

УДК 004.82 (045)

ОБРОБКА ТЕКСТОВОЇ ІНФОРМАЦІЇ ЧЕРЕЗ ПРИЗМУ АНАЛІЗУ ТА ІНТЕРПРЕТАЦІЇ ЕЛЕМЕНТІВ ФОРМАЛЬНОЇ СИСТЕМИ

А.І. Вавіленкова

Національний авіаційний університет

e-mail: a_vavilenkova@mail.ru

При обробці знань найбільш фундаментальною і важливою проблемою є перш за все опис змістовного вмісту проблем широкого діапазону, а також наявність такої форми опису знань, яка гарантує, що обробка їх змісту формальними правилами перетворення буде здійснюватися правильно. Однак, системи інженерії знань, які являються результатом досліджень в області штучного інтелекту і які знайшли на сьогоднішній день практичне втілення, представляють собою звичайні програмні системи, в яких функції дедуктивних виводів об'єднані з відносно простими функціями діалогового вводу-виводу, але саме представлення знань в цих системах є надзвичайно простим. В існуючих системах інженерії знань можливе необмежене нарощування об'єму знань, але на практиці при цьому суттєво зростають проблеми управління знаннями. Для того, щоб створення складної системи стало дійсно можливим, необхідно розробити такі методи представлення інформації та управління її обробкою, щоб система функціонувала незалежно (в логічному розумінні) від об'єму циркулюючої в ній інформації [1].

Першими формальними засобами в 50-ті роки для автоматизації «розуміння» текстів на природній мові були КС-граматики Хомського. Однак, породжуючи правильні синтаксичні конструкції, за потужністю вони надлишкові для аналізу синтаксису і в той же час недостатні для встановлення змісту фрази. У зв'язку з цим з'явилися граматика для аналізу текстів, які є модифікацією або розширенням КС-граматик Хомського, що дозволяє враховувати семантичні ознаки. Проте універсальність семантичних ознак, їх мовна орієнтація і недостатній рівень розробки ускладнюють застосування даного підходу в аналізаторах експертних систем.

Спеціалісти в області семантики виокремлюють два рівня представлення висловлювань: а) семантичний (глибинно-синтаксичний); б) поверхнево-синтаксичний. Поверхневі структури відображають будову речення в формі узгодження граматичних одиниць (дієслово, іменник, прикметник). Глибинні ж структури, що виражають семантичне відношення, відображають логічну, а не граматичну структуру речень на природній мові. Тобто глибинні структури – це опис тестів за допомогою предикатно-аргументних структур.

Мову для запису глибинної структури і деякі правила перекладу глибинної структури в поверхневу запропонував Ч.Філмор. Він встановлює одиниці глибинних структур або семантичні ролі аргументів:

- 1) агент – живий ініціатор подій, що описується відповідним дієсловом (*Джон відкрив двері*);
- 2) контрагент – сила, що проти якої направлена дія (*чинити опір комусь*);
- 3) об'єкт – істота або не істота, яка пересувається або змінюється, положення або існування якої є предметом уваги (*розбити вікно*);
- 4) місце – фізичне тіло, на яке спричиняють дію зі сторони (*зачепити когось по носі*);
- 5) адресат – особа, над яким виконується певна дія (*вчити когось*);
- 6) пацієнт – істота або не істота, на якій випробовують якісь дії (*обкратати когось*);
- 7) результат – річ, яка виникає в результаті дії (*вичавити сік*);
- 8) інструмент – безпосередня фізична причина події (*вдарити когось палкою*);
- 9) джерело – місце, від якого направлена дія (*він віддає книгу*).

Між «ролями» (елементами глибинної структури) і аргументами – елементами тексту та елементами поверхневої структури, немає взаємно однозначної відповідності. Філмором запропоновано декілька відмінкових систем, що охоплюють різні аспекти значень певних дієслів, проте немає прямого зв'язку між ролями та граматикою, яка лежить в основі аналізу вхідного тексту, тобто в основному здійснюється незалежний аналіз поверхневих та глибинних структур тексту; опис предметної області громіздкий; «ролі» розповсюджуються тільки на поняття; аналіз проводиться на основі єдиної граматики для всіх вхідних текстів без врахування їх структурної і змістової специфікації.

