УДК 004.93:159.95

О. В. Бісікало, І. В. Богач
ВНТУ, м.Вінниця, obisikalo@gmail.com, ilona.bogach@gmail.com

ФОРМАЛЬНЕ ВВЕДЕННЯ ОБРАЗНОГО РІВНЯ ДО ТРАДИЦІЙНОЇ ЛІНГВІСТИЧНОЇ ТРІАДИ МОРФОЛОГІЯ-СИНТАКСИС-СЕМАНТИКА

В роботі обґрунтовано підхід до формалізації образного аналізу та синтезу речення. Визначено поліноміальну оцінку обчислювальної складності внутрішніх процедур рівня образного аналізу та синтезу (РОАС) для базових природно-мовних конструкцій з ступенем не вище 3-го. Отримана формальна специфікація зовнішніх процедур РОАС забезпечує корисні функції зняття багатозначності слів, знаходження асоціацій для мовних образів та верифікації дерева синтаксичного розбору речення.

ПРИРОДНО-МОВНІ КОНСТРУКЦІЇ (ПМК), ВЕРБАЛЬНА ІНФОРМАЦІЯ, ОБРАЗНИЙ СЕНС, АНАЛІЗ І СИНТЕЗ, КОМП'ЮТЕРНА ЛІНГВІСТИКА, ВІДНОШЕННЯ, БУЛЕАН

Вступ

Об'єктом моделювання у комп'ютерній лінгвістиці найчастіше обирають окремі процеси аналізу та синтезу природно-мовних конструкцій (ПМК). Класичний підхід до аналізу речень тексту в лінгвістиці загалом та комп'ютерній лінгвістиці зокрема передбачає послідовне виконання операцій обробки вербальної інформації на 3-х рівнях [1]: морфологічному, синтаксичному та семантичному. Синтез ПМК базується на зворотному порядку тих же самих рівнів. Проте значні складності у формалізації семантичних правил змушують дослідників вдаватися до введення штучних рівнів на зразок глибинного синтаксичного або поверхневого семантичного [2], що на практиці так і не забезпечує бажаного розуміння текстового контенту [3].

Основна формальна проблема розуміння сенсу полягає у обчислювальній складності відповідних процедур обробки вербальної інформації, що відповідає NP-повному класу задач [4]. Але нейропсихологічний аналіз мовленнєвої діяльності людини демонструє паралельний розвиток процесів різної, у т.ч. невербальної природи [5] та дозволяє висунути гіпотезу щодо існування базового образного рівня розуміння [6] природної мови. Актуальним предметом дослідження є обґрунтування формальних обмежень, властивостей та можливостей рівня образного аналізу та синтезу (РОАС) для речень та інших ПМК.

Отже, мета роботи полягає у розв'язанні таких задач: визначенні формальних особливостей РОАС для базових ПМК; обґрунтуванні доцільності запропонованого підходу з точки зору обчислювальної складності алгоритмів обробки інформації; формалізації процедур взаємодії РОАС з іншими рівнями лінгвістичної тріади.

1. Образний аналіз та синтез речення

Основним змістовним елементом тексту та базовою ПМК у подальшому будемо вважати речення. Такий вибір пов'язано з тим, що в процесі комунікації речення є природно-мовною формою відображення, насамперед, окремої події [7], а вже потім – думки. Формальні задачі відокремлення речення з тексту та морфологічного і синтаксичного аналізу слів окремого речення на даний час з прийнятною якістю вирішені для більшості природних мов.

Мовленнєва практика показує, що розуміння основного сенсу речення не вимагає його повного семантичного аналізу. Наприклад, ті ж самі мімічка та жестикуляція значно покращують розуміння висловлювань співбесідника навіть в умовах невизначеності окремих почутих слів. Отже, існує образний рівень сприйняття світу людиною, для якого вербальні ознаки є лише одним з можливих типів формальних ознак. У певному колі задач комп'ютерної лінгвістики достатньо досягти оцього загального розуміння ПМК, який у [6] формально введено у вигляді поняття образного сенсу. Тоді зображений на рис. 1 РОАС забезпечить розв'язок таких задач, оминаючи труднощі семантичного рівня.

