УДК 004.94:159.95

О.В. БІСІКАЛО

ФОРМАЛІЗАЦІЯ ПОНЯТЬ МОВНОГО ОБРАЗУ ТА ОБРАЗНОГО СЕНСУ ПРИРОДНО-МОВНИХ КОНСТРУКЦІЙ

Анотація. *У статті обґрунтовано підхід до формалізації понять мовного образу та образного сенсу природно-мовних конструкцій. Інтерпретовано формальні ознаки асоціативної мережі образів, отримано чисельну оцінку одиниці образного сенсу на основі поняття ентропії.*
Ключові слова: *асоціативна мережа образів, мовний образ, образний сенс, ентропія.*

Аннотация. *В статье обосновывается подход к формализации понятий языкового образа и образного смысла естественно-языковых конструкций. Интерпретированы формальные признаки ассоциативной сети образов, получена численная оценка единицы образного смысла на основе понятия энтропии.*
Ключевые слова: *ассоциативная сеть образов, языковый образ, образный смысл, энтропия.*

Abstract. *The approach to formalization of language image concept and figurative meaning of the natural and language constructions is substantiated in the article. The formal characteristics of the associative network of images are interpreted. The numerical estimate of the unit of figurative meaning on the basis of the concept of entropy is received.*
Keywords: *associative network of images, language image, figurative meaning, entropy.*

1. Вступ

Поняття інформації, як і поняття знань, не мають однозначного трактування, що підтверджується існуванням значної кількості різних визначень. Цінність застосування цих понять у сучасних інформаційних технологіях базується на формальних обмеженнях і, головне, кількісних оцінках існуючих баз даних та баз знань. Очевидно, що найбільш загальний характер має класична міра інформації К. Шеннона, в основу якої покладено поняття ентропії [1]. Проте оцінка знань у вигляді наукового тексту або бази знань в одиницях інформації виглядає, як мінімум, неінформативно або, швидше, парадоксально. Тому, в залежності від типу бази знань, використовують такі показники, як кількість аксіом, правил за типом ЯКЩО-ТО, вузлів семантичної мережі, фреймів тощо [2].

Виходитимемо з того, що шлях від загального поняття інформації до більш складного поняття знань передбачає накладення певних формальних обмежень. На основі моделі асоціативного образного мислення людини [3] було формалізовано поняття образного сенсу з визначенням відповідної одиниці *Сав* (Синтагматичної асоціації вага) [4]. Запропоноване в [5] поняття мовного образу, що узагальнює лексеми за ознакою спільного кореня, грає роль проміжної ланки між образними концептами та природно-мовними конструкціями (ПМК), а також забезпечує початкове ущільнення текстової інформації [6].

Мета дослідження полягає в отриманні чисельної оцінки образного сенсу ПМК на основі поняття ентропії. Така постановка задачі передбачає, що закладена в основу моделі [3] семантична мережа з образів та асоціативних зв'язків між ними і є тією природною системою обмежень, що породжує образний сенс з понять ентропії та інформації.

2. Асоціативна мережа мовних образів

Системою S будемо вважати формальну модель процесів асоціативного образного мислення людини, функції якої налаштовані на розв'язок прикладних задач комп'ютерної лінгвістики [7]. Нехай S здатна розпізнавати окремі образи з нескінченної множини $I = \{i_1, i_2, ..., i_n, ...\}$ аналогічно тому, як людина розпізнає гештальт. S також може сприйма-

ти асоціативні зв'язки між парами образів як елементи множини $\omega \in \Omega$, де $\Omega \subseteq I \times I$ – довільна множина упорядкованих пар. Образною конструкцією будемо вважати будь-яку підмножину $\gamma \subseteq \Omega$, яка є елементом F – σ-алгебри підмножин з Ω.

Нехай система S обмінюється інформацією з зовнішнім світом виключно у вигляді γ, з яких розрізнятимемо послідовність вхідних подій $X = \{x_1, x_2, ...\}$, де $x_i \in \mathsf{F}$. Внаслідок цього формується база знань системи як семантична мережа, що задається матрицею A_Q. У зв'язку з переважно вербальним характером вхідної інформації системи [3], що повинна обробляти електронний контент, поставимо у відповідність вузлам мережі мовні образи.

Формально підмножину $I' \subset I$, доступних для сприйняття системою S мовних образів, представимо за допомогою четвірки основних змістовних концептів:

$$I' = \langle N; O; M; Q' \rangle,$$

де N – поняття, O – об'єкт, M – метод, Q' – якість. Тоді конструкція з мовних образів $\gamma' \subset \gamma$ є деревом орієнтованого графа, що узагальнює дерево синтаксичного розбору речення та у вершинах якого знаходяться відповідні словам речення мовні образи. Цим самим з'являється можливість накопичувати інформацію з тексту в матриці A_Q через асоціативну мережу мовних образів (АММО), що будується як $B = \bigcup_{i=1}^{m'} \gamma'_i$, де m' – загальна кількість сприйнятих системою на даний час вхідних конструкцій з мовних образів.

Задамо деяку АММО на певний момент часу такими параметрами: $k_{\lg}$ – кількість виявлених системою зв'язків між l-м та g-м мовними образами, m – кількість ненульових елементів матриці A_Q. Маємо статистичну оцінку математичного сподівання кількості повторень одного зв'язку як $\lambda = k_\Sigma / m$, де $k_\Sigma = \sum_{l=1}^{n} \sum_{g=1}^{n} k_{\lg}$. Тоді образний сенс пари (l, g) нормується сигмоїдальною функцією як $\mu_Q(< i_l, i_g >) = 1/(1 + e^{-k_{\lg} + \lambda})$ [4], що дозволяє знайти оцінку його середнього значення для всієї АММО у вигляді

$$\overline{\mu_Q} = \frac{1}{m} \sum_{j=1}^{m} \mu_{Qj} = 0{,}5 \ \ [Сав]. \qquad (1)$$

3. Чисельна оцінка одиниці образного сенсу на основі поняття ентропії

Згідно з запропонованим підходом, одиниця образного сенсу розміром один *Сав* характеризує максимальну вагу (l, g)-пари АММО як $\mu_Q(< i_l, i_g >) = 1$. У той же час факт появи на вході системи S кожної j-ї пари мовних образів з імовірністю $p(x_j)$ дозволяє оцінити ентропію цієї системи. Для отримання верхньої межі ентропії будемо вважати, що ОК складається з незалежних пар образів, хоча в реальних природно-мовних конструкціях це не зовсім так. Відомо, що у цьому випадку загальна ентропія або кількість інформації [1] системи S дорівнює

$$H = -\sum_{j=1}^{m} n_j \cdot \log p(x_j), \qquad (2)$$

де значення n_j відповідає $k_{\lg}$, як кількості зв'язків між l-м та g-м образами.

Також можна визначити середню ентропію, що припадає на одну пару. З цією метою розділимо (2) на k_Σ:

$$H_1 = -\sum_{j=1}^{m} \frac{n_j}{k_\Sigma} \cdot \log p(x_j).$$

Врахуємо, що для великих значень n_j та k_Σ імовірність j-ї пари мовних образів $p(x_j) = \lim\limits_{k_\Sigma \to \infty} \frac{n_j}{k_\Sigma}$. Тоді середня ентропія однієї пари дорівнює

$$H_1 = -\sum_{j=1}^{m} p(x_j) \cdot \log p(x_j). \qquad (3)$$

Оцінка (3) вже завищена умовою незалежності образних пар, але максимального значення середня ентропія пари досягає за додатковою умовою [1]: якщо поява кожної з m можливих пар мовних образів на вході S рівноімовірна, то

$$\overline{H_1} = \log_2 m \ \ [Біт]. \qquad (4)$$

Зрозуміло, що побудова матриці A_Q на основі реального текстового матеріалу не призведе до максимального значення ентропії (4). Але, з суто формальної точки зору, кількісні оцінки $\overline{\mu_Q}$ та $\overline{H_1}$ є різними інтерпретаціями тієї ж самої чисельної характеристики АММО – середньої ваги однієї пари. Отже, з урахуванням (1), можна отримати верхню оцінку співвідношення одиниць образного сенсу та інформації як логарифмічну згортку:

$$1 \ \ [Сав] = 2\log_2 m \ \ [Біт]. \qquad (5)$$

Тепер визначимо нижню межу розглянутого оператора згортки інформації в образний сенс (2–5), прийнявши до уваги те, що кожне речення та відповідні до нього конструкції з мовних образів є підграфами–деревами загального орієнтованого графа АММО. Тому поява тільки першої (l,g)-пари образів для S може бути незалежною, а всі наступні вже мають бути пов'язаними з l-м чи g-м образами або (рекурсивно) з новоприєднаними до підграфа мовними образами. Для оцінки середньої ентропії однієї пари образів у таких умовах скористаємося виразом

$$H = -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{n} p(x_j / x_i) \cdot \log p(x_j / x_i), \qquad (6)$$