Другий підхід запропоновано Р.Шенком і полягає в представленні ідентичних речень, виражених різними поверхневими структурами, єдиною концептуальною конструкцією. В концептуальній моделі використано знання про мову, що зберігаються в словнику, детальна класифікація слів, комплекс правил, що дозволяє робити висновки про текст на основі моделі знань [2].

Дослідження текстів зі спільним аналізом поверхневих та глибинних структур дозволяє простежити всі закономірності електронних текстів і вилучити з них знання, тобто виокремити зміст фрази.

На основі вивчення вище наведеної інформації, аналізу поверхневих та глибинних структур, існуючих систем управління та обробки інформації, а також дослідження структури та різноманітних закономірностей природної мови (української), запропоновано метод автоматизованого формування логіко-лінгвістичних моделей текстової інформації.

Нехай речення представляє собою систему, що складається з елементів (словоформ): 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 $S_i, i=1, n$ та з m складних взаємодіючих елементів $S_j^i, i=1, n; j=1, m$ (Рис.1). Зв'язки між елементами системи можна прослідкувати тільки після виконання наступних кроків алгоритму згідно методу автоматизованого формування логіко-лінгвістичних моделей тексту, так як для кожної конкретної системи зв'язки будуть різні.

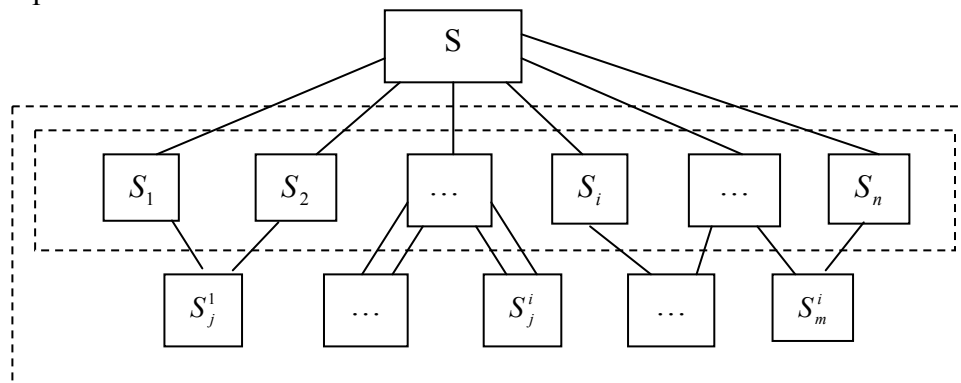


Рис.1 – Узагальнена структура формальної системи

$S_i, i=1, n$ - прості елементи системи (слова), які можуть входити або не входити до структури складних елементів;

$S_j^i, i=1, n; j=1, m$ - складні елементи системи (словосполучення), головною складовою якого являється простий елемент $S_i, i=1, n$.

Тобто, на етапі ідентифікації вхідна текстова інформація, яка представляє собою сукупність синтаксичних одиниць природної мови, ототожнюється з формальною

системою $S = \bigwedge_{i \in n} S_i = \bigwedge_{i \in n} \bigwedge_{j \in m} S_j^i$. Таке перетворення дає змогу розглядати будь-яке речення (як загодно складної структури) як систему операцій над послідовностями символів, а самі операції - формально, без змістовної інтерпретації.

Кожен простий елемент системи описується вектором значень характеристик $z_i = (z_{k1}(s_i); k1 = \overline{1, m1})$, 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. Під час морфологічного розбору кожного слова за відсутності певних морфологічних ознак у того чи іншого елемента, такі характеристики будуть замінені нулями, що дасть більше теоретичної інформації для подальшого встановлення взаємозв'язків між елементами в складних синтаксичних конструкціях. Тобто визначення характеристик елементів системи дає первинне концептуальне пояснення, що носить інтерпретаційний (розпізнавальний) характер.