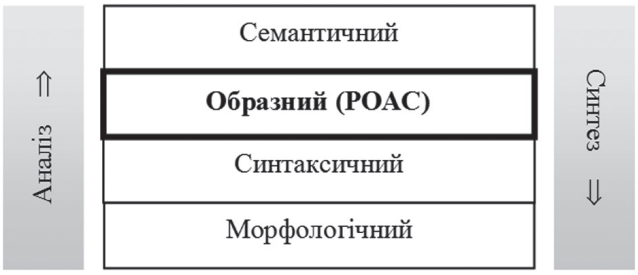

Рис. 1. Образний рівень у традиційній тріаді

Прикладами можливого класу задач образного рівня можна вважати такі: пошук за змістом ПМК, побудову природно-мовних онтологій, кластеризацію, анотування та реферування текстів, переклад, підтримку спрощених типів діалогу, коли відповіді надаються у вигляді множини слів, асоціативно пов'язаних з питанням («дельфійський оракул»), цитат з літературних джерел («магістр Йода»), з недотриманням синтаксичних правил («Basic English»).

Останні приклади задач підтримки спрощених типів діалогу є обмеженими випадками відомого тесту Тьюринга, який відносять до так званих AI-повних задач. Запропонований підхід до формалізації рівня образного аналізу та синтезу ПМК має забезпечити можливість отримання таких обмежених поняттям образного сенсу випадків класу AI-повних задач, для розв'язку яких достатньо поліноміальної складності обчислювальних процедур. Ще одним перспективним, на думку авторів, прикладом дослідження AI-повної задачі на основі окресленого підходу є формалізація жарту та інших природно-мовних форм гумору.

Концептуально рівень образного аналізу та синтезу ПМК базується на використанні тезауруса мовних образів [7]. Останніми вважатимемо множини однокореневих слів, які характеризують окремий образ з нескінченної множини $I = \{i_1, i_2, ..., i_n, ...\}$, що забезпечує морфемну класифікацію та гніздовий принцип об'єднання словоформ у тезаурусі. Потрібну для РОАС лінгвістичну інформацію щодо мовних образів доцільно зберігати у реляційній базі даних [8], що складається з відношень:

$$RE = \begin{Bmatrix} \mathrm{Im}\,age, Assoc-Twice, Construct, Event, \\ Interrogative-\mathrm{Pr}\,onoun, Link, Text, Words, Role \end{Bmatrix}. \quad (1)$$

Такий склад бази даних отримано шляхом побудови булевої алгебри сенсу як двоосновної алгебраїчної системи $BAS = \langle B; \Omega_b \rangle$, де $B = \{Word, Number\}$ – основи, а $\Omega_b = \{OP, RE, IF\}$ – сигнатура системи.

Множину-ступінь або булеан $P(I)$ множини образів I, що містить в собі n елементів, було покладено в основу кодування мовних образів [9] для бази даних у вигляді сукупності бінарних послідовностей (чисел), яка побудована на основі таких правил:

- кількість розрядів кожного числа відповідає n;
- кількість всіх чисел дорівнює 2^n;
- якщо i-й елемент I входить у цю підмножину $P(I)$, то в i-му розряді відповідного коду знаходиться 1, а інакше – 0;
- пуста множина $\varnothing$ також входить в $P(I)$ та позначається як 000...0 (кількість розрядів дорівнює n).

2. Оцінка обчислювальної складності внутрішніх процедур РОАС ПМК

Розглянемо декілька аргументів щодо доцільності введення нового РОАС ПМК на основі оцінки складності внутрішніх обчислювальних процедур, що забезпечують розв'язання розглянутих вище корисних для практики обробки текстової інформації задач.

Надамо оцінку інформаційної збитковості потенційних технологічних засобів підтримки образного сенсу ПМК. З цією метою проведемо порівняння задіяних інформаційних ресурсів з відомими підходами [10, 11].

1. Збитковість бази даних [8]. Оскільки можна вважати, що потік вхідної інформації деякої системи обробки ПМК [7] представляє собою множину текстових файлів, то потрібний для їх зберігання обсяг інформації є пропорційним загальній кількості слів n_v в усіх цих файлах. Запропонований підхід передбачає визначення n мовних образів, довжина яких пропорційна середній довжині слова, причому в загальному випадку $n_v \geq n$, а при $n_v \to \infty$ за рахунок ефекту дублювання слів $n_v >> n$. Загальний обсяг інформації у Бт для відношень бази даних (1) оцінимо з таких міркувань:

– загальна кількість записів у відношеннях *Interrogative-Pronoun*, *Link*, *Text*, *Role* дорівнює $n_{const} << n$, отже, всі вони додають до БД обмежену деякою константою E [Бт] кількість інформації;

– кількість записів у відношенні *Image* дорівнює n, але в зв'язку з необхідністю зберігання поля з ідентифікаційним кодом та додаткових полів, всього інформації потрібно $n \cdot (C' + \log_{256} n)$, де C' – деяка константа [Бт];

– аналогічним чином оцінимо кількість інформації у відношеннях *Words*, *Event*, *Construct* як $n \cdot (C'' + \log_{256} n)$, де C'' – константа [Бт], $C'' > C'$;

– найбільшу кількість інформації додає до БД відношення $Assoc-Twice$ з кількістю записів n^2, а саме $n^2 \cdot (A + 4 \cdot \log_{256} n)$, де A – константа [Бт], $A < C'$.

Отже, сумарний обсяг інформації для відношень БД складає

$$V_\Sigma = A \cdot n^2 + B \cdot n^2 \cdot \log_{256} n + C \cdot n + D \cdot n \cdot \log_{256} n + E,$$

де A, B, C, D, E – константи [Бт]. Зрозуміло, що з моменту досягнення співвідношення $n_v > n^2$ кількість інформації V_Σ БД наблизиться, а потім стане меншим від обсягу інформації, потрібної для зберігання множини вхідних текстових файлів.

2. Збитковість бази знань. На відміну від отриманої шляхом нормалізації відношень схеми БД запропоновані засоби підтримки образного сенсу електронного контенту використовують базу знань у вигляді модифікованих згідно з запропонованими у [7] методами моделювання відношень BAS. З інформаційної точки зору різниця між БД та базою знань полягає у додаванні до частини відношень останньої полів з n-розрядним бінарним кодом [9], що відповідає булеану, а саме:

відношення *Text* з кодом *Bi-Te* має розмір $n_{const} \cdot (E' + n/256) \approx n'_{const} \cdot n + E'$, де n_{const}, n'_{const}, E' – деякі константи;

відношення $\mathrm{Im}\,age$ з кодом *Bi-I* має розмір $n \cdot (C' + n/256) \approx n^2/256 + C' \cdot n$, де C' – деяка константа [Бт];

відношення *Event* з кодом *Bi-Sy* має розмір $n \cdot (C'' + \log_{256} n + n/256) \approx n^2/256 + C'' \cdot n + n \cdot \log_{256} n$, де C'' – константа [Бт];

відношення $Assoc-Twice$ з кодами $Bi\text{-}I_1$ та $Bi\text{-}I_2$ має розмір:

$$n^2 \cdot (A' + 2 \cdot \log_{256} n + n/128) \approx$$
$$\approx F \cdot n^3 + A' \cdot n^2 + B' \cdot n^2 \cdot \log_{256} n,$$

де F, A', B' – деякі константи [Бт].

З урахуванням співвідношень БД загальний обсяг бази знань складає :

$$V_{\Sigma}^{BAS} = F \cdot n^3 + A \cdot n^2 + B \cdot n^2 \cdot \log_{256} n +$$
$$+C \cdot n + D \cdot n \cdot \log_{256} n + E.$$

Отже, V_{Σ}^{BAS} збільшилося порівняно з V_{Σ} на член полінома третього степеня.

3. Збитковість інформаційного забезпечення графових моделей. Відомо, що орієнтований або неорієнтований граф розмірністю n задається матрицею суміжності розмірністю n^2. У базі знань на основі *BAS* роль матриці суміжності грає відношення *Assoc – Twice*, розмірність якого, як було показано раніше, досягає $F \cdot n^3 + A' \cdot n^2 + B' \cdot n^2 \cdot \log_{256} n$, де F, A', B' – константи [Бт]. Маємо некритичне уповільнення роботи відомих алгоритмів оброблення графів, тим більше, що врахування бінарних кодів необхідне не для всіх функцій системи обробки ПМК.

Додатково зауважимо, що оброблення n-вимірних бінарних кодів булеану за допомогою операцій та предикатів *BAS* [7] потребує значно більше пам'яті (2^n), ніж їх зберігання (n^3). Але таке обмеження рівня складності *NP*-повної проблеми для алгебраїчної системи *BAS* не варто вважати критичним, оскільки механізм побітової послідовності виконання всіх розглянутих у [7] операцій та предикатів *BAS* дозволяє застосувати для їх реалізації будь-який необхідний ступінь паралелізму відповідних обчислювальних процесів.

3. Визначення зовнішніх процедур і операцій підтримки РОАС ПМК

Забезпечення корисних функцій РОАС ПМК досягається шляхом специфікації обміну його даних, у першу чергу, з попереднім синтаксичним, а також з наступним семантичними рівнями обробки ПМК (див. рис. 1). Отже, формально потрібно визначити можливості операторів *СинА* → *ОбрА*, *ОбрА* → *СемА*, *СемС* → *ОбрС*, та *ОбрС* → *СинС*, де *СинА* і *СемА* – результати синтаксичного та семантичного аналізу, а *СинС* та *СемС* – результати синтаксичного і семантичного синтезу.

Найважливішою складовою *СинА* пропонується вважати дерево підлеглості (синтаксичного розбору) речення як базової ПМК. Тоді на РОАС ПМК з'являється можливість отримати *ОбрА* як дерево підлеглості мовних образів за допомогою тезауруса. Цим самим буде реалізована корисна функція зняття багатозначності слів та створені передумови для забезпечення функцій *СемА* – верифікації дерева синтаксичного розбору речення з метою поповнення бази знань та можливого логічного виведення.

Зворотний процес синтезу передбачає представлення певної (фактичної) інформації з бази знань у вигляді речення. З цією метою внаслідок застосування *ОбрС* мають бути знайдені асоціації для мовних образів, що відповідають формальним конструктам логічних висновків *СемС* з бази знань. Отримане таким чином дерево підлеглості мовних образів, у свою чергу, є необхідною складовою для *СинС* – того ж самого дерева підлеглості слів речення.

На думку авторів, навіть «полегшений» варіант аналізу та синтезу ПМК, що не виходить на семантичний рівень і зациклюється на РОАС, окрім визначених у п.1 задач, може бути корисним для формальної оцінки загального задуму тексту. Проте це твердження потребує подальших експериментальних досліджень з урахуванням асоціативних властивостей вербальних ознак мовних образів [12].

Важливою характеристикою запропонованих процедур обміну інформацією є суттєве зменшення пошукового простору на рівні мовних образів. Будемо вважати, що 1 мовний образ у середньому об'єднує через підлеглі лексеми та відповідні до них словоформи приблизно 100 різних за написанням слів типової флексійної, наприклад, української мови. Тоді кількість можливих зв'язків n^2 (розмірність матриці суміжності орієнтованого графу) для n мовних образів зменшиться на 10^4.

Позитивний ефект, який зі свого боку вносить у напрямку зменшення складності інформаційних процедур задекларована функція зняття багатозначності слів, пропонується оцінити таким чином. Нехай середня кількість значень однієї словоформи у тлумачному словнику дорівнює s, а середня кількість значущих слів у реченні – k. Тоді пара слів з синтаксичного дерева розбору речення має s^2 всіх можливих значень, а, відповідно, все дерево, що об'єднує $k-1$ таку пару, формально потребує розгляду $(k-1) \cdot s^2$ можливих значень. Наприклад, для наукової лексики української мови з нижніми кордонами значень $s = 3 \div 5$ та $k \approx 7$ така оцінка дає зменшення простору на два порядки тільки для одного речення.

Отже, ми бачимо, що введення РОАС ПМК з формальної точки зору сприяє зменшенню складності статистичного методу обробки текстової інформації. Зрозуміло, що людині навіть на думку не спадає перебирати 100 різних варіантів „прочитання" одного речення, а правильний варіант миттєво з'являється навіть на підсвідомому рівні. Проте ця зовнішня „легкість" розуміння сенсу вербальної інформації базується на потужній основі досвіду та знань людини. Наша мета – отримати різновид гібридного методу за рахунок застосування бази знань мовних образів, який, на відміну від відомих, імітує природний шлях розвитку і потребує мінімум експертної лінгвістичної інформації.

Для технологічної реалізації запропонованого підходу до аналізу та синтезу текстового контенту в

подальшому потрібно забезпечити такі формальні операції різного ступеню складності:

– відокремлення речення або автономної частини речення з тексту;

– перетворення речення у список слів, що ідентифікуються модулем тезауруса;

– побудова та модифікація тезауруса мовних образів обраної природної мови;

– розробка парсеру обраної мови для отримання дерева підлеглості (залежностей) речення через синтаксичні зв'язки;

– образна індексація множини текстів за рахунок накопичення синтаксичних зв'язків між мовними образами тезауруса у вигляді семантичної мережі;

– перетворення базових словоформ з дерева мовних образів у речення або ПМК за допомогою модулів синтаксичного та морфологічного рівня.

Зауважимо, що останню операцію доцільно реалізувати за допомогою вже існуючих морфологічних та синтаксичних бібліотек з відкритих лінгвістичних ресурсів. Найбільш складною задачею залишається автоматизація парсерінгу флексійних природних мов з метою отримання дерева підлеглості (залежностей) речення через синтаксичні зв'язки.

Висновки

Доцільність введення рівня образного аналізу та синтезу ПМК у традиційну тріаду морфологія–синтаксис–семантика пов'язана з можливістю отримання прийнятних рішень актуального кола задач комп'ютерної лінгвістики. При цьому зникає необхідність долати відомі труднощі семантичного рівня. З формальної точки зору підхід базується на отриманні обчислювальних процедур поліноміальної складності для обмежених поняттям образного сенсу випадків класу AI-повних задач.

Необхідність застосування експоненційних обсягів пам'яті для оброблення в межах операцій та предикатів *BAS* n-вимірних бінарних кодів булеану долається шляхом розпаралелювання обчислювальних процесів. Проте отримані формалізми забезпечують важливі для практики функції зняття багатозначності слів, знаходження асоціацій для мовних образів та верифікації дерева синтаксичного розбору речення.

У цілому дослідження спрямовано на обґрунтування одного з можливих підходів до моделювання природного шляху створення лінгвоінтелекту. З метою подальшого розвитку запропонованого підходу необхідно реалізувати шість формальних операцій, найважливішою з яких є розробка парсеру обраної природної мови для отримання дерева підлеглості (залежностей) речення через синтаксичні зв'язки.

Список літератури: 1. *Apresyan J.* ETAP-3 Linguistic Processor: a full-fledged NLP implementation of the MTT / J. Apresyan, I. Boguslavsky, L. Iomdin etc. // MTT 2003, First International Conference on Meaning – Text theory (Paris, June 16-18, 2003).– Paris, ENS, 2003. – P. 279–288. **2.** *Кобозева И.М.* Лингвистическая семантика: учебное пособие / И.М. Кобозева– М.: Эдиториал УРСС, 2000. — 352 с. – ISBN 5-8360-0165-0. **3.** *Стюарт Р.* Искусственный интеллект: современный подход / Р. Стюарт, Н. Питер – 2-е изд. – М. : Вильямс, 2006. – 1408 с. **4.** *Шлезингер М.И.* Решение (MAX,+)-задач структурного распознавания с помощью их эквивалентных преобразований. Часть 1 / М.И. Шлезингер, В.В. Гигиняк // Управляющие системы и машины. – 2007. – № 1. – С. 3-15. **5.** *Лурия А.Р.* Язык и сознание / А.Р. Лурия; под ред. Е.Д. Хомской. – М.: Издательство Московского университета, 1979. – 320 с. **6.** *Бісікало О.В.* Формалізація понять мовного образу та образного сенсу природномовних конструкцій / О.В. Бісікало // Математичні машини і системи. – 2012. – № 2. – С. 70–73. – ISSN 1028-9763. **7.** *Бісікало О.В.* Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія / О. В. Бісікало. – Вінниця : ВНТУ, 2013. – 316 с. – ISBN 978-966-641-528-1. **8.** *Бісікало О. В.* Структура блоку пам'яті на основі моделі образного мислення людини / О. В. Бісікало // Искусственный интеллект. – 2007. – № 3. – С. 461–468. – ISSN 1561-5359. **9.** *Бісікало О. В.* Дослідження простору асоціативних пар в контексті бази знань електронного підручника / О. В. Бісікало // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2006. – № 2 (28). – С. 109–113. **10.** *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Теоретико-модельная формализация онтологии и рефлексии / Д. Е. Пальчунов // Философия науки. – 2006. – № 4 (31). – С. 86–113. **11.** *Широков В. А.* Феноменологія лексикографічних систем / В. А. Широков ; НАН України, Укр. мов.-інформ. фонд. – К. : Наукова думка, 2004. – 327 с. **12.** *Горошко Е. И.* Интегративная модель свободного ассоциативного эксперимента / Е. И. Горошко. – М. : Харьков : ИЯ РАН ; Каравелла, 2001. – 318 с. – ISBN 966-7012-09-3.

Надійшла до редколегії 28.05.2013

УДК 004.93:159.95

Формальное введение образного уровня традиционной лингвистической триады морфология-синтаксис-семантика / О.В. Бисикало, И.В. Богач // Бионика интеллекта: науч.-техн. журнал. – 2013. – № 2 (81). – С. 27-30.

В статье рассматривается подход к формализации уровня образного анализа и синтеза в составе традиционной триады морфологии, синтаксиса и семантики. Предлагаются новые возможности использования образного анализа и синтеза предложений из текста и других естественно-языковых конструкций. Обсуждаются характеристики и требования к формальным процедурам данного подхода.

Ил. 1. Библиогр.: 7 назв.

UDC 004.93:159.95

Formal entry-level imagery of the traditional triad of linguistic morphology, syntax, semantics / O.V. Bisikalo, I.V. Bogach // Bionica Intellecta: Sci. Mag. – 2013. – № 2 (81). – С. 27-30.

The level of figurative analysis and synthesis in the traditional triad of morphology, syntax and semantics is examined in the article. New possibilities of using figurative analysis and synthesis the sentences of text are suggested. The characteristics and requirements to the formal procedures of the obtained approach are discussed.

Fig. 1. Ref.: 7 items.