НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.9

Бісікало О. В.[1], Василевський О. М.[2]

[1]Д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики Вінницького національного технічного університету, Вінниця, Україна

[2]Канд. техн. наук, доцент, професор кафедри метрології та промислової автоматики Вінницького національного технічного університету, Вінниця, Україна

ОЦІНКА НЕВИЗНАЧЕНОСТІ ВИМІРЮВАННЯ СЕНСУ ПРИРОДНО-МОВНИХ КОНСТРУКЦІЙ

Вирішено завдання оцінки невизначеності вимірювання сенсу природно-мовних конструкцій (ПМК) на основі формалізації понять мовного образу, штучної когнітивної системи та одиниці сенсу. В основу моделі бази знань штучної когнітивної системи закладено статистичну інформацію про асоціативну сполучуваність мовних образів, що надає підстави для уніфікованої оцінки одиниці та кількості сенсу ПМК. Запропоновано метод вимірювання сенсу ПМК на основі нечіткого відношення сенсу, який забезпечує врахування інформації про зв'язки між лемами тексту, що дозволяє отримати оцінку двох типів невизначеності вимірювання ознак сенсу. Отримано та інтерпретовано формальні оцінки невизначеності результатів вимірювання сенсу ПМК, що дозволяє врахувати інформацію про зв'язки між лемами для розв'язання прикладних задач комп'ютерної лінгвістики.

За допомогою розробленого на основі пакету DKPro Core програмного забезпечення проведено експерименти з дослідження запропонованого методу в задачі виявлення інформативних ознак тексту. В результаті проведених експериментів отримано залежності параметрів виявленого Парето-подібного закону розподілу зв'язків між лемами, аналіз яких дозволяє вважати показник середньої кількості зв'язків мовного образу найбільш інформативною чисельною ознакою тексту.

Ключові слова: сенс, невизначеність, текст, природно-мовна конструкція, штучна когнітивна системи, мовний образ, лема.

НОМЕНКЛАТУРА

$[\alpha_-;\alpha_+]$ – границі апріорно визначеного закону розподілу;

γ, ПМК – природно-мовна конструкція (образна конструкція);

λ, $\overline{q}$ – математичне сподівання;

μ – цілочисельний показник емоційного стану системи;

$\mu_Q(<i_l,i_j>)$ – функція належності нечіткого відношення Q;

σ, σ_i, U_j – середньоквадратичне відхилення (СКВ);

$\Omega \subseteq I \times I$ – простір упорядкованих пар образів $I=\{i_1,i_2,...,i_L,...\}$;

$\omega \in \Omega$ – асоціативний зв'язок між парами образів (словосполучення);

A_Q – матриця суміжності розмірністю $L \times L$ на момент часу t_L;

$G_Q(V,E)$ – граф Бержа;

k – параметр розподілу Парето;

k_{lj} – значення (частота зв'язку) ненульового lj-го елементу матриці A_Q;

k_Σ – сумарне значення всіх зв'язків системи;

m – кількість ненульових елементів матриці суміжності A_Q;

N – кількість відомих ШКС;

n – кількість спостережень;

nt – кількість образів, що розрізняє ШКС;

p – довірчий рівень;

S, S_i, ШКС – штучна когнітивна система;

$u_A(X)$ – оцінка невизначеності за типом А;

$u_B(X)$ – оцінка невизначеності за типом В;

x_i – спостереження стану бінарного нечіткого відношення образного сенсу Q;

Сав – (синтагматичної асоціації вага) – одиниця вимірювання образного сенсу.

DOI 10.15588/1607-3274-2016-2-4

ВСТУП

Складність задач семантичного аналізу текстової інформації вважається однією з головних перешкод на шляху побудови штучного інтелекту в цілому та розв'язання з належною якістю значної частини задач комп'ютерної лінгвістики зокрема. В процесі онтогенезу людина вчиться та набуває нових знань все своє життя, внаслідок цього кожний природний інтелект є унікальним та динамічним явищем, здатним самовдосконалюватися та добре розуміти собі подібних. Тому конструювання лінгвістичних баз знань має базуватися саме на таких принципах, а проблема отримання нових формальних методів семантичного аналізу природно-мовних конструкцій на основі баз знань є актуальною. Потребують обґрунтування формальні підходи до створення штучних когнітивних систем, здатних імітувати діяльність людини в процесах оброблення, розуміння смислу та ефективного застосування вхідної текстової інформації.

В роботах [1, 2] було запропоновано та обґрунтовано введення одиниці вимірювання образного сенсу 1 *Сав* з метою розв'язання задач комп'ютерної лінгвістики, пов'язаних з моделюванням образного мислення людини. Але в процесі такого моделювання обов'язково потрібно врахувати суб'єктивний та динамічний характер онтогенезу пізнавальної, у тому числі мовленнєвої діяльності людини. Формально це можна зробити різними шляхами, одним з яких є оцінка невизначеності результату вимірювання сенсу як окремих природно-мовних конструкцій (ПМК), так і текстів та штучної когнітивної системи (ШКС) у цілому на певний момент часу. Відомо [3], що невизначеність вимірювання – це параметр, пов'язаний з результатом вимірювання, який характеризує дисперсію значень, що можуть бути достатньо обґрунтовано приписані вимірюваній величині. Але важливо, щоб величина, яка безпосередньо використовується для вираження невизначеності, має бути внутрішньо узгоджена: безпосередньо виведена з компонентів, які її утворюють, а також не повинна залежати від групування цих компонентів і від їх розкладу на субкомпоненти [4]. У відомих літературних джерелах, де розглянуті стандартні невизначеності вимірювання типів А та В, не було застосовано поняття невизначеності та основні вимоги до нього для розв'язання задач семантичного аналізу тексту.

Об'єктом дослідження обрано процес побудови лінгвістичних баз знань когнітивної системи, предмет дослідження – оцінка невизначеності формальних ознак сенсу ПМК. Мета роботи полягає в отриманні оцінки невизначеності вимірювання сенсу ПМК як компонентів ШКС. Для досягнення поставленої мети необхідно ввести формальне поняття ШКС, обґрунтувати метод вимірювання сенсу ПМК на основі нечіткого відношення, отримати та інтерпретувати формальні оцінки невизначеності результатів вимірювання сенсу ПМК.

1 ПОСТАНОВКА ЗАДАЧІ

На вхід будь-якої системи S_i з N відомих подається деякий потік $X = \{x_1, x_2, ...\}$, що на момент часу t_L може бути визначений графом $G_Q(V, E)$ та відповідною матрицею суміжності A_Q розмірністю $L \times L$. Відомо також, що в розрідженій матриці A_Q кількість ненульових lj-х елементів дорівнює m, а кожний з них набуває значення k_{lj}. Потрібно отримати оцінки невизначеності σ результатів спостережень k_{lj} кожної системи S_i, а також обчислити стандартні невизначеності типу А – $u_A(X)$ та типу В – $u_B(X)$ для всіх систем. З огляду на мету дослідження необхідно інтерпретувати та проаналізувати формальні результати у термінах предметної області комп'ютерної лінгвістики.

2 ОГЛЯД ЛІТЕРАТУРИ

Розглянемо основні вимоги до поняття невизначеність вимірювання, викладені у [4, 5]. Ідеальний метод оцінювання невизначеності результату вимірювання повинен бути універсальним: придатним для всіх видів вимірювань і для всіх типів вхідних даних, що використовуються у вимірюваннях. Внутрішня узгодженість величини, що безпосередньо використовується для вираження невизначеності, передбачає можливість прямого використання невизначеності одного результату як компонента оцінювання невизначеності іншого, в якому використовується перший результат.

Невизначеність результату вимірювання у загальному випадку складається з кількох компонентів, які можна згрупувати у дві категорії, залежно від способу оцінювання їх числового значення: тип А – компоненти, оцінені статистичними методами; тип В – компоненти, оцінені іншими способами. Кожний детальний звіт про невизначеності повинен містити повний перелік компонентів і для кожного з них – метод, який використовувався при одержанні його числового значення.

Компоненти категорії А зазвичай характеризуються оціненими дисперсіями σ_i^2 (або оціненими «стандартними відхиленнями» σ_i) і числом степенів вільності. У випадку необхідності слід зазначати коваріації. Компоненти категорії В повинні характеризуватися величинами U_j^2, які можна розглядати як наближення до відповідних дисперсій, існування яких допускається. Величини U_j^2 можна розглядати як дисперсії, а U_j – як стандартні відхилення. При необхідності, коваріації повинні розглядатися аналогічно.

Комбінована невизначеність повинна характеризуватися числовим значенням, одержаним при застосуванні звичайного методу для складання дисперсій. Комбінована невизначеність і її компоненти повинні виражатися у формі «стандартних відхилень». Якщо в окремих випадках для одержання загальної невизначеності комбіновану невизначеність необхідно множити на коефіцієнт, то коефіцієнт множення повинен бути завжди зазначений. Загалом слово невизначеність (*uncertainty*) означає сумнів, і, таким чином, у широкому сенсі «невизначеність вимірювання» означає сумнів щодо вірогідності результату вимірювання (*uncertainty measuring*).

Отже, невизначеність результату вимірювання необов'язково є свідченням правдоподібності того, що результат вимірювання близький до значення вимірюваної величини; це просто оцінювання близькості результату вимірювання до найкращого значення, що відповідає наявним на цей час знанням. Введення поняття «невиз-

наченість вимірювання» є вимушеною мірою, необхідною для одноманітного і спрощеного оцінювання достовірності вимірювання (*evaluation of measuring authenticity*), оскільки її визначення здійснюється на основі одержуваних результатів вимірювання, відомих умов вимірювань і характеристик використовуваної апаратури, а не на невідомому дійсному значенні вимірюваної величини [6].

Для оцінювання x_i вхідної величини X_i, яка не була отримана в результаті повторних спостережень, пов'язані з ними оцінена дисперсія $u^2(x_i)$ або стандартна невизначеність $u(x_i)$ визначаються на базі наукового судження, що базується на всій доступній інформації про можливу змінність X_i. Тобто, стандартну невизначеність типу В одержують із передбачуваної функції щільності ймовірності, заснованої на мірі впевненості в тому, що подія обов'язково відбудеться (ця ймовірність часто називається суб'єктивною ймовірністю).

Оскільки інформацію для оцінки невизначеності вимірювання можуть складати дані попередніх вимірювань, розглянутий у [2] підхід дозволяє забезпечити процес вимірювання сенсу ПМК на основі нечіткої міри. Так, в [1] бінарне нечітке відношення, що задане на одній базисній множині (універсумі) мовних образів I, визначено як нечітке відношення

$$Q = \{< i_l, i_j >, \mu_Q(< i_l, i_j >)\}, \quad (1)$$

де $\mu_Q(< i_l, i_j >)$ – функція належності бінарного нечіткого відношення, що задається як відображення $\mu_Q : I \times I \to [0,\ 1]$. У виразі (1) через $< i_l, i_j >$ позначено кортеж з двох елементів, причому $i_l \in I$, $i_j \in I$. Якщо носій Q_s нечіткого відношення Q є скінченним, то потужність цього нечіткого відношення чисельно дорівнює кількості кортежів його носія і позначається як $card(Q_s)$.

Якщо бінарне нечітке відношення (1) є базовою когнітивною характеристикою ШКС, тоді функцію належності $\mu_Q(< i_l, i_j >)$ варто вважати природною чисельною мірою сенсу. Значення $\mu_Q(< i_l, i_j >) = 1$, згідно з [1], отримало назву одиниці сенсу розміром один *Сав*. В загальному вигляді функція належності нечіткого відношення сенсу для пари мовних образів (на базовому рівні) задається як

$$\mu_Q(< i_l, i_j >) = f(k_{lj}, t_L), \quad (2)$$

де k_{lj} – кількість зафіксованих ШКС зв'язків між l-м та j-м образами на момент часу t_L. Значення k_{lj} неважко отримати шляхом підрахунку кількості зафіксованих ШКС кортежів $< i_l,\ i_j >$ на основі технологічних можливостей сучасних лінгвістичних пакетів, що дозволяє вперше застосувати та обгрунтувати поняття невизначеності вимірювання сенсу ПМК.

3 МАТЕРІАЛИ І МЕТОДИ

3.1 Поняття штучної когнітивної системи: формалізація та інтерпретація

Розглянемо систему S, яку в подальшому будемо називати штучною когнітивною системою, з точки зору процесів накопичення її бази знань. Нехай S здатна розпізнавати образи з нескінченної множини $I = \{i_1, i_2, ..., i_L, ...\}$ та сприймати асоціативні зв'язки між парами образів як елементи множини $\omega \in \Omega$, де $\Omega \subseteq I \times I$ – простір упорядкованих пар. Для визначення образної конструкції застосуємо поняття F – сигма-алгебри (σ-алгебри) підмножин з Ω. Далі будемо вважати образною конструкцією будь-яку підмножину $\gamma \subseteq \Omega$, що має властивість $\gamma \in \mathbf{F}$. Якщо, згідно з властивостями у-алгебри [7], множини $A, B \in \mathbf{F}$, то об'єднання, перетин і різниця A та B у теоретико-множинному сенсі також належать F.

Припустимо, що система S обмінюється інформацією із зовнішнім світом як чорним ящиком виключно у вигляді образних конструкцій, з яких розрізняють послідовність вхідних подій $X = \{x_1, x_2, ...\}$ та множину образних реакцій системи $Y = \{y_1, y_2, ...\}$, причому $x_i \in \mathbf{F}$, $y_i \in \mathbf{F}$. На рис.1 зображено схему абстрактної моделі когнітивної діяльності, що включає у свій склад зовнішній «чорний ящик» та внутрішню ШКС, на вхід якої неперервно подається множина образів подій у вигляді потоку X. На виході ШКС з'являються образи Y, які є реакцією цієї системи на зовнішню ситуацію X згідно з підходом до моделювання образного мислення людини [2].

Закладемо як базовий *онтогенетичний принцип* побудови ШКС – когнітивний ресурс Ω системи S, що визначає сенс її функціонування, отримується виключного шляхом послідовного накопичення параметрів чергових ω з зовнішнього «чорного ящика» та подальшого самовдосконалення множини Ω. Формально онтогенетичний принцип відображається в тому, що базу знань системи S будуємо як $C = \bigcup_{i=1}^{m'} x_i$, де m' – загальна кількість сприйнятих системою на даний час вхідних образних конструкцій.

З метою розв'язання прикладних задач комп'ютерної лінгвістики інтерпретуємо складові отриманої абстрактної моделі когнітивної діяльності. Для ШКС лінгвістичного типу образом i пропонується вважати мовний образ, що наближено задається лексемою або словоформою [8]. Тоді аналогом асоціативного зв'язку між парами образів ω є словосполучення, а образної конструкції γ – речення, мовне висловлювання, загалом – ПМК. Накопичення ШКС когнітивного ресурсу Ω відбувається шляхом опрацювання множини текстів, а наслідком цього є побудова лінгвістичної бази знань C.

На відміну від відомих моделей знань комп'ютерної лінгвістики, в яких словник словоформ поєднується з

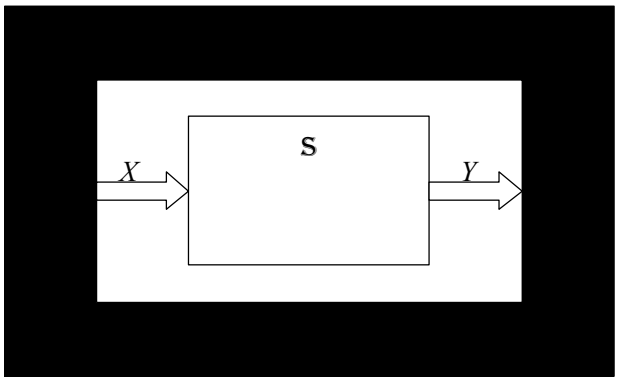

Рисунок 1 – Схема абстрактної моделі когнітивної діяльності

множинами морфологічних, синтаксичних та семантичних правил, основу бази знань C складають виключно асоціативні знання про сполучуваність мовних образів i. Це надає підстави для уніфікованої оцінки одиниці та кількості сенсу ПМК.

3.2 Метод вимірювання сенсу ПМК на основі нечіткого відношення

У відповідності до запропонованого підходу [9] деталізуємо функцію належності, що породжує бінарне нечітке відношення сенсу (1) на таких 3-х послідовних рівнях, побудованих на базовому (2):

1. Рівень імовірнісного прогнозування – з метою нормування функції належності у проміжку [0, 1] передбачено розрахунок статистичної оцінки λ (математичного сподівання): якщо для nt відомих ШКС на момент часу t_L образів $k_\Sigma = \sum_{l=1}^{nt}\sum_{j=1}^{nt} k_{lj}$, а m – кількість усіх ненульових кортежів $< i_l,\ i_j >$, то $\lambda = k_\Sigma / m$ – в цьому випадку застосуємо відому сигмоїдальну функцію [10]

$$\mu_Q(< i_l, i_j >) = f_1(k_{lj}, \lambda) = 1/(1 + e^{-k_{lj}+\lambda}), \qquad (3)$$

Внаслідок нормування з'являється характерна властивість функції належності, отриманої за методом, що пропонується – середнє значення $\overline{\mu_Q} = \frac{1}{m}\sum_{j=1}^{m}\mu_{Qj} = 0{,}5$.

2. Рівень врахування емоційного стану – введено можливість врахування бінарної моделі емоцій ШКС [9] за рахунок показника $\mu = \{...,-2,-1,1,2,...\}$, тоді

$$\mu_Q(< i_l, i_j >) = f_2(k_{lj}, \lambda, \mu) = 1/(1 + e^{-\frac{k_{lj}-\lambda}{|\mu|}}). \qquad (4)$$

При $\mu = -1 \vee 1$ емоції не впливають на сенс функціонування ШКС, а функція належності (4) вироджується у функцію (3). Збільшення показника μ симетрично згладжує сигмоїдальну функцію f_2, що продемонстровано на рис. 2.

3. Рівень врахування мотиваційної компоненти на основі образів-центрів потреб – запропоновано моделлю мотиву ШКС на момент часу t_L вважати досягнення образу-центру потреби j', а також розрахувати дисперсію та середньоквадратичне відхилення результатів спостережень k_{lj} як

$$D = \frac{1}{m}\sum_{l=1}^{nt}\sum_{j=1}^{nt}(k_{lj} - \lambda)^2 \mid \ k_{lj} > 0 \ \text{і} \ \sigma = \sqrt{D}. \qquad (5)$$

Отримане значення σ будемо вважати невизначеністю, що обумовлена недосконалістю моделі мотиву ШКС. Характеризує цю невизначеність зокрема недосконалість базової залежності (3), на основі якої пропонується врахувати мотиваційну компоненту на основі образів-центрів потреб.

В залежності від ступеня наближення r пари образів $< i_l,\ i_j >$ до j', функцію (4) можна зміщувати вліво за віссю абсцис шляхом зменшення математичного сподівання для цієї пари $\lambda_{lj} = \lambda - r \cdot \sigma$, де $r = \{0,\ 1,\ 2,\ 3\}$, зрештою маємо

$$\mu_Q(< i_l, i_j >) = f_3(k_{lj}, \lambda_{lj}, \sigma, \mu, i') = 1/(1 + e^{-\frac{k_{lj}-\lambda_{lj}}{|\mu|}}). \qquad (6)$$

Питання побудови окремого алгоритму для визначення ступеня наближеності r пари $< i_l,\ i_j >$ до образу-потреби j' та введення додаткового рівня врахування рефлексів та результатів зовнішнього навчання розглянуто у [9]. Зауважимо, що, на відміну від (3) та (4), у функції належності відношення сенсу (6) внаслідок локальних зсувів математичного сподівання зникає властивість $\overline{\mu_Q} = 0{,}5$, що, на думку авторів, свідчить про належну формальну інтерпретацію відомих фактів з психології та фізіології щодо протиріч між загальноприйнятим

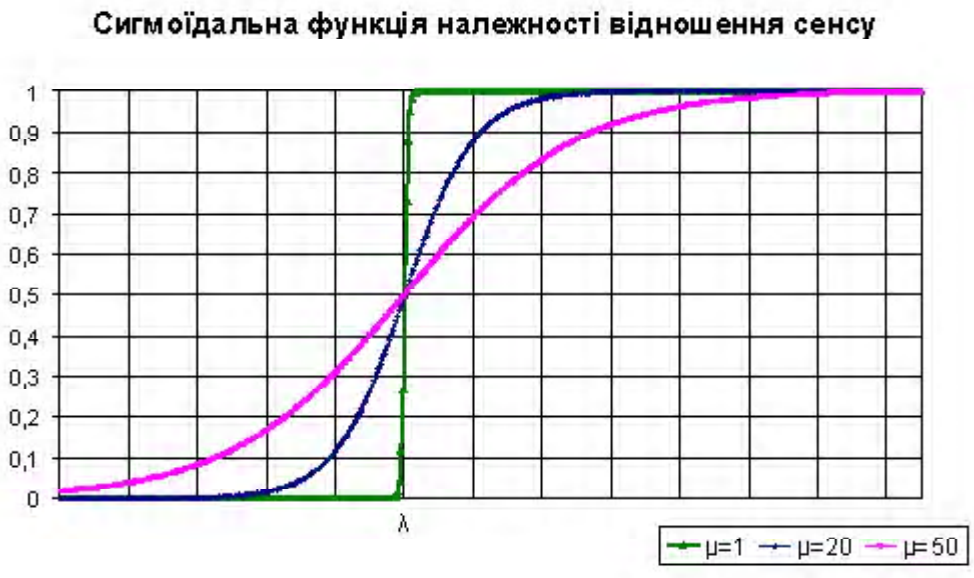

Рисунок 2 – Вплив показника μ на функцію належності (4)

(середньостатистичним) сенсом і діями під впливом сильних мотивів.

3.3 Невизначеність результатів вимірювання сенсу ПМК

Розглянутий підхід до вимірювання сенсу відповідає лінгвістичній базі знань однієї ШКС, вихідними даними якої може бути як окремий текст, так і деяка унікальна множина текстів. При цьому потрібно розуміти, що за кожним текстом стоїть так само унікальний світогляд автора, втілений у його мові. Для розв'язання задачі виявлення інформативних ознак тексту важливим є визначення достовірності бази знань у цілому та сенсу однієї пари мовних образів у вигляді $\mu_Q(< i_l, i_j >)$, як базової складової цих знань зокрема. Оскільки фактично йдеться про вимірювання сенсу, то для оцінки достовірності пропонується застосувати поняття невизначеності результатів множинного вимірювання сенсу ПМК.

У першому наближенні будемо вважати, що суб'єктивна оцінка кількості сенсу однієї пари мовних образів втілюється у статистичний ряд чисельних значень для N різних ШКС. Отже, для довільного кортежу $< i_l, i_j >$ вимірювана згідно (3) величина $Y = \mu_Q(< i_l, i_j >)$ функціонально залежить від результатів її багаторазових вимірювань $X_1, X_2, ..., X_N$ для різних ШКЛ та, в загальному випадку, має такий вигляд

$$Y = f(X_1, X_2, ..., X_N). \quad (7)$$

Оцінку вимірюваної величини Y, позначену y, одержимо із загального рівняння (7), використовуючи вхідні оцінки $x_1, x_2, ..., x_N$ для N значень величин $X_1, X_2, ..., X_N$. Отже, вихідна оцінка y, яка є результатом вимірювання, виражається таким чином

$$y = f(x_1, x_2, ..., x_N).$$

Базовою оцінкою математичного сподівання або очікуваного значення μ_Q величини q, що змінюється випадковим чином, є середнє арифметичне або середнє значення $\overline{q}$ із n спостережень

$$\overline{q} = \frac{1}{n}\sum_{k=1}^{n} q_k . \quad (8)$$

Експериментальне стандартне відхилення, що характеризує змінність значень q_k, або, точніше, їхню дисперсію σ^2 щодо середнього значення $\overline{q}$, розраховують за формулою [6]

$$u_A(q_k) = \sqrt{\frac{\sum_{k=1}^{n}(q_k - \overline{q})^2}{n-1}} . \quad (9)$$

Оскільки за результат багаторазових вимірювань приймають середнє значення $\overline{q}$, то важливо оцінити його дисперсію. Найкраща оцінка $\sigma^2(\overline{q}) = \sigma^2/n$ дисперсії середнього значення $u_A^2(\overline{q})$ виражається як

$$u_A^2(\overline{q}) = \frac{u_A^2(q_k)}{n} . \quad (10)$$

Експериментальна дисперсія середнього $u_A^2(\overline{q})$ і експериментальне стандартне відхилення середнього значення $u_A(\overline{q})$, що дорівнює позитивному квадратному кореню з оцінки дисперсії $u_A^2(\overline{q})$, кількісно визначають, наскільки добре $\overline{q}$ оцінює очікування μ_Q величини q. З урахуванням виразів (9) та (10) експериментальне стандартне відхилення середнього значення $u_A(\overline{q})$ розраховується за формулою [6]

$$u_A(\overline{q}) = \sqrt{\frac{\sum_{k=1}^{n}(q_k - \overline{q})^2}{n(n-1)}} . \quad (11)$$

Для більш глибокого врахування суб'єктивного характеру вимірюваного сенсу кортежів у функції (7) застосуємо складові стандартної невизначеності типу В, які, як правило визначають|обчисляють,вичисляють| на основі інформації про верхні і нижні границі $[\alpha_-; \alpha_+]$ передбачуваного закону розподілу чи через інтервал U, що має заданий довірчий рівень довіри p.

Для визначення стандартної невизначеності типу В потрібно взяти позитивний квадратний корінь з добутку довірчого рівня кожного значення та квадрата відхилення цього значення і всі добутки такого виду додати. В результаті загальний вигляд формули для обчислення стандартної невизначеності типу В при дискретних даних має вигляд

$$u_B(X) = \sqrt{\sum_{i=1}^{n}\left(x_i - \sum_{i=1}^{n} x_i p_i\right)^2 p_i} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 p_i} . \quad (12)$$

Якщо для значення величини X_i можна оцінити верхню та нижню границю $[\alpha_-; \alpha_+]$, то стандартні невизначеності типу В, в припущені про можливий вигляд закону розподілу, можна визначити за формулами [4, 5, 6]:

а) для трикутного закону розподілу

$$u_B(X_i) = \frac{\alpha_+ - \alpha_-}{\sqrt{24}} ; \quad (13)$$

б) для експоненціального закону розподілу

$$u_B(X_i) = \sqrt{\frac{(\alpha_+ - x)(x - \alpha_-) - (\alpha_+ - 2x + \alpha_-)}{\lambda}} , \quad (14)$$

де x – очікуване значення, а λ – параметр розподілу;

в) для закону розподілу Парето

$$u_B(X_i) = \frac{x_m}{k-1}\sqrt{\frac{k}{k-2}}, \quad (15)$$

де x_m – початкове значення x, а k – параметр розподілу (значення щільності для x_m);

д) для рівномірного закону розподілу

$$u_B(X_i) = \frac{\alpha_+ - \alpha_-}{\sqrt{12}}. \quad (16)$$

Для заданих інтервалів U_p із відомим рівнем довіри p, в припущенні нормального закону розподілу, невизначеність типу В визначається за формулою

$$u_B(X_i) = \frac{U_p}{k_p},$$

де k_p – коефіцієнт охоплення, який для нормального закону розподілу, дорівнює 1,64; 1,96; 2,58 і 3 для довірчих рівнів 0,9; 0,95; 0,99 і 0,9973. За відсутності інформації про наявність законів (13)–(16) розподілу вхідної величини X_i для симетричних границь $\pm\alpha_i$ стандартну невизначеність типу В визначають за формулою

$$u_B(X_i) = \frac{2\alpha_i}{\sqrt{12}} = \frac{\alpha_i}{\sqrt{3}}, \quad (17)$$

яка може бути застосована на початковому етапі експериментального дослідження ШКС.

4 ЕКСПЕРИМЕНТИ

З метою експериментальної перевірки результатів оцінки невизначеності вимірювання сенсу ПМК як компонентів ШКС за допомогою запропонованого методу було застосовано відомий лінгвістичний пакет DKPro Core, який базується на платформі Apache UIMA framework [12]. Для реалізації серії експериментів було розроблено додаткову *Java*-програму (додаток 1), що використовує та удосконалює колекцію програмних компонентів для обробки природної мови DKPro Core [13]. Особливість розробленої програми, що орієнтована на технологію Java/Maven/Eclipse, полягає у визначенні списку лем тексту та складних залежностей згідно [14] між цими лемами у вигляді списку з m зв'язків.

Експериментальною базою було обрано три відомі літературні твори з відкритого джерела *Project Gutenberg* [15], а саме англомовні (авторські) варіанти 4-х текстів різного обсягу: «Аліса в країні див» (Л. Керол, 1 – уривок з 4204 слів та 2 – повна версія з 26690 слів), 3 – «Біле ікло» (Дж. Лондон, 48907 слів) та 4 – «Троє у човні без врахування собаки» (Дж. К. Джером, 67328 слів). Мета серії експериментів полягала у дослідженні базових характеристик невизначеності кожного з 4-х текстів, а також у отриманні оцінки невизначеності множини спільних для всіх текстів пар мовних образів $< i_l, i_j >$ згідно з запропонованим методом.

5 РЕЗУЛЬТАТИ

У результаті дослідження формалізовано та інтерпретовано для предметної галузі комп'ютерної лінгвістики поняття штучної когнітивної системи, закладено базовий *онтогенетичний принцип* побудови ШКС. Отримано формальні характеристики методу створення бінарного нечіткого відношення образного сенсу Q ШКС S_Q шляхом моделювання понять мотиваційної мети та емоційного стану. Запропоновано принципи послідовної багаторівневої побудови функції належності $\mu_Q(< i_l,\ i_j >)$, що породжує нечітке відношення Q, визначено характерну властивість $\overline{\mu_Q} = 0{,}5$ методу вимірювання сенсу ПМК. Згідно з ним для задачі виявлення інформативних ознак тексту отримано формальні теоретичні оцінки невизначеності σ результатів спостережень k_{lj} кожної ШКС S_i, а також розраховані стандартні невизначеності типу А – $u_A(X)$ та типу В – $u_B(X)$ для всіх ШКС.

За допомогою розробленого в [13] програмного забезпечення на основі пакету DKPro Core було отримано результати обробки 4-х обраних англомовних текстів, що можуть інтерпретуватися як 4 різні ШКС. Основні результати обробки у відповідності до (5) представлено в табл. 1, де 3 останні стовпці вміщують такі дані:

– відсоток СКВ σ від оцінки математичного сподівання λ;

– кількість визначених засобами DKPro Core лем тексту;

– середня кількість різних зв'язків для однієї леми тексту.

Отримані гістограми експериментальних законів щільності розподілу показали значну зовнішню схожість до розподілу за законом Парето, що демонструє приклад порівняння експериментального результату для тексту 1 (Carrol_part) з теоретичною щільністю розподілу Парето зі значенням параметру $k = 2108$.

Аналіз відсортованих за спаданням k_{lj} списків пар мовних образів $< i_l, i_j >$ дозволив виявити 4 спільні пари у верхній частині списків, вихідні дані та результати оцінки $\overline{q}$ за (8) та невизначеності за типами А та В згідно з (11) і (12) яких представлено в табл. 2.

Таблиця 1 – основні результати обробки 4-х англомовних текстів

Текст		m	k_Σ	λ	σ	%	Кільк. лем	Сер. кільк. зв'язків
1	*Carrol_part*	2360	2812	1,191525424	0,778805721	65,36%	762	3,0971
2	*Carrol_full*	12156	17786	1,463145772	2,245695112	153,48%	2121	5,7313
3	*London*	25244	31234	1,237284107	1,259517221	101,80%	5702	4,4272
4	*Jerom*	33316	47091	1,413465002	2,044626970	144,65%	6048	5,5086

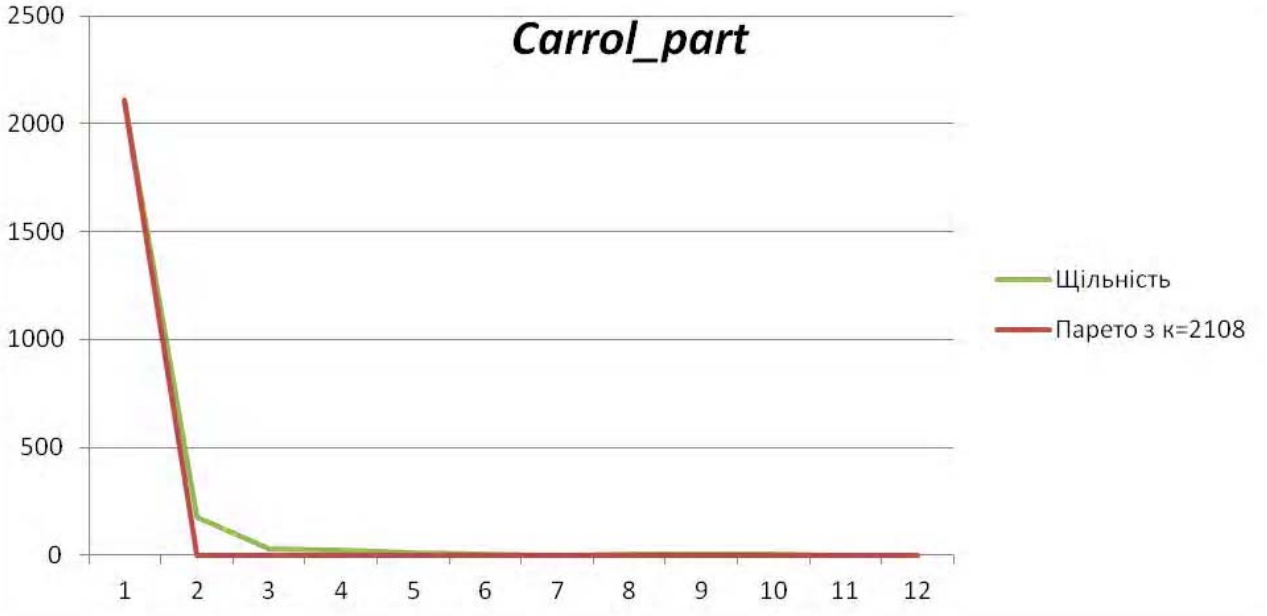

Рисунок 3 – Аналіз експериментального закону щільності розподілу для тексту 1

Таблиця 2 – Результати оцінки невизначеності 4-х обраних пар мовних образів

Текст		*go-back*	*say-I*	*know-I*	*see-I*
1	*Carrol_part*	0,859177412	0,943132060	0,943132060	0,943132060
2	*Carrol_full*	0,998553057	0,999999514	0,386239825	0,998553057
3	*London*	0,999999948	1,000000000	1,000000000	0,999999993
4	*Jerom*	0,999999537	1,000000000	1,000000000	1,000000000
$\overline{q}$		0,928865234	0,971565787	0,664685943	0,970842558
$u_A(X)$		0,028449934	0,01160802	0,113675151	0,011312764
		0,036641087	0,01495014	0,146403893	0,014569874

6 ОБГОВОРЕННЯ

Отримані в результаті експерименту чисельні оцінки невизначеності результатів вимірювання сенсу пар мовних образів дозволяють отримати нову інформацію щодо текстів, що аналізуються. Представлення кожного тексту, як окремої ШКС демонструє, що експериментальний закон щільності розподілу для характеристики k_{lj} пар мовних образів дуже подібний до розподілу Парето. Однак такому висновку не відповідають оцінки математичного сподівання λ, яке б мало зменшуватися та наближатися до 1 ($\lambda_{Pareto} = \frac{k \cdot x_m}{k-1}$) зі збільшенням числа пар [16], а також СКВ σ, яке занадто велике для розподілу Парето. Наприклад, для тексту 1 згідно з (5) $\sigma = 0{,}7788$, що складає 65,36% від λ. Аналогічні оцінки відповідно до залежностей (15) для розподілу Парето та (17) для загального випадку з малим значенням $\alpha_i = \pm 0{,}01$: $\sigma_1 = 0{,}0004748$ (0,04%) та $\sigma_2 = 0{,}58$ (0,48%).

Проте, аналіз даних таблиці 1 надає формальні підстави для висунення *гіпотези* – найбільш інформативною характеристикою ШКС є середня кількість зв'язків для однієї леми (мовного образу). Обґрунтування – коефіцієнт кореляції Пірсона для стовпчиків з λ та «Кількість лем» для всіх 4-х ШКС дорівнює 0,198, але для пар стовпчиків λ та «Середня кількість зв'язків» – 0,945. Одночасно для пар стовпчиків σ та «Середня кількість зв'язків» коефіцієнт кореляції дорівнює 0,984, а для пар стовпчиків «%» та «Середня кількість зв'язків» – 0,996. Це дозволяє вважати, що закон розподілу є лише Парето-подібним, проте невизначеність сенсу ШКС (параметр σ) прямо пропорційна середній кількості зв'язків. Висунута *гіпотеза* потребує масштабнішої експериментальної перевірки та уточнення.

Дані таблиці 2 демонструють високу ступінь сенсоподібності згідно з висунутим підходом для 4-х обраних пар мовних образів, що використовувалися 3-ма різними авторами. Загальна тенденція полягає у тому, що оцінки невизначеності $u_A(X)$ за типом А менші відповідних оцінок $u_B(X)$ за типом В для всіх ШКС приблизно у 1,5 рази. При цьому відсоток невизначеності не перевищує 4% від оцінки математичного сподівання $\overline{q}$ для всіх пар $\mu_Q(<i_l, i_j>)$, окрім пари «know-I» (до 22,03%), що має зрозуміле пояснення – в обраному уривку 1 тексту Л. Керола ця пара зустрічається відносно набагато частіше, аніж у цьому ж творі 2 («Аліса у країні казок») у цілому. Такі результати дозволяють сподіватися, що зап-

ропонований підхід дозволить підвищити якість розв'язання задач автоматичного семантичного аналізу текстів, зокрема визначення авторства. Однак цілком ймовірно, що аналогічне порівняння пар, що зустрічаються рідко (знаходяться у нижній частині відсортованих списків) продемонструє високу невизначеність.

Подальших досліджень потребує також визначення законів розподілу експериментальних значень $\mu_Q(<i_l,i_j>)$ та отримання суб'єктивних характеристик бази знань ШКС у представленні динамічної невизначеності вимірювань.

ВИСНОВКИ

Внаслідок проведених досліджень розв'язано актуальну задачу оцінки невизначеності вимірювання сенсу ПМК як компонентів ШКС, яка безпосередньо пов'язана з проблемою розуміння смислу текстової інформації. Набув подальшого розвитку метод вимірювання сенсу ПМК на основі нечіткого відношення, який, на відміну від існуючих, базується на введених формальних понятях штучної когнітивної системи та мовного образу, що дозволяє отримати вихідні статистичні дані для оцінки результатів невизначеності вимірювання типів А та В. Уперше отримано та інтерпретовано формальні оцінки невизначеності результатів вимірювання сенсу ПМК, що дозволяє врахувати інформацію про зв'язки між лемами тексту для розв'язання задачі виявлення інформативних ознак тексту.

Практичне значення отриманих результатів полягає в отриманні програмного технологічного інструментарію на основі лінгвістичного пакету DKPro Core, що дозволяє реалізувати запропонований метод для семантичного аналізу англомовних текстів. За результатами проведеної серії експериментів виявлено, що закон розподілу зв'язків між лемами тексту є Парето-подібним, проте має суттєві формальні відмінності від класичного розподілу Парето, зокрема суттєво більші оцінки математичного сподівання λ (до 46,3 %) та СКВ σ (на кілька порядків).

З точки зору запропонованого підходу до визначення сенсу ПМК збільшення розмірів тексту за кількістю слів та, відповідно, його словникового складу за кількістю лем не впливає на параметри закону розподілу та невизначеність сенсу окремої ШКС. Аналіз отриманих результатів дозволяє вважати показник середньої кількості зв'язків мовного образу найбільш інформативною ознакою тексту, оскільки коефіцієнт кореляції Пірсона між ним та параметрами, пов'язаними з невизначеністю сенсу більший за 0,945.

Порівняння оцінок невизначеності 4-х пар мовних образів, що використовувалися 3-ма різними авторами, показало високу ступінь сенсоподібності таких пар згідно з висунутим підходом. При цьому оцінки невизначеності $u_A(X)$ за типом А пропорційно менші відповідних оцінок $u_B(X)$ за типом В для всіх ШКС приблизно у 1,5 рази, що дозволяє обмежитися знаходженням тільки однієї оцінки невизначеності $u_A(X)$.

Отримані результати досліджень, а саме формальні показники невизначеності сенсу та середня кількість зв'язків мовного образу, мають перспективи використання в задачах семантичного аналізу ПМК, зокрема кластеризації, класифікації та визначення авторства текстів.

ПОДЯКИ

Проведені дослідження здійснювались у межах держбюджетної науково-дослідної роботи Вінницького національного технічного університету за темою «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-мовного контенту» (№ держреєстрації 0114U003462), а також у відповідності до плану досліджень науково-дослідного центру прикладної та комп'ютерної лінгвістики ВНТУ.

ДОДАТОК 1

Текст *Java*-програми *MyBasePipeline3*:

```
package de.tudarmstadt.ukp.tutorial.gscl2013.dkpro;
import static org.apache.uima.fit.factory.AnalysisEngineFactory.createEngineDescription;
import static org.apache.uima.fit.factory.CollectionReaderFactory.createReaderDescription;
import static org.apache.uima.fit.util.JCasUtil.select;
import java.io.FileNotFoundException;
import java.io.PrintWriter;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.Collection;
import java.util.HashMap;
import java.util.HashSet;
import java.util.List;
import java.util.Map;
import java.util.Set;
import javax.xml.transform.TransformerConfigurationException;
import org.apache.uima.fit.pipeline.JCasIterable;
import org.apache.uima.fit.util.JCasUtil;
import org.apache.uima.jcas.JCas;
import org.jgrapht.ext.GraphMLExporter;
import org.jgrapht.ext.IntegerEdgeNameProvider;
import org.jgrapht.ext.IntegerNameProvider;
import org.jgrapht.ext.StringEdgeNameProvider;
import org.jgrapht.ext.StringNameProvider;
import org.jgrapht.graph.ClassBasedEdgeFactory;
import org.jgrapht.graph.DefaultDirectedWeightedGraph;
import org.xml.sax.SAXException;
import de.tudarmstadt.ukp.dkpro.core.api.coref.type.CoreferenceChain;
import de.tudarmstadt.ukp.dkpro.core.api.coref.type.CoreferenceLink;
import de.tudarmstadt.ukp.dkpro.core.api.segmentation.type.Lemma;
import de.tudarmstadt.ukp.dkpro.core.api.segmentation.type.Sentence;
import de.tudarmstadt.ukp.dkpro.core.api.segmentation.type.Token;
import de.tudarmstadt.ukp.dkpro.core.api.syntax.type.dependency.Dependency;
import de.tudarmstadt.ukp.dkpro.core.io.text.TextReader;
import de.tudarmstadt.ukp.dkpro.core.opennlp.OpenNlpPosTagger;
import de.tudarmstadt.ukp.dkpro.core.stanfordnlp.StanfordCoreferenceResolver;
import de.tudarmstadt.ukp.dkpro.core.stanfordnlp.StanfordParser;
import de.tudarmstadt.ukp.dkpro.core.tokit.BreakIteratorSegmenter;
public class MyBasePipeline3 {
private static class LinkCounter {// класс для связи Map с
графом DefaultDirectedWeightedGraph
private Map<String, Map<String, Integer>> links;
private   DefaultDirectedWeightedGraph<String,
RelationshipEdge<String>> g;
LinkCounter() {//
links = new HashMap<String, Map<String, Integer>>();
g  =  new  DefaultDirectedWeightedGraph<String,
RelationshipEdge<String>>(
new ClassBasedEdgeFactory(RelationshipEdge.class));
}
void addLink(String from, String to, String type) {// метод для
добавления новой связи
from = from.toLowerCase();
to = to.toLowerCase();
Map<String, Integer> mapFrom = links.get(from);
if (mapFrom == null) {
mapFrom = new HashMap<String, Integer>();
links.put(from, mapFrom);
}
```

```
Integer countTo = mapFrom.get(to);
mapFrom.put(to, (countTo == null) ? 1 : countTo + 1);
if (!g.containsVertex(from)) {
g.addVertex(from);
}
if (!g.containsVertex(to)) {
g.addVertex(to);
}
RelationshipEdge<String> edge = new  RelationshipEdge<String>(from,
to, type);
g.addEdge(from, to, edge);
}
void saveLinks(String filename)
throws FileNotFoundException, UnsupportedEncodingException { /
/ метод для запоминания статистики связей в файле links.csv
PrintWriter writer = new PrintWriter(filename, "UTF-8");
for (String keyFrom : links.keySet()) {
Map<String, Integer> mapFrom = links.get(keyFrom);
for (String keyTo : mapFrom.keySet()) {
Integer count = mapFrom.get(keyTo);
writer.println(keyFrom + ", " + keyTo + ", " + count);
}
}
writer.close();
}
/* метод - в графе искать v2 и заменять программно на v1, т.е.
найти все связи для v2, запомнить, удалить v2, вставить узел
v1, добавить связи */
void V1ChangeV2(String v1, String v2) {
try {
v1 = v1.toLowerCase();
v2 = v2.toLowerCase();
Set<RelationshipEdge<String>> targ1 = g.outgoingEdgesOf(v2);
if(targ1!=null) {
Set<RelationshipEdge<String>>  targ1copy  =  new
HashSet<RelationshipEdge<String>>();
targ1copy.addAll(targ1);
for(RelationshipEdge v : targ1copy) {
addLink(v1, g.getEdgeTarget(v), v.toString());
System.out.printf("%n"+v1+" "+g.getEdgeTarget(v)+" "+v);
g.removeEdge(v);
//delLink(v2, g.getEdgeTarget(v), v.toString());
}
}
Set<RelationshipEdge<String>> targ2 = g.incomingEdgesOf(v2);
if(targ2!=null) {
 for(RelationshipEdge v : targ2) {
addLink(g.getEdgeSource(v), v1, v.toString());
System.out.printf("%n"+g.getEdgeSource(v)+" "+v1+" "+v);
//g.removeEdge(v);
//delLink(g.getEdgeSource(v), v2, v.toString());
}
}
g.removeVertex(v2);
} catch (Exception e) {
System.out.println("\nНесподіванка спіткала українських науковців
під час зміни " + v1 + " на " + v2);
//e.printStackTrace();
}
}
private void saveGraph(String filename)
throws FileNotFoundException, UnsupportedEncodingException,
SAXException, TransformerConfigurationException {// метод для
запоминания графа в файле graph.xml
GraphMLExporter<String, RelationshipEdge<String>> me =
new GraphMLExporter<String, RelationshipEdge<String>>(
new IntegerNameProvider<String>(),
new StringNameProvider<String>(),
new IntegerEdgeNameProvider<RelationshipEdge<String>>(),
new StringEdgeNameProvider<RelationshipEdge<String>>());
PrintWriter writer = new PrintWriter(filename, "UTF-8");
me.export(writer, g);
}
Map<String, Map<String, Integer>> getLinks() {
return links;
}
DefaultDirectedWeightedGraph<String, RelationshipEdge<String>>
getGraph() {
return g;
}
}
public static void main(String[] args) throws Exception {//
главный метод класса MyBasePipeline3
JCasIterable pipeline = new JCasIterable(// запуск программного
конвейера для последовательного аннотирования
// (создания многоуровневой разметки) текста
createReaderDescription(TextReader.class,
TextReader.PARAM_SOURCE_LOCATION, "input/Obama.txt",
TextReader.PARAM_LANGUAGE, "en"), // чтение текста
createEngineDescription(BreakIteratorSegmenter.class),// сегмен-
тирование текста
createEngineDescription(OpenNlpPosTagger.class),// морфологичес-
кая разметка
createEngineDescription(StanfordParser.class,
StanfordParser.PARAM_VARIANT, "rnn",
StanfordParser.PARAM_MODE,
StanfordParser.DependenciesMode.CC_PROPAGATED), // синтаксическая
разметка, учитывающая
// сложные зависимости между парами лемм
// Stem
//createEngineDescription(SnowballStemmer.class),
// Lemma
//createEngineDescription(MateLemmatizer.class),
// NamedEntity
//createEngineDescription(OpenNlpNameFinder.class,
// OpenNlpNameFinder.PARAM_VARIANT, "person"),
//createEngineDescription(OpenNlpNameFinder.class,
//OpenNlpNameFinder.PARAM_VARIANT, "organization"),
//CoreferenceChain, CoreferenceLink
createEngineDescription(StanfordCoreferenceResolver.class)//
поиск соответствия местоимений
// SemanticPredicate, SemanticArgument
//createEngineDescription(ClearNlpSemanticRoleLabeler.class)
);
PrintWriter writer = new PrintWriter("output/output.txt", "UTF-
8");// запись зависимостей по предложениям в файл output.txt
for (JCas jcas : pipeline) {
LinkCounter linkCounter = new LinkCounter();
for (Sentence sentence : select(jcas, Sentence.class)) {
writer.println("sentence: " + sentence.getCoveredText());    /
/ запись исходного предложения
Collection<Token>  tokens  =  JCasUtil.selectCovered(jcas,
Token.class, sentence);
List<String> ts = new ArrayList<String>(tokens.size());
/* исключение неинформативных для анализа типов зависимостей */
List<String> excludes = Arrays.asList(new String[]{"det",
"punct", "cop", "cc", "aux", "auxpass", "expl", "mark", "num",
"number", "quantmod", "ref"/**/});
for  (Dependency  dep  :  JCasUtil.selectCovered(jcas,
Dependency.class, sentence)) {
String type = dep.getDependencyType();
if (!excludes.contains(type)) {
String govLemma = safeVal(dep.getGovernor());
String depLemma = safeVal(dep.getDependent());
String depn1 = dep.getDependencyType();
writer.println(depn1 + "(" + govLemma + ", " + depLemma + ")");
// запись очередной зависимости
linkCounter.addLink(govLemma, depLemma, depn1);
}
}
//}
//System.out.printf("%n — Semantic structure —%n");
//for (SemanticPredicate pred : selectCovered(
//SemanticPredicate.class, sentence)) {
//System.out.printf("  %-16s %-10s", pred.getCoveredText(),
//pred.getCategory());
//for (SemanticArgument arg : select(pred.getArguments(),
//SemanticArgument.class)) {
//System.out.printf("\t%s:%s", arg.getRole(),
//arg.getCoveredText());
//}
//System.out.printf("%n");
//}
}
System.out.printf("%n== Coreference chains (for the whole
document) ==%n");
for  (CoreferenceChain  chain  :  select(jcas,
CoreferenceChain.class)) {// цикл по все найденным кореференци-
ям для имен и местоимений
CoreferenceLink link = chain.getFirst();
//System.out.println(link);
String v1 = "#";
String v2 = "$";
```

```
while (link != null) {
//String v = link.getCoveredText().trim();
String v = link.getCoveredText();
while (v.indexOf(" ")!=-1) v = v.substring(1+v.indexOf(" "));
System.out.printf("\n:: %s |%s|", link.getCoveredText(),
link.getReferenceType());
if     (link.getReferenceType()=="PROPER"     ||
link.getReferenceType()=="NOMINAL") v1=v;
if (link.getReferenceType()=="PRONOMINAL") v2=v;
if (link.getReferenceRelation() != null) {
//System.out.printf("-[%s]", link.getReferenceRelation());
}
link = link.getNext();
}
if (!v1.equals("#") & !v2.equals("$")) {// если кореференции
найдены, то проводим замену местоимений на номиналы
System.out.printf(v1+" "+v2);
linkCounter.V1ChangeV2(v1, v2);// вызов метода V2ChangeV1
}
System.out.printf("%n");
}
linkCounter.saveLinks("output/links.csv");// вызов метода
saveLinks
linkCounter.saveGraph("output/graph.xml");// вызов метода
saveGraph
//GraphAnalyzer.showGraph(linkCounter.getGraph());   // визуа-
лизация графа через вызов метода showGraph класса GraphAnalyzer
}
writer.close();
}
private static String safeVal(Token t) {// функция для опреде-
ления леммы для слова (токена)
Lemma l = t.getLemma();
return l != null ? l.getValue() : "";
}
}
```

СПИСОК ЛІТЕРАТУРИ

1. 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.
2. Бісікало О. В. Інфологічний підхід до моделювання образного мислення людини [Електронний ресурс] / О. В. Бісікало // Вісник СумДУ (Серія «Технічні науки»). – 2009. – № 2. – С. 15–20. – Режим доступу: http://visnyk.sumdu.edu.ua/arhiv/2009/Tech_2_09/09bovoml.pdf.
3. Vasilevskyi O. M. Calibration method to assess the accuracy of measurement devices using the theory of uncertainty. International Journal of Metrology and Quality Engineering, 2014, 5.04: 403. – № 3 (7). – 2006. – P. 147–151.
4. Руководство по выражению неопределенностей измерения = Guide to the Expression of Uncertainly in Measurement : [науч. редактор Слаев В. А.]. – Санкт-Петербург : НПО ВНИИМ им. Д. М. Менделеева, 1999. – 134 с.
5. Василевський О. М. Алгоритм оцінювання невизначеності у вимірюваннях при виконанні метрологічних робіт / О. М. Василевський // Інформаційні технології та комп'ютерна інженерія. – № 3 (7). – 2006. – С. 147–151.
6. Применение «Руководства по выражению неопределенности измерений» : МИ 2552-99. – Офиц. изд. – Санкт-Петербург : ВНИИМ им. Д. И. Менделеева, 1999. – 27 с.
7. Колмогоров А. Н. Основные понятия теории вероятностей / А. Н. Колмогоров. – 2-е изд. – М. : Наука, 1974. – 120 с.
8. Бісікало О. В. Формалізація понять мовного образу та образного сенсу природно-мовних конструкцій / О. В. Бісікало // Математичні машини і системи. – 2012. – № 2. – С. 70–73.
9. Бісікало О. В. Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія / О. В. Бісікало. – Вінниця : ВНТУ, 2013. – 316 с.
10. Раскин Л. Г. Нечеткая математика. Основы теории. Приложения / Л. Г. Раскин, О. В. Серая. – Х. : Парус, 2008. – 352 с.
11. Загальні вимоги до компетентності випробувальних та калібрувальних лабораторій : ДСТУ ISO/IEC 17025-2001. – [Чинний від 2001 – 01 – 01]. – К. : Держстандарт України, 2001. – 31 с. – (Національний стандарт України).
12. Gurevych I. Darmstadt Knowledge Processing Repository Based on UIMA [Electronic resource] / I. Gurevych, M. Muhlhauser, Ch. Muller, J. Steimle, M. Weimer, T. Zesch. – February 9, 2007. – Available at: \www/URL: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf.
13. Бісікало О. В. Метод вилучення образних знань з англомовного тексту на основі інструментальних засобів пакету DKPro Core / О. В. Бісікало, І. Гуревич // Контроль і управління в складних системах: XII міжнар. конф., 14–16 жовтня 2014 р.: тези доповідей. – Вінниця, 2014. – С. 51.
14. Stanford parser [Електронний ресурс] // Stanford Dependencies. – Назва з екрану. – Режим доступу: http://nlp.stanford.edu/software/stanford-dependencies.shtml.
15. Free ebooks – Project Gutenberg [Електронний ресурс] / Project Gutenberg Literary Archive Foundation. – Режим доступу: https://www.gutenberg.org/.
16. Бісікало О. В. Статистичний аналіз складних залежностей у тексті / О. В. Бісікало // Вісник Нац. ун-ту «Львівська політехніка» : Інформаційні системи та мережі. – 2015. – № 814. – С. 228–236.

Стаття надійшла до редакції 19.11.2015.
Після доробки 25.12.2016.

Бисикало О. В.[1], Василевский А. Н.[2]

[1]Д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики Винницкого национального технического университета, Винница, Украина

[2]Канд. техн. наук, доцент, профессор кафедры метрологии и промышленной автоматики Винницкого национального технического университета, Винница, Украина

ОЦЕНКА НЕОПРЕДЕЛЕННОСТИ ИЗМЕРЕНИЯ СМЫСЛА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ КОНСТРУКЦИЙ

Решена задача оценки неопределенности измерения смысла естественно-языковых конструкций (ЕЯК) на основе формализации понятий лингвистического образа, искусственной когнитивной системы и единицы смысла. В основу модели базы знаний искусственной когнитивной системы заложена статистическая информация относительно ассоциативной сочетаемости лингвистических образов, что обеспечивает возможность унифицированной оценки единицы и количества смысла ЕЯК. Предложен метод измерения смысла ЕЯК на основе нечеткого отношения смысла, обеспечивающий использование информации про связи между леммами текста, что позволяет получить оценку двух типов неопределенности измерения формальных признаков смысла. Получены и интерпретированы формальные оценки неопределенности результатов измерения смысла ЕЯК, что позволяет учитывать информацию про связи между леммами для решения прикладных задач компьютерной лингвистики.

С помощью разработанного на основе пакета DKPro Core программного обеспечения проведены эксперименты с целью исследования предложенного метода в задаче определения информативных признаков текста. В результате проведенных экспериментов полу-

чены зависимости параметров обнаруженного Парето-подобного закона распределения связей между леммами, анализ которых позволяет считать показатель среднего количества связей лингвистического образа наиболее информативным численным признаком текста.

Ключевые слова: смысл, неопределенность, текст, естественно-языковая конструкция, искусственная когнитивная система, лингвистический образ, лемма.

Bisikalo O. V.[1], Vasilevskyi O. M.[2]

[1]Dr.Sc., Professor, Dean of faculty for computer systems and automation, Vinnytsia National Technical University, Vinnytsia, Ukraine

[2]PhD, Associate professor, Professor of department of metrology and industrial automatics, Vinnytsia National Technical University, Vinnytsia, Ukraine

EVALUATION OF UNCERTAINTY MEASURING OF SENSE OF THE NATURAL LANGUAGE CONSTRUCTS

The task of evaluation of measurement uncertainty meaning of natural language constructs (NLC) based on formalization of the concepts of linguistic image, artificial cognitive systems and unit of sense is resolved. The basis of model the knowledge base of artificial cognitive system laid down statistical information regarding the associative compatibility of linguistic images, which enables unified evaluation the unit and the quantity of sense NLC. The method for measuring the sense of NLC based on fuzzy relation of meaning is offered. It provides to use information about the links between lemmas of text that allows you to estimate the measurement uncertainty of two types of sense signs. The results of the formal evaluation of the uncertainty of measurement sense of NLC are received and interpreted what enables into account information about the relationship between lemmas for solve tasks of computational linguistics.

With developed on the basis of the package DKPro Core Software conducted experiments to study the proposed method in the problem of the definition of informative features of the text. The experiments obtained dependence of the parameters detected Pareto-like distribution law relations between lemmas, whose analysis suggests that average number of connections of linguistic image is the most informative numerical feature for the text.

Keywords: sense, uncertainty, text, natural language construct, artificial cognitive system, linguistic image, lemma.

REFERENCES

1. Bisikalo O.V. Sub'ektivnaya edinitsa smyisla obraznyih konstruktsiy, Nauka: teoria i praktyka – 2009: materialy V miedzynar. naukowi-praktycznej konf., (Przemysl, 7–15 sierpnia 2009). Przemysl, Nauka i studia, 2009, Vol. 6, pp. 9–12
2. Bisikalo O. V. Infologichniy pidhid do modelyuvannya obraznogo mislennya lyudini [Elektronniy resurs], Visnik SumDU (Seriya "Tehnichni nauki"), 2009, No. 2, pp. 15–20. Rezhim dostupu: http://visnyk.sumdu.edu.ua/arhiv/2009/Tech_2_09/09bovoml.pdf.
3. Vasilevskyi O. M. Calibration method to assess the accuracy of measurement devices using the theory of uncertainty, International Journal of Metrology and Quality Engineering, 2014, 5.04: 403, No. 3 (7), 2006, pp. 147–151.
4. Rukovodstvo po vyirazheniyu neopredelennostey izmereniya = Guide to the Expression of Uncertainly in Measurement : [nauchn. redaktor Slaev V. A.]. Sankt-Peterburg, NPO VNIIM im. D. M. Mendeleeva, 1999, 134 p.
5. Vasilevskiy O. M. Algoritm otsinyuvannya neviznachenosti u vimiryuvannyah pri vikonanni metrologichnih robit, Informatsiyni tehnologiyi ta komp'yuterna inzheneriya, 2006, No. 3 (7), pp. 147–151.
6. Primenenie «Rukovodstva po vyirazheniyu neopredelennosti izmereniy», MI 2552-99. Ofits. izd. Sankt-Peterburg, VNIIM im. D. I. Mendeleeva, 1999, 27 p.
7. Kolmogorov A. N. Osnovnyie ponyatiya teorii veroyatnostey 2-e izd. Moscow, Nauka, 1974, 120 p.
8. Bisikalo O.V. Formalizatsiya ponyat movnogo obrazu ta obraznogo sensu prirodno-movnih konstruktsiy, Matematichni mashini i sistemi, 2012, No. 2, pp. 70–73.
9. Bisikalo O.V. Formalni metodi obraznogo analizu ta sintezu prirodno-movnih konstruktsiy : monografiya. Vinnitsya, VNTU, 2013, 316 p.
10. Raskin L. G., Seraya O. V. Nechetkaya matematika. Osnovyi teorii. Prilozheniya. Har'kov, Parus, 2008, 352 p.
11. Zagalni vimogi do kompetentnosti viprobuvalnih ta kalibruvalnih laboratoriy : DSTU ISO/IEC 17025-2001. [Chinniy vid 2001 – 01 – 01]. Kiev, Derzhstandart UkraYini, 2001, 31 p. (Natsionalniy standart UkraYini).
12. Gurevych I., Muhlhauser M., Muller Ch., Steimle J., Weimer M., Zesch T. Darmstadt Knowledge Processing Repository Based on UIMA [Electronic resource]. February 9, 2007, Available at: \www/ URL: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf.
13. Bisikalo O. V., Gurevich I. Metod viluchennya obraznih znan z anglomovnogo tekstu na osnovi instrumentalnih zasobiv paketu DKPro Core, Kontrol i upravlinnya v skladnih sistemah: XII mizhnar. konf., 14–16 zhovtnya 2014 r.: tezi dopovidey. Vinnitsya, 2014, pp. 51.
14. Stanford parser [Elektronniy resurs]. Stanford Dependencies. Nazva z ekranu. Rezhim dostupu: http://nlp.stanford.edu/software/stanford-dependencies.shtml.
15. Free ebooks – Project Gutenberg [Elektronniy resurs]. Project Gutenberg Literary Archive Foundation. Rezhim dostupu: https://www.gutenberg.org/.
16. Bisikalo O. V. Statistichniy analiz skladnih zalezhnostey u teksti, Visnik Nats. un-tu «Lvivska politehnika», Informatsiyni sistemi ta merezhi, 2015, No. 814, pp. 228–236.