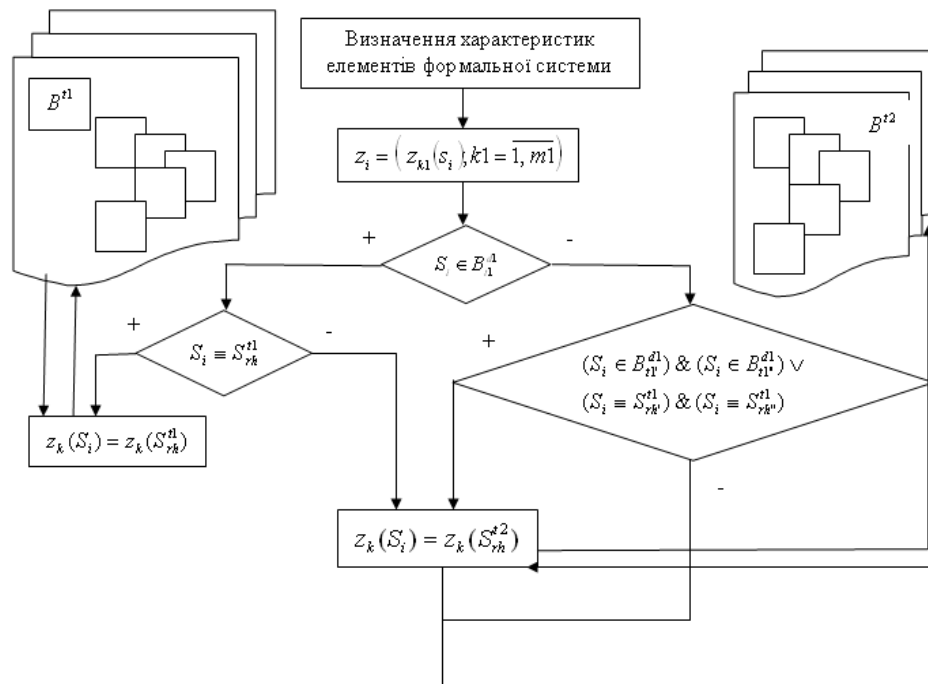


Рис. 3. Концептуальна схема для визначення характеристик кожного елемента системи

де B^{d1} – множина слів бази даних; B_{t1}^{d1} – множина слів таблиці бази даних, $B^{d1} \in B_{t1}^{d1}$; B^{t1} – множина таблиць бази даних; S_{rh}^{t1} – слово t_1 -ої таблиці із множини таблиць бази даних B^{t1} ; B^{d2} – множина слів бази даних; B_{t2}^{d2} – множина

слів таблиці бази даних, $B^{d2} \in B_{t2}^{d2}$; B^{t2} – множина таблиць бази даних; S_{rh}^{t2} – слово t_2 -ої таблиці із множини таблиць бази даних B^{t2} .

Формалізація

Етап формалізації відповідає процесу побудови лінгвістичного інтерпретато-

ра, який в свою чергу включає в себе два етапи методу автоматизованого формування логіко-лінгвістичних моделей: інтерпретацію і визначення ролей, синтаксичний аналізатор та формалізацію процесу побудови логіко-лінгвістичної моделі.

Вся задача розуміння висловлювання поділяється на два етапи: аналіз та інтерпретацію. На етапі концептуалізації було проаналізовано кожний елемент системи і визначено його характеристики. На основі отриманих даних на етапі інтерпретації передається запит до бази знань, яка представлена у вигляді аксіом формальної системи, внаслідок чого визначається структура складних елементів системи $S_j^i, i = 1, n; j = 1, m$.

Набір аксіом представляє собою сукупність правил української мови, згідно яких слова об'єднуються в словосполучення. Правила утворення словосполучень взяті з української (природної мови) та інтерпретовані у формули логіки предикатів першого порядку, так $S_i^{(cm_{i1}, s_{i2}, n_{i3}, k_{i4}, r_{i5}, l_{i6}, h_{i7}, l_{i8})}$ – простий елемент системи.

На основі даних, отриманих на попередніх етапах виконання методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації, а саме граматичних характеристик простих та складних елементів системи, аналізатор визначає синтаксичну роль кожного з елементів. В українській мові існує п'ять членів речення: підмет, присудок, додаток, означення, обставина. Саме такі значення можуть приймати елементи масиву ролей G , для якого елементи формальної системи $S_i, i = 1, n$ являються прототипами.

Кожне речення природної мови характеризується граматичною організацією: будовою речення та граматичним вираженням (оформленням) членів речення. В реченні виділяють граматичну основу (центр, ядро), без якого не можна обійтися жодне речення. Граматичну основу складають головні члени речення (підмет та

присудок). Вважається, що відношення між підметом та присудком, а також їх загальне відношення до того, що вони виражають в дійсності, і формують найважливішу граматичну властивість речення (його суть) – предикативність.

Формально роль слова в речення визначається значенням елемента з множини $G = \bigwedge_{i \in n} G_i$.

Логіко-лінгвістична модель в загальному вигляді для будь-якої складної синтаксичної конструкції (речення) представляє собою набір аксіом (1)-(2) [3]:

$$(A \& B) \vee (A \rightarrow B) \vee (A \vee B) \vee (A \sim B) \vee A, (1)$$

де A – складне логічне висловлювання, яке описує просте речення, що входить до складу простого і являється головним або описує будь-яке односкладне речення; B – складне логічне висловлювання, яке описує просте речення, що входить до складу просто і являється залежним.

$$\forall_{S_j^i} S_i = M_{S_j^i}(S_k), k/i \in K_j, j = \overline{1, m}, (2)$$

де K_j – множина елементів (слів), що входять до складного елемента S_j^i , тобто до кожного j -го словосполучення; $M_{S_j^i}$ – множина значень, які може приймати властивість залежного слова в словосполученні (допоміжних частин складного елемента S_j^i), наприклад, головне слово може позначати об'єкт, кількість, якість, що залежить від типу словосполучення; k – частина складного елемента S_j^i .

Кількість складових логіко-лінгвістичної моделі дорівнює $(m+1)$, тобто кожна аксіома (2) для $j = \overline{2, m+1}$ відображає зміст словосполучення.

Даний набір аксіом дає змогу побудувати логіко-лінгвістичну модель будь-якого речення за чітко окресленими правилами.

Складні висловлювання A і B представляють собою структуровану за певними правилами послідовність простих елементів формальної системи. В основі структуризації лежить онтологія, що

на базі характеристик кожного з елементів системи, а також значеннях елементів масиву $G = \bigwedge_{i \in n} G_i$ описує взаємозв'язки між членами речення:

$$P(x, U(x), Y(\{D(x)\}) \{Y(\{D(x)\}), \{Q(D(x))\}, \{O(x)\})$$

де P – предикат, що позначає присудок; x – підмет; $\{U(x)\}$ – множина означень, що характеризують підмет; $\{D(x)\}$ – множина додатків; $\{Q(D(x))\}$ – множина означень, що характеризують додаток речення; $\{Y(\{D(x)\}), \{Q(D(x))\}\}$ – множина лексично пов'язаних додатків і означень; $\{O(x)\}$ – множина обставин.

Виконання та тестування

Етап виконання та тестування тотожній процесу безпосередньої реалізації та перевірки методу автоматизованого перетворення речення в логіко-лінгвістичну модель.

Задача етапу виконання полягає в створенні одного або декількох прототипів експертної системи, які розв'язуватимуть необхідні задачі. Потім за результатами тестування та дослідної експлуатації на даному етапі створюється кінцевий продукт. Розробка прототипу полягає в програмуванні його компонент (або виборі необхідних інструментальних засобів) та наповнення бази знань [6].

Побудова такої експертної системи базується на підході використання декількох парадигм: продукційна парадигма, в рамках якої відбувається взаємодія правил - продукцій, спирається на функціональну, а уніфікація умов застосування правил – на логічну парадигму [7], таким чином забезпечуючи повну та ефективну обробку текстової інформації, що потрапляє на вхід експертної системи.

Висновки

Створення такої експертної системи на базі методу автоматизованого формування логіко-лінгвістичних моделей до-

зволить вирішити проблему вилучення знань з текстової інформації. Подані на вхід системи речення (будь-яка текстова інформація) перетворюються в логіко-лінгвістичну модель, яка представляє собою формулу логіки предикатів, для якої надалі можна застосовувати різноманітні алгоритми з метою відшукування в текстах логічних протиріч та їх порівняння. Логіко-лінгвістична модель згідно використаного методу представляє собою набір аксіом, кожна з яких характеризує зв'язки між словоформами речення, що дозволяє більш точно визначити зміст речення, вилучити необхідні знання.

Список літератури

1. Шабанов-Кушнарєнко Ю.П. Теорія інтелекта. Технические средства. – Х.: Вища шк. Изд-во при Харьк. Ун-те, 1986. – 134 с.
2. Шабанов-Кушнарєнко Ю.П. Теорія інтелекта. Математические средства. – Х.: Вища шк. Изд-во при Харьк. Ун-те, 1984. – 144 с.
3. Вавіленкова А.І. Автоматизація процесу побудови простого речення // Проблеми інформатизації та управління: Зб. наук. пр. Вип. 2(24) – К.: НАУ, 2008. – С.28–31.
4. Таунсенд К., Фохт Д. Проектирование и программная реализация экспертных систем на персональных ЭВМ: Пер. с англ./ Предисл. Г.С.Осипова. – М.: Финансы и статистика, 1990. – 320 с.
5. Нейлор К. Как построить свою экспертную систему: Пер. с англ. – М.: Энергоатомиздат, 1991. – 286 с.
6. Попов Э.В. Экспертные системы: решение неформализованных задач в диалоге с ЭВМ. – М.: Наука, 1987. – 288 с.
7. Искусственный интеллект: В 3-х кн. Кн.1 Системы общения и экспертные системы: Справочник / Под ред. Э.В. Попова. – М.: Радио и связь, 1990. – 464 с.