ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 004.94: 159.95

О. В. Бісікало[1]
І. В. Богач[1]

СКЛАДНІСТЬ КЛАСУ СЕМАНТИКО-ЗАЛЕЖНИХ ЗАДАЧ ОБРОБКИ ТЕКСТУ

[1]Вінницький національний технічний університет

Розглянуто формальні ознаки класу семантико-залежних задач обробки тексту, обґрунтовано його NP-повну процедурну складність. На основі аналогії між задачами про рюкзак та автоматичного реферування тексту показано доцільність використання формальних лінгвістичних знань з огляду на зменшення процедурної складності. Запропоновано універсальний підхід до обробки тексту з урахуванням зв'язків між сутностями, отримано його інформаційну оцінку та визначено шляхи удосконалення.

Ключові слова: процедурна складність, *NP*-повнота, семантико-залежні задачі, обробка тексту, інформаційна оцінка.

Вступ

Серед усіх задач комп'ютерної лінгвістики найбільшою складністю відрізняються такі, що пов'язані з семантичним аналізом та синтезом природно-мовної інформації. З огляду на деякі спільні формальні ознаки, на думку авторів, такі задачі варто виокремити в клас семантико-залежних задач обробки тексту, актуальними прикладами яких є анотування та реферування тексту, пошук ключових слів, підтримка діалогу тощо. Підвищення інтересу до такого класу задач та збільшення кількості відповідних наукових досліджень викликано зростаючим попитом на лінгвістичні Інтернет-технології практично в усіх країнах світу.

Розглянемо приклад семантико-залежних задач анотування та реферування тексту, достатня якість розв'язання яких спроможна вивільнити значний обсяг часу експертів. На відміну від анотації, яка є коротким викладом вмісту документа із загальним уявленням про його тему, реферат — це зв'язаний текст, що коротко виражає не лише центральну тему документа, але і мету, методи, основні результати викладеного дослідження чи розробки. Якщо мета анотації здебільшого полягає у привернення уваги читача до відповідного тексту, то на основі реферату, який містить всього лише 10—23 % тексту, користувачі можуть зробити висновки про текст так само точно, як із нього самого, але витративши на це вдвічі менше часу [1].

Основні складнощі розв'язання задач розглянутого класу виникають завдяки багатозначності слів природної мови — головній проблемі комп'ютерної лінгвістики. Важливо отримати оцінки складності різних підходів до розв'язання семантико-залежних задач, починаючи від прямого перебору та закінчуючи застосуванням таких евристичних процедур, які дають змогу людині швидко та ефективно розуміти смисл нової для неї текстової інформації. Це дозволить визначити доцільність та ефективність додаткових процедур лінгвістичного аналізу тексту, вилучення так званих стоп-слів, залучення спеціальних експертних знань тощо. Тому актуальною задачею є визначення формальних ознак, зокрема інформаційних, та загальна оцінка складності класу семантико-залежних задач обробки тексту.

Мета дослідження полягає в отриманні оцінки процедурної складності класу семантико-залежних задач та визначенні перспективних підходів до їх розв'язання.

1. Аналіз предметної області дослідження

З метою формального виокремлення класу семантико-залежних задач передусім розглянемо загальне поняття класів складності. В теорії алгоритмів класами складності називаються множини обчислювальних задач, приблизно однакових за складністю обчислення. Інакше кажучи, класи складності — це множини предикатів (функцій, що отримують на вхід слово і повертають резуль-

тат 0 або 1), що використовують для обчислення приблизно однакові кількості ресурсів [2].

Для кожного класу існує категорія задач, які є «найскладнішими». Це означає, що будь-яка задача з класу зводиться до такої задачі, причому сама задача лежить в класі. Такі задачі називають повними задачами для даного класу. Найвідомішими є *NP*-повні задачі.