Як було сказано вище, $S_i^{(cm_{i1}, g_{i2}, n_{i3}, k2_{i4}, t_{i5}, j_{i6}, j_{i7}, j_{i8})}$ - простий елемент системи, де $cm_{i1}, i1 = \overline{1, 11}$ - граматична характеристика, що позначає частину мови, може приймати одне із значень:
 cm_1 - іменник – відповідає таблиця бази даних імен, це самостійна частина мови, що означає предмет;
 cm_2 - прикметник – відповідає таблиця бази даних прик, це самостійна частина мови, що виражає ознаку предмета;
 cm_3 - числівник – відповідають таблиці бази даних chis, chis1, chis2, це самостійна частина мови, що означає число, певну кількість предметів, їх порядок при лічбі;
 cm_4 - займенник – відповідає таблиця бази даних zajm, zajmpr, це самостійна частина мови, що вказує на предмети, ознаки або кількість не називаючи їх;
 cm_5 - дієслово – відповідає таблиця бази даних die2, diedok, dienedok, це самостійна частина мови, що означає дію або стан предмета;
 cm_6 - дієприкметник – відповідає таблиця бази даних diepr, це самостійна частина мови, яка являється формою дієслова і виражає ознаку за дією;
 cm_7 - дієприслівник – відповідає таблиця бази даних die2, diedok, dienedok, це самостійна частина мови, що являється незмінюваною дієслівною формою, що виражає додаткову дію, пояснюючи в реченні основне дієслово - присудок;
 cm_8 - прислівник – відповідає таблиця бази даних prsl, це самостійна частина мови, що виражає ознаку дії або стану та ознаку якості;
 cm_9 - прийменник – відповідає таблиця бази даних prum, це службова частина мови, що виступає разом з відмінковою формою іменника, займенника або числівника для вираження залежності її від інших слів у словосполученні;
 cm_{10} - сполучник – відповідає таблиця бази даних spol, це службова частина мови, що служить для зв'язку однорідних членів речення та частин складного речення;
 cm_{11} - частка – відповідає таблиця бази даних chast, це службова частина мови, що надає слову або реченню відтінку або служить засобом творення форм слів чи похідних слів;
 $g_{i2}, i2 = \overline{1, 7}$ - морфологічна ознака, що відповідає за відмінок;
 $n_{i3}, i3 = \overline{1, 2}$ - граматичний параметр, що означає число;
 $k2_{i4}, i4 = \overline{1, 4}$ - граматичний параметр, що означає рід;
 $t_{i5}, i5 = \overline{1, 3}$ - граматичний параметр, що означає час;

$h_{i6}, i6 = \overline{1,3}$ - граматичний параметр, що означає спосіб;

$l_{i7}, i7 = \overline{1,3}$ - граматичний параметр, що означає особу.

На основі отриманих даних на рівні концептуалізації текстової інформації формується логіко-лінгвістична модель.

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

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

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

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.

K_j - множина елементів (слів), що входять до складного елементу 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 елементу S_j^i), наприклад, головне слово може позначати об'єкт, кількість, якість, що залежить від типу словосполучення;

k - частина складного елементу 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 [3]:

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

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.

Таким чином, на основі аналізу та інтерпретації кожного з елементів формальної системи можна побудувати логіко-лінгвістичну модель як заведено складної синтаксичної конструкції природної мови.

Список використаної літератури

1. Осуга С. Обработка знаний/ Пер. с япон. – М.: Мир, 1989. – 293с.
2. Бухараев Р.Г., Сулейманов Д.Ш. Семантический анализ в вопросно-ответных системах. – Казань.: Издательство Казанского университета, 1990. – 124с.
3. Вавіленкова А.І. Проблеми інформатизації та управління: Збірник наукових праць: Випуск 2 (24) / Автоматизація процесу побудови простого речення – К.: НАУ, 2008. – С.28-31.