де $p(x_j / x_i)$ – умовна імовірність появи в системі S пари x_j, якщо попередньою парою була x_i [1]. Отже, поставлена задача зводиться до рекурсивного визначення $p(x_j / x_i)$ з елементів матриці A_Q. Позначимо (l,g)-пару як x_i, а у парі x_j разом з h-м має бути або l-й, або g-й образ. Тоді маємо 4 варіанти:

$$p(x_j / x_i) = \lim_{k_\Sigma \to \infty} \left(\frac{k_{hg} \mid k_{lh} \mid k_{gh} \mid k_{hl}}{k_{\Sigma hg} + k_{\Sigma lh} + k_{\Sigma gh} + k_{\Sigma hl}} \right), \qquad (7)$$

де вибір у чисельнику елементів матриці A_Q здійснюється $\mid$ – оператором АБО в нотації Бекуса-Наура, а знаменник складають часткові суми матриці A_Q: $k_{\Sigma hg}$ – сума елементів

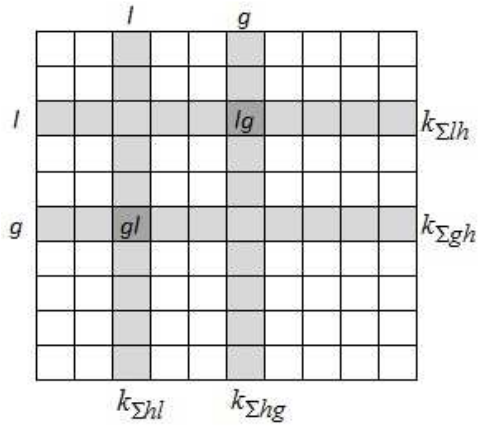

Рис. 1. Часткові суми матриці A_Q

g-го стовпчика, $k_{\Sigma lh}$ – сума елементів l-го рядка, $k_{\Sigma gh}$ – сума елементів g-го рядка, $k_{\Sigma hl}$ – сума елементів l-го стовпчика. На рис. 1 представлені часткові суми матриці A_Q для довільної першої (l,g)-пари x_i. Зрозуміло, що рекурсивне визначення часткових сум для 3-ї та подальших пар конструкції з мовних образів призведе до збільшення знаменника (7) та зменшення оцінки нижньої межі (6) оператора згортки.

4. Висновки

Запропоновану в [4] чисельну міру образного сенсу 1 *Сав* можна вважати логарифмічною згорткою ентропії образних пар конструкцій з мовних образів, що обумовлена системою обмежень у вигляді семантичної мережі АММО. Отримано оцінки верхньої та нижньої меж оператора згортки інформації в образний сенс.

СПИСОК ЛІТЕРАТУРИ

1. Кузьмин И.В. Основы теории информации и кодирования / И.В. Кузьмин, В.А. Кедрус. – К.: Вища школа, 1986. – 238 с.
2. Стюарт Р. Искусственный интеллект: современный подход / Р. Стюарт, Н. Питер; пер. с англ. – [2-е изд.]. – М.: Вильямс, 2006. – 1408 с.
3. Бісікало О.В. Концептуальні основи моделювання образного мислення людини / Бісікало О.В. – Вінниця: ПП Балюк І.Б., ВДАУ, 2009. – 163 с.
4. Бисикало О.В. Субъективная единица смысла образных конструкций / О.В. Бисикало // Nauka: teoria i praktyka – 2009: materialy V miedzynar. naukowi-praktycznej konf., (Przemysl, 7–15 sierpnia 2009). – Przemysl: Nauka i studia, 2009. – Vol. 6. – P. 9 – 12.
5. Бісікало О.В. Концептуальне поєднання понять образного мислення та мовленнєвої діяльності / О.В. Бісікало // Інформаційні технології та комп'ютерна інженерія. – 2010. – № 1 (17). – С. 72 – 77.
6. Бісікало О.В. Ущільнення інформації шляхом побудови бази знань з морфології / Олег Володимирович Бісікало, Ірина Анатоліївна Кравчук // Методи та засоби кодування, захисту та ущільнення інформації: тези доповідей ІІІ міжнар. наук.-практ. конф., (Вінниця, 20–22 квітня 2011 р.). – Вінниця, 2011. – С. 186 – 187.
7. Бісікало О.В. Онтогенетичний метод побудови нечіткого відношення сенсу / О.В. Бісікало // Штучний інтелект. – 2011. – № 1. – С. 134 – 140.

Стаття надійшла до редакції 29.08.2011