УДК 519.62

Моделювання фрагментів природної мови алгебро-логічними засобами (на прикладі *k*-значної реалізації морфологічного аналізу)

Г.Г. Четвериков[1], Т.М. Федорова[2], І.Д. Вечірська[3], Д.Ф. Богданов[4]

[1]ХНУРЕ, м. Харків, Україна, chetvergg@kture.kharkov.ua.
[2]ХНУРЕ, м. Харків, Україна, tatanic_fedorova@mail.ru.
[3]ХНУРЕ, м. Харків, Україна, ira_sc@list.ru.
[4]Харківський національний політехнічний університет «ХПІ», м. Харків, Україна

Здійснено перехід до системного та структурного синтезу функціонально-орієнтованих процесорів на основі структур алгебри скінченних предикатів для аналізу та синтезу мови. На прикладі задачі морфологічного аналізу запропоновано методику проектування лінгвістичних об'єктів *k*-значними засобами структур алгебри скінченних предикатів просторового типу та наведено інтерпретацію їх функціонування.

АЛГЕБРА СКІНЧЕННИХ ПРЕДИКАТІВ, ВІДНОШЕННЯ, ЛІНГВІСТИЧНИЙ ОБ'ЄКТ, МОРФОЛОГІЧНИЙ АНАЛІЗ, СТРУКТУРА ПРОСТОРОВОГО ТИПУ, ФУНКЦІОНАЛЬНО-ОРІЄНТОВАНИЙ ПРОЦЕСОР

Вступ

Останнім часом в науці багато уваги приділяється питанням розробки систем автоматичної та автоматизованої обробки мовної субстанції, її аналізу, синтезу, нормалізації, тощо. Досягнення найбільшого економічного та соціального успіху стає можливим за умови активного використання сучасних алгебро-логічних методів опису структури інформаційних процесів, зокрема, «інтелектуального опрацювання текстів на предмет екстракції з них певних знань» [1]. Наразі у світі вже існують програмні засоби, які певною мірою можуть виконувати ці операції [2]. Але такі засоби і відповідні їм лінгвістичні технології повинні мати комплексний характер. Актуальною є задача забезпечення їх достатньо «високоорганізованим інтелектом», що базується як на програмних, так і на апаратних засобах формального моделювання природної мови, наближених за своїми можливостями та принципами організації до роботи мозку людини [3].

Наявність алгебри скінченних предикатів (АСП) [3] дає можливість переходу від алгоритмічного опису інформаційних процесів до опису їх у вигляді рівнянь. Рівняння ж у свою чергу задають відношення між змінними. Усі змінні у рівнянні рівноправні, будь-які з них можуть виступати як у ролі незалежних, так і в ролі залежних. Рівняння дають перевагу перед алгоритмами, яка полягає у можливості передбачення реакції системи навіть при неповній визначеності вхідних сигналів, у той час як неповністю розроблений алгоритм є непрацездатним. Крім того, зміна знань про об'єкт у вигляді системи рівнянь АСП, що описують її структуру, на противагу алгоритму, не вимагає докорінної зміни структури самої системи. За допомогою формул АСП можна побудувати АСП-структури, які реалізують відповідні скінченні предикати. Це є аналогом побудови комбінаційних схем за формулами алгебри логіки. В залежності від рівня функціонально-структурної реалізації отримуємо АСП-структури різної складності та призначення [4].

Стаття є логічним продовженням досліджень, які наведені у роботах [5-7].

1. Стан проблеми та постановка задачі

Насамперед необхідно визначитись, що таке «природна мова» або «мова людська»? Однозначної відповіді на це запитання немає. Нашим дослідженням відповідає наступне означення: «Язык – это находящаяся в распоряжении того или иного коллектива система элементов – единиц разных ярусов (морфем, слов, словосочетаний, предложения и тому подобного) плюс система правил функционирования этих единиц. Систему единиц называют инвентарем языка, а систему правил функционирования, то есть правил генерации и понимания осмысленных высказываний – грамматикой (в широком смысле) этого языка» [8], тобто мова – це система знаків, де мовний знак часто відображають у вигляді трикутника Фреге [9].

Таким чином, об'єктами мови виступають певні психофізичні стани та процеси, які відбуваються в мовнорозумовому апараті людини, а усна та писемна її репрезентації слугують елементами інфраструктури мовного процесу [1].