Зазвичай клас складності визначається через множину предикатів, що володіють деякими властивостями. Типове визначення класу складності виглядає так: класом складності X називається множина предикатів $P(x)$, що обчислюються на машинах Тьюринга і використовують для цього обчислення $O(f(n))$ ресурсу, де n — довжина слова x.

Як ресурси найчастіше обирають час обчислення (кількість робочих тактів машини Тьюринга) або робочу зону. Мови, що розпізнаються предикатами з деякого класу (тобто множини слів, на які предикат повертає 1), також називаються тими, що належать тому ж класу.

До задач класу P (з англ. *Polynomial*) відносять множину задач, для якої існують відносно «швидкі» алгоритми розв'язання. Час розв'язання задач класу P поліноміально залежить від розміру вхідних даних. Клас P включений у ширші класи складності алгоритмів.

Прикладами задач із класу P є цілочислове додавання, ділення, множення матриць, визначення зв'язності графів, сортування множини з n чисел.

Задачі класу *NP* (з англ. *Non-deterministic polynomial*) — множина задач розпізнавання, розв'язок яких за наявності деяких додаткових відомостей (сертифікати рішень) можна так само «швидко» перевірити на машині Тьюринга. Екавівалентно клас *NP* можна визначити як клас, що містить задачі, які допускають поліноміальний час розв'язання на недетермінованій машині Тьюринга. Серед прикладів задач, про які на сьогоднішній день невідомо, чи належать вони *P*, але відомо, що вони належать *NP* наведемо:

— задачу виконання булевых формул — дізнатись по певній булевій формулі, чи існує для неї набір вхідних змінних, що повертає 1. Сертифікат — такий набір.

— задачу про повні підграфи — за даними графу дізнатись, чи є в ньому повні підграфи заданого розміру. Сертифікат — номери вершин, що утворюють повний підграф.

— визначення наявності в графі Гамільтонового циклу. Сертифікат — послідовність вершин, що утворюють Гамільтонів цикл.

Серед всіх задач класу NP можна виділити «найскладніші» — *NP*-повні задачі. Якщо вдасться розв'язати будь-яку з них за поліноміальний час, то усі задачі класу *NP* також можна буде розв'язати за поліноміальний час. Прикладами *NP*-повних задач є задача комівояжера, проблема Штейнера, задача про незалежну множину, ігри Сапер та Тетріс, задача про рюкзак тощо. На сьогодні всі ці задачі потребують експоненційних алгоритмів розв'язку.

Для оцінки складності класу семантико-залежних задач обробки тексту, що пропонується ввести, потрібно врахувати суттєві специфічні ознаки результатів розуміння текстової інформації. Для цього розглянемо проблему багатозначності слів природної мови з формальної точки зору словарних значень. В тлумачних словниках зазвичай подаються усі можливі словарні значення кожної словоформи з відповідним лексемним знаком, що об'єднує певну множину слів. Однакове написання слів, які належать до різних словоформ якраз і є тією причиною, що різко збільшує обсяг можливого перебору під час визначення потрібного значення (полісемічного) слова у кожному реченні тексту. Формально для r_i лексемних знаків у i-му реченні обраного тексту загальний обсяг перебору дорівнює всім можливим варіантам значень $(k)^{r_i}$, з яких лише один правильний з точки зору автора (k — середній коефіцієнт полісемії відповідної мови).

Внаслідок лінгвістичних досліджень підтверджено гіпотезу: чим вищий ступінь аналітичності мови, тим частіше той самий лексемний знак виконує різні функції, тим більший середній коефіцієнт полісемії. Наприклад, іспанська мова більш аналітична за німецьку, її коефіцієнт полісемії складає 6,9 значень на одну лексему, а для німецької мови коефіцієнт полісемії — 5,6 значень на одну лексему [3]. Для більш синтетичних слов'янських мов середні коефіцієнти полісемії суттєво різняться для різних частин мови, наприклад, для іменників — 4,32 значення на лексему, для прийменників — 5 для конкретних та 3,5 для абстрактних, а середній коефіцієнт полісемії для російської мови складає 3,1 значення на одну лексему [4]. Отже, можна вважати, що нижня межа загального обсягу перебору V для деякого тексту не менша

$$V \geq \sum_{i=1}^{m} 3^{r_i}, \qquad (1)$$

де m — кількість речень у цьому тексті.