Відомо, що семантичному аналізу речення або зв'язного тексту передує морфологічний та синтаксичний аналіз. В прикладній лінгвістиці під морфологічним аналізом (МА) розуміють аналіз окремих словоформ поза контекстом. В автоматичній обробці природномовного тексту МА означає процедуру, в результаті виконання якої з форми, зовнішнього оформлення слова в тексті можна одержати відомості про будь-які рівні мовної структури.

Задачею морфологічного аналізу є ідентифікація словоформ та присвоєння кожній словоформі комп-

лексу морфологічної інформації (КМІ) [8]. Такий комплекс складається із морфологічно-інформаційних рядків (МІ-рядків), структура яких така: номер, <(основа чи ознаки основи), МІ>, де номер – порядковий номер даної словоформи у висловлюванні; основа (ознака основи) – код семантичної ознаки, номер синтаксичної чи семантичної моделі керування, що присвоєні даній основі у словнику основ; МІ – частина мови та її граматичні категорії: рід, число, відмінок, час, особа, тощо.

У зв'язку з цим вважається доцільним вирішити задачу побудови і дослідження АСП – структур шляхом декомпозиції багатозначних структур на дискретно-аналогові та цифрові субблоки (у вигляді універсальних функціональних перетворювачів просторового типу [10]) для схемної реалізації математичних моделей мовної субстанції на прикладі k-значної реалізації морфологічного аналізу.

2. Джерела розробки та алгоритмічна реалізація

Існує два методи реалізації [6, 7] морфологічного аналізу: словниковий (декларативний), що використовується для аналізу мов із нерозвинутим відмінюванням слів (англійська, французька, тощо) та алгоритмічний (процедурний) – для аналізу мов із розвинутим відмінюванням слів (наприклад, українська). В МА розчленовують словоформи на основу та закінчення, після чого в словниках зберігаються як основи, так і їх закінчення. Для МА виконують пошук у складі словоформи, яку аналізують, деякої словникової основи та певного словникового закінчення. Потім порівнюють інформацію про основу та закінчення і отримують КМІ для всієї словоформи.

Для МА змінюваної словоформи її кінцеву частину по черзі порівнюють із закінченнями словника. Після порівняння ту частину словоформи, що збіглася, відокремлюють і отримують припустиму основу (ПОС), припустиме закінчення (ПЗК) та припустиму морфологічну інформацію (ПМІ).

Дані про ПЗК (ПМІ) зчитують із словника закінчень (морфологічної інформації). Потім переходять до пошуку інших ПЗК, ПОС та ПМІ.

На другому кроці аналізу словоформи виконують ідентифікацію її можливих основ шляхом перевірки збігів отриманих ПОС із основами машинного словника основ.

На третьому кроці МА словоформи порівнюють інформацію з тими ПОС та ПЗК, що отримали підтвердження за допомогою словника основ.

Ефективність МА суттєво залежить від виду подання машинних словників у пам'яті комп'ютера та способу їх кодування. Тому доцільно мати окремий допоміжний словник перенумерованих основ, що наявні в одному примірнику та розташовані в алфавітному порядку.

Розглянемо приклад можливої системи кодування. Для подання значень граматичних категорій будь-якої словоформи використаємо 9-розрядний 10-значний код. Позначимо його розряди: $P(1)$, $P(2)$, ..., $P(9)$. Призначення розрядів:

$P(1)$, $P(2)$ – для кодування частин мови словоформи: 01 – іменник; 02 – повна форма прикметника; 03 – коротка форма прикметника; 04 – дієслово і так далі.

$P(3)$ – для позначення типу (прислівника, числівника, сполучника, частки), класу займенника або параметра словозміни (іменника, повного прикметника) зі значенням «живий», «неживий». $P(3) = 0$ позначає відсутність значень будь-якої з перелічених категорій або їх неозначеність. Якщо $P(3) \neq 0$, то його значення розглядається разом з $P(1)$ і $P(2)$.

$P(4)$ – для передачі значення граматичного роду словоформи: 0 – його відсутність або неозначеність; 1, 2, 3, – чоловічий, жіночий, середній рід відповідно.

$P(5)$ – для кодування значення числа: 0, 1, 2, – відсутність, однина, множина.

$P(6)$ – для позначення відмінку: 0, 1, 2, 3, 4, 5, 6, 7 – відсутність, називний, родовий, давальний, знахідний, орудний, місцевий, кличний.

$P(7)$ – для передачі значення (коду) категорії пасивності-активності: 0, 1, 2 – відсутність / неозначеність, активність, пасивність.

$P(8)$ – для вираження значення (коду) часу; 0, 1, 2, 3 – відсутність, минулий, теперішній, майбутній.

$P(9)$ – для передачі значення (коду) категорії виду: 0, 1, 2 – відсутність / неозначеність (закінчення), доконаний, недоконаний.

Кодування даних тезауруса (машиннго словника) закінчень можна представити як файл запису таким чином:

$$< XXX, P(1), P(2), ..., P(9)>,$$

де XXX – номер морфологічного типу; $P(1)$, $P(2)$, ..., $P(9)$ – код морфологічних ознак закінчень.

Усі записи, що відносяться до одного й того ж закінчення, треба розміщувати один за одним; а для спрощення процедури обробки файла наприкінці кожної групи записів треба додати один чи декілька записів, заповнених нулями. Це спрощує процедуру поповнення файла новою інформацією [11].

Для формування одного МІ-рядка для всієї словоформи порівнюють код основи та код закінчення на відповідність їх перших п'яти розрядів; якщо вони не співпадають, то дані несумісні. Для порівняння вибирають черговий код закінчення. Якщо відповідність визначена, то решта розрядів результуючого коду формуються згідно з правилами 10-значної диз'юнкції / кон'юнкції (див. табл. 1, 2) значень відповідних розрядів кодів основи та закінчення. Крім того, попередньо перевіряють умова збігів операндів чи рівність одного з них нулеві.

Приклад

Словоформа «аналізу» має ПЗК = «у», ПОС = = «аналіз»:

«аналіз» – <001 01 21 00 000>

«у» – <001 01 01 12 000>

«аналізу» – <001 01 21 12 000>

Отримана МІ відповідає таким ознакам: 001 – морфологічний тип, 01 – іменник, неживий, чоловічий рід, однина, родовий відмінок.

Таблиця 1

Таблиця істинності 10-значної функції *max* (X_1, X_2)

$f(X_1,X_2)=max\ (X_1,X_2)$		X_2									
		0	1	2	3	4	5	6	7	8	9
X_1	0	0	1	2	3	4	5	6	7	8	9
	1	1	1	2	3	4	5	6	7	8	9
	2	2	2	2	3	4	5	6	7	8	9
	3	3	3	3	3	4	5	6	7	8	9
	4	4	4	4	4	4	5	6	7	8	9
	5	5	5	5	5	5	5	6	7	8	9
	6	6	6	6	6	6	6	6	7	8	9
	7	7	7	7	7	7	7	7	7	8	9
	8	8	8	8	8	8	8	8	8	8	9
	9	9	9	9	9	9	9	9	9	9	9

Таблиця 2

Таблиця істинності 10-значної функції *min* (X_1, X_2)

$f(X_1,X_2)=min\ (X_1,X_2)$		X_2									
		0	1	2	3	4	5	6	7	8	9
X_1	0	0	0	0	0	0	0	0	0	0	0
	1	0	1	1	1	1	1	1	1	1	1
	2	0	1	2	2	2	2	2	2	2	2
	3	0	1	2	3	3	3	3	3	3	3
	4	0	1	2	3	4	4	4	4	4	4
	5	0	1	2	3	4	5	5	5	5	5
	6	0	1	2	3	4	5	6	6	6	6
	7	0	1	2	3	4	5	6	7	7	7
	8	0	1	2	3	4	5	6	7	8	8
	9	0	1	2	3	4	5	6	7	8	9

3. Апаратні платформи та побудови

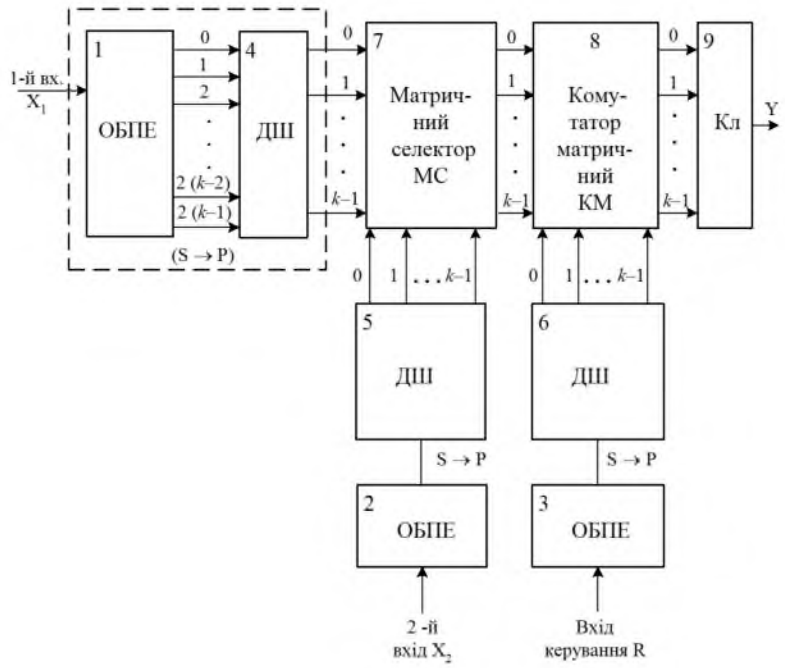
Стосовно апаратної реалізації алгоритму в розглянутому прикладі МА: найближчим підходом є застосування неоднорідних систем типу «базова машина + функціонально-орієнтований процесор» (БМ+ФОП) (рис.1) [12]. ФОП – це процесори ефективної реалізації достатньо складних функцій, які мають значну питому вагу в поставлених задачах.

ФОП поділяються на програмовані та непрограмовані. Непрограмовані ФОП називають спеціалізованими процесорами. Вони обчислюють елементарні функції, здійснюють перетворення координат, тощо. Програмовані – мають свою систему команд, яка відрізняється від системи команд БМ.

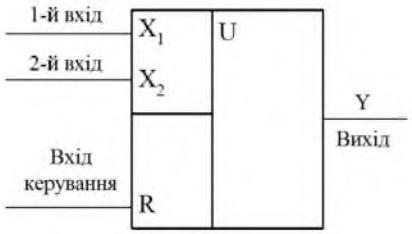
Очевидно, що в основі апаратної побудови проблемно-орієнтованої системи для нашого випадку буде лежати структура «базова машина + програмований ФОП на базі універсального двовходового десятизначного елемента» (рис. 2 *а*, *б*) [13, 14].

Оскільки алгоритм МА передбачає ряд кроків автоматичного перебору, порівняння та покрокового аналізу результатів обчислень, до апаратних засобів БМ відносяться: пам'ять словників основ та закінчень, а також оперативна пам'ять припустимих основ, припустимих закінчень та припустимої морфологічної інформації, оперативна пам'ять налагоджень універсального елемента на виконання певної логічної функції (у нашому прикладі це функція вибору максимального з двох значень). Стикування десятизначного елемента з БМ, що є на сьогоднішній день двозначною, вимагає наявності перетворювачів кодів (ПК) з двозначного коду в десятизначний та навпаки (ПК 2→10 та ПК 10→2). Процедура морфо-логічного аналізу вимагає також окремого лічильника на п'ять кроків для перевірки даних на сумісність (див.рис. 1) [15].

Рис. 1. Неоднорідна система типу «БМ+ФОП»



а



б

Рис. 2. Двовходовий універсальний багатозначний логічний елемент:
а – структурна схема;
б – умовне позначення

Таким чином, основні переваги застосування такого апаратного вирішення задачі МА полягають у наступному: підвищенні міри регулярності структури ФОП; скорочення термінів розроблення системи; можливості використання працівників нижчої кваліфікації під час проектування; зниження трудомісткості під час конструювання; гнучкість архітектури; спрощення ремонту та заміни модулів, що вийшли з ладу; зручність від використання скороченої експлуатаційної документації; зменшення числа типів кристалів та підвищення регулярності внутрішньої структури великих та надвеликих інтегральних схем; підвищення швидкодії за рахунок граничного паралелізму в роботі АСП-структури просторового типу.

Висновки

На підставі вивчення даних морфології української мови та за допомогою її опису у вигляді рівнянь АСП здійснено перехід до системного та структурного синтезу функціонально-орієнтованих процесорів на базі АСП-структур для аналізу та синтезу мови.

Запропоновано методику проектування лінгвістичних об'єктів (на прикладі задачі морфологічного аналізу) k-значними засобами на базі АСП-структур просторового типу та наведено інтерпретацію їх функціонування.

Наведені вище результати дозволяють зробити наступний важливий висновок: використання нових алгебро-логічних засобів моделювання природномовних конструкцій у вигляді системи рівнянь мовою АСП та явного способу завдання скінченного алфавітного оператора, який лежить в основі апаратурного методу розв'язання цих рівнянь, забезпечують реалізацію властивості оборотності АСП-структур і широке розпаралелення обробки символьної інформації. Фундаментальні дослідження алгебро-логічної структури природної мови, а також алгебро-логічних засобів її моделювання у вигляді відповідних АСП-структур дають можливість впритул підійти до розв'язання важливої наукової проблеми: досягнення високоякісних технологій обробки символьної інформації на основі концепції уніфікації і методів синтезу оборотних просторових багатозначних структур мовних систем.

Список літератури: **1.** *Широков В. А.* Елементи лексикографіїї. –К.: Довіра, 2005. - 304 с. **2.** *Широков В. А.* Феноменологія лексикографічних систем. –К.: Наукова думка, 2004.– 326 с. **3.** *Бондаренко М. Ф., Шабанов-Кушнаренко Ю.П.* Теория интеллекта. –Х.: СМИТ, 2006. –576 с. **4.** *Четвериков Г. Г.* Формалізація принципів побудови універсальних k-значних структур мовних систем штучного інтелекту // Доповіді НАН України.–2001.–№1 (41).–С. 76 –79. **5.** *Четвериков Г. Г.* Алгебро-логічні засоби моделювання природної мови. Част. 1 // Бионика интеллекта.–2004. –№1(61). –С. 42 –49. **6.** *Четвериков Г.Г.* Алгебро-логічні засоби моделювання природної мови. Част. 2 // Бионика интеллекта: науч.-техн. журнал.– 2005. –№2(62). –С. 25 –31. **7.** *Четвериков Г.Г.* Алгебро-логічні засоби моделювання природної мови. Част. 3 // Бионика интеллекта.–2007.–№1(66).–С. 48 –52. **8.** *Апресян Ю.Д.* Лексическая семантика. –М.: Наука, 1974.–324 с. **9.** *Попов Э.В.* Общение с ЭВМ на естественном языке.–М.: Наука, 1982.– 360 с. **10.** *Четвериков Г.Г., Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.* Модуль для преобразования кодов, заданных логическими уравнениями // А.с. 1241483 СССР, МКИ Н03М 7/00 / (СССР).–№3844735/24; Заявлено 11.12.84; Опубл. 30.06.86. Бюл. №4. **11.** *Бондаренко М.Ф., Коноплянко З.Д., Четвериков Г.Г.* Основи теорії синтезу надшвидкодіючих структур мовних систем штучного інтелекту.–К.: ІЗМН, 1997.– 264 с. **12.** Функционально ориентированные процессоры / А.И.Водяхо, В.Б.Смолов, В.У.Плюснин и др. / Под ред. В. Б. Смолова.–Л.: Машиностроение, 1988.–224 с. **13.** *Бондаренко М. Ф., Коноплянко З.Д., Четвериков Г.Г.* Основи теорії багатозначних структур і кодування в системах штучного інтелекту.–Х.: Фактор-друк, 2003.–336 с. **14.** *Бондаренко М.Ф., Коноплянко З.Д., Четвериков Г.Г.* Функциональный преобразователь с многозначным кодированием // Пат. 2147789 РФ, МКВ Н 03 К 19/02, Н 03 М 1/00. (Україна). –№97101717/09; Заявл. 04.02.97; Опубл. 24.04.2000. Бюл.№11. **15.** *Бондаренко М.Ф., Осыка А.Ф.* Автоматическая обработка информации на естественном языке: Учеб. пособ. –К.: УМК ВО, 1991. –142 с.

Надійшла до редколегії 13.02.2009

УДК 519.62

Моделирование фрагментов естественного языка алгебро-логическими средствами (на примере k-значной реализации морфологического анализа) / Г. Г. Четвериков, Т. Н. Федорова, И. Д. Вечирская, Д. Ф. Богданов // Бионика интеллекта: науч.-техн. журнал. –2009. –№ 1 (70). –С. 55–59.

В статье рассмотрены основные принципы проектирования лингвистических объектов k-значными средствами структур пространственного типа. Исследования проводились на данных морфологического анализа украинского языка средствами алгебры конечных предикатов.

Табл.: 2. Ил.: 2. Библиогр.: 15 назв.

UDK 519.62

Modeling fragments of natural language by algebra-logic means (on an example of k-unit realization of the morphological analysis) / G. G. Chetverikov, T. N. Fedorova, I. D. Vechirsky, D. F. Bogdanov // Bionics of Intelligence: Sci. Mag. –2009. –№ 1 (70). –P. 55–59.

In article main principles designing of linguistic objects are considered by k-unit means structures of spatial type. Researches were spent on the data of the morphological analysis of the Ukrainian language by means of algebra final predicates.

Tab.: 2. Fig.: 2. Ref.: 15 items.