Окрім ступеня аналітичності мови на середній коефіцієнт полісемії можуть впливати характер та предметна область тексту — зменшують коефіцієнт термінологічна сталість певної предметної області та строгий (науковий) стиль викладення матеріалу, а збільшують — застосування займенників, метафор, елементів так званої «Езопової мови» тощо. Але, у будь-якому випадку, зрозуміло, що задача розуміння тексту формально відповідає *NP*-повній складності завдяки ступеневому характеру функції (1). При цьому для людини розуміння добре відомої мови, у т. ч. незнайомого тексту на цій мові не викликає помітних труднощів, що свідчить про наявність природних механізмів ефективного вибору найімовірніших комбінацій значень всіх лексем речення у противагу повному перебору можливих значень.

Врахуємо також відомі підходи до семантичного аналізу текстової інформації, що розрізняють поняття лексичних функцій та семантичних відношень. З точки зору семантики окремого речення лінгвістами виявлено 40...60 (в залежності від мови) лексичних функцій, які пов'язують, як правило, окремі пари слів або словосполучення. Точно розрізняти всі можливі випадки означає, як мінімум, складність за кількістю сполук з r_i по 2 з коефіцієнтом 40, тобто $V' \geq 40 \cdot \sum_{i=1}^{m} \frac{r_i!}{2! \cdot (r_i - 2)!}$. Наступним кроком формального узагальнення змісту речення є поняття семантичного відношення (схеми), наприклад, у [5] обґрунтовується агрегація 21 відношення у 6 типів, що задаються 9-ма три- чотиримісними предикатами. Складність такого підходу пропорційна вже кількості розміщень з r_i по 3 з коефіцієнтом 9, а саме $V'' \geq 9 \cdot \sum_{i=1}^{m} \frac{r_i!}{(r_i - 3)!}$. Але, потрібно зважити на те, що значна, якщо не більша частина людства ніколи не чула або не переймалася існуванням лексичних функцій та семантичних відношень, що зовсім не заважало всім цим людям добре розуміти власну мову.

Отже, доцільність виокремлення класу семантико-залежних задач полягає в наступному — з одного боку, він характеризується *NP*-повною складністю, оскільки $V'' \geq V' \geq V$, але, з іншого боку, об'єктивно існують природні алгоритми мислення людини, що дозволяють ефективно розв'язувати задачі цього класу.

2. Автоматичне реферування як приклад семантико-залежної задачі обробки тексту

Попередній аналіз дає підстави стверджувати, що окремі задачі з класу семантико-залежних не тільки відповідають знаним *NP*-повним за процедурною складністю, але й подібні до них навіть за постановкою. Для підтвердження такої тези розглянемо вже згадану задачу про рюкзак (англ. *Knapsack problem*) як зручну аналогію для порівняння та оцінки процедурної складності задачі автоматичного реферування тексту [6]. В загальному вигляді задачу можна сформулювати так: із заданої множини предметів з властивостями «цінність» і «вага» маємо відібрати деяке число предметів таким чином, щоб отримати максимальну сумарну цінність з одночасним дотриманням обмеження на сумарну вагу.

Без урахування додаткової інформації аналогами параметрів «цінність» і «вага» в задачі про рюкзак очевидно є параметри «важливість» і «розмір» фрагментів тексту в задачі автоматичного реферування. Тобто, в загальному випадку результатом реферування має бути мінімальний обсяг тексту за наявності у ньому найважливіших фраз (речень), причому текст має зберегти свою сутність — остання додаткова вимога ще більше ускладнює задачу автоматичного реферування. За допомогою отриманої аналогії можна припустити, що задача автоматичного реферування тексту відноситься до задач *NP*-повного класу.

Відомо, що віднесення певної обчислювальної задачі до класу *NP*-повних зосереджує дослідників на знаходженні наближених алгоритмів її розв'язку [7], оскільки відсутність поліноміальних рішень зводить до нуля практичну цінність роботи. Розглянемо ситуацію з відомими розв'язками згаданих та використаних у аналогії задач. Для задачі комбінаторної оптимізації пакування рюкзаку маємо класичну картину — незадовільний час розв'язання точними методами повного перебору або (за рахунок збільшення необхідної пам'яті) динамічного програмування чи гілок і меж призводить до зосередження уваги на отриманні наближених результатів жадібним алгоритмом, генетичними алгоритмами або іншими методами дискретної оптимізації. На відміну від свого аналогу підходи до розв'язання задачі автоматичного реферування тексту історично будувалися як наближені методи [8], що враховували, виходячи зі специфіки задачі, додаткову інформацію лінгвістич-

ного характеру — як приклад можна навести метод карт текстових відношень (з англ. *TRM — Text Relation Map*).

Класичний метод *TRM* враховує зважені вектори слів, що відповідають фрагментам (реченням) обраного документа, при цьому використовується граф у якості формальної моделі семантичних відношень між структурними одиницями тексту. Вершинами графу є текстові фрагменти, ребра з'єднують вершини з високої мірою подібності (семантичного зв'язку). Визначення ключових текстових фрагментів (вершин графу) для формування реферату відбувається на основі критерію кількості семантичних зв'язків певного фрагменту з іншими (ребер, які виходять з вершини графу). В різновидах методу пропонується комбінувати метод *TRM* зі статистичними методами *TFIDF* та *TLTF* з метою додаткового визначення ваги окремих слів документу [9].

Проведемо оцінку процедурної складності традиційного методу *TRM*. Нехай n — кількість слів тексту, а m — кількість фрагментів (наприклад, речень). Не втрачаючи загальності міркувань, вважатимемо, що кількість слів у кожному реченні однакова і дорівнює $n' = n/m \approx r_i$. Тоді одна операція знаходження скалярного добутку двох векторів розмірністю n' (для 2-х речень) вимагає обчислень

$$k_1 = 2\frac{n}{m} - 1 . \tag{2}$$

Оскільки загальна кількість фрагментів *m*, то кількість операцій скалярного добутку їх векторів дорівнює

$$k_2 = \sum_{j=1}^{m} j . \tag{3}$$

Сума членів арифметичної прогресії (3)

$$k_2 = \sum_{j=1}^{m} j = \frac{m(m+1)}{2} . \tag{4}$$

Виходячи з (2) і (4), загальна кількість обчислень для визначення мір семантичної подібності текстових фрагментів за методом *TRM* дорівнює

$$K_2 = k_1 k_2 = \left(2\frac{n}{m} - 1\right)\frac{m(m+1)}{2} . \tag{5}$$

Зрештою маємо [10], що для методу TRM оцінка процедурної складності обмежується $O(nm)$ — це не перевищує складності полінома 2-го порядку для кількості слів тексту n та свідчить про дієвість застосування процедур його лінгвістичного аналізу. Проте потрібно визнати, що найкращі результати автоматичного реферування все ще значно поступаються авторським або експертним варіантам.

Отже, для типової семантико-залежної задачі автоматичного реферування ефективним є а) врахування суттєвих лінгвістичних ознак та параметрів окремих слів, тексту в цілому або колекції текстів; б) визначення найінформативніших метрик для оцінки якості реферату з огляду на специфіку тексту.

3. Інформаційна оцінка підходу до обробки тексту з урахуванням зв'язків між сутностями

Альтернативним шляхом оцінки складності класу семантико-залежних задач можна вважати аналіз інформаційних потоків, необхідних для забезпечення розв'язку таких задач. На відміну від процедурної складності, що визначається як загальна оцінка без конкретизації та врахування особливостей методу (процедури) розв'язку, інформаційна оцінка за визначенням є процедурно-орієнтованою. З огляду на це пропонується розглянути інформаційну оцінку універсального підходу до обробки тексту з урахуванням зв'язків між його змістовними сутностями (лексемами).

Виходитимемо з того, що ключова ознака семантико-залежних задач полягає у визначенні та обробленні множини змістовних сутностей тексту. З інформаційної точки зору розуміння сенсу речення окремим суб'єктом супроводжується розпізнаванням окремих слів, з яких воно складається та зв'язків між парами цих слів з відповідною побудовою дерева таких зв'язків [11]. Вважатимемо, що всі ці процеси відбуваються шляхом порівняльного аналізу та залучення інформації з деякої загальнолінгвістичної бази знань суб'єкта розуміння. Якщо кожен з цих етапів супроводжується збільшенням інформації, то гіпотетично для такого універсального підходу:

— рівень загального розуміння тексту *T* може змінюватися від мінімально можливого до максима-

льного в залежності від обсягу та інших параметрів загальнолінгвістичної бази знань суб'єкта;

— якість визначення змістовних сутностей пропорційна рівню загального розуміння тексту, що має підтверджуватися формальними ознаками.

Надамо інформаційну оцінку цим процесам, а саме:

1. Визначимо обсяг ентропії одного слова тексту для випадку, коли поява цього слова є незалежною випадковою подією x з l можливими станами як

$$H(x) = -\sum_{j=1}^{l} p\left(x_j\right) \cdot \log p\left(x_j\right),$$

а максимальну усереднену оцінку отримуємо для рівноймовірного випадку

$$H_w = \log_2 l \ \ [\text{Біт}].$$

Змінною l можна вважати кількість різних слів (лексем) тексту T, очевидно, що $l \le n$.

2. Визначимо також максимальну оцінку обсягу ентропії всіх слів речення за умови, що поява наступного слова з n' слів цього речення не залежить від попереднього

$$H(x) = -n' \cdot \sum_{j=1}^{l} p\left(x_j\right) \cdot \log p\left(x_j\right)$$

або для рівноймовірного випадку

$$H_{sw} = n' \cdot \log_2 l \ \ [\text{Біт}]. \tag{6}$$

3. Тепер визначимо обсяг ентропії одного парного зв'язку за умови, що слова речення у вигляді деякої множини $X = \{x_1, ..., x_{n'}\}$ суб'єктом вже розпізнані та відомі. Для незалежної появи пари як випадкової події y потенційна кількість таких пар може бути $n' \times (n' - 1) = (n')^2 - n'$, оскільки речення з n' слів складає дерево синтаксичного розбору з n' пар, враховуючи двосторонній зв'язок підмет-присудок. З іншого боку, головна діагональ такої матриці зв'язків виключається, оскільки слово у реченні не може бути пов'язане само з собою. Отже

$$H(y) = -\sum_{i=1}^{(n')^2 - n'} p\left(y_i \mid X\right) \cdot \log p\left(y_i \mid X\right),$$

відповідно для рівноймовірного випадку

$$H_p \approx \log_2 (n')^2 = 2 \cdot \log_2 n' \ \ [\text{Біт}].$$

4. Наостанок визначимо обсяг ентропії всіх пар одного окремого речення, наприклад, за комбінаційними характеристиками побудови дерева з n' пар, які обрані з $n' \times (n' - 1)$ можливих. За найжорсткішої умови незалежного єднання слів речення у n' пар маємо

$$H(y) = -n' \cdot \sum_{i=1}^{(n')^2 - n'} p\left(y_i \mid X\right) \cdot \log p\left(y_i \mid X\right),$$

а для рівноймовірного випадку

$$H_{sp} \approx n' \cdot \log_2 (n')^2 = 2n' \cdot \log_2 n' \ \ [\text{Біт}]. \tag{7}$$

Внаслідок збільшення базового обсягу ентропії слів речення (6) додатковою ентропією його пар (7) отримуємо максимальну загальну ентропію одного речення тексту

$$H_{sent} = H_{sw} + H_{sp} = n'\left(\log_2 l + 2 \cdot \log_2 n'\right) \ \ [\text{Біт}]. \tag{8}$$

Отже, застосування запропонованого універсального підходу до обробки тексту з m речень збільшує загальнолінгвістичну базу знань суб'єкта на $mn'\left(\log_2 l + 2 \cdot \log_2 n'\right)$ [Біт].

Аналіз виразу (8) показує, що за умови відносно невеликих коливань кількості значимих слів n' у реченні, ключовим параметром загальнолінгвістичної бази знань суб'єкта залишається l — кількість розпізнаних словоформ (лексем) тексту. Другим важливим висновком є прийнятна межа інформаційної оцінки $O(mn'l)$ складності універсального підходу, яка співмірна процедурній складності методу TRM для типової семантико-залежної задачі автоматичного реферування.

Для зменшення процедурної складності запропонованого підходу, що може бути перспектив-

ним для розв'язання цілої низки семантико-залежних задач [11], варто розглянути частотні характеристики словарного складу природної мови. Оскільки основну змістовну інформацію несуть значимі словоформи (лексеми) речення, то безпосереднє вилучення в процесі пасерування тексту так званих стоп-слів, а також визначення кореференцій для займенників суттєво зменшує значення l. Так, для синтетичної російської мови згідно з [12, 13] питома вага у корпусах таких частин мови, як вступні слова, займенники (для іменників, прикметників та прислівників), прийменники, сполучники, частки досягає 38,1 %. Для аналітичних мов, зокрема англійської ситуація інша, а саме, згідно з [14], питома вагома іменників, дієслів, прикметників та прислівників з найвживаніших слів досягає 96,4 %. З іншого боку, вилучення збиткових зв'язків для відносно великої кількості артиклів та прийменників англійського речення дозволить помітно зменшити значення n' — всі ці процеси забезпечуються сучасними парсерами.

Висновки

Обґрунтовано доцільність виокремлення класу семантико-залежних задач, який характеризується *NP*-повною складністю з одночасною наявністю природних алгоритмів мислення людини, що дозволяють ефективно розв'язувати задачі цього класу. На прикладі аналогії між задачами про рюкзак та автоматичного реферування тексту показано, що використання знань про мову у тій чи іншій формі традиційно закладається в алгоритми обробки тексту і дозволяє зменшити процедурну складність до поліноміальної.

На основі отриманої в роботі інформаційної оцінки універсального підходу до обробки тексту з урахуванням зв'язків між сутностями (словами, лексемами) визначено максимальну загальну ентропію одного речення тексту. Оскільки складність запропонованого підходу також є поліноміальною, а технологічні можливості сучасних парсерів забезпечують відповідні процедури лінгвістичного аналізу тексту, обробка тексту з m речень за таким підходом практично може збільшити загальнолінгвістичну базу знань суб'єкта аналізу на $mn'\left(\log_2 l + 2\cdot\log_2 n'\right)$ [Біт].

Перспективним напрямом розвитку задач дослідження є визначення впливу накопичення знань про зв'язки між змістовними сутностями тексту на точність оцінки ймовірностей $p\left(y_i \mid X\right)$.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРА

1. Hahn U. The Challenges of Automatic Summarization / U. Hahn, I. Mani // IEEE Computer Cociety. — 2000. — Vol. 33, № 11. — P. 29—36.

2. Cormen T. H. Introduction to Algorithms (2nd ed.). MIT Press and McGraw-Hill / T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein. — Chapter 34: NP–Completeness. — 2001. — P. 966—1021.

3. Яковлева Т. А. Сопоставительное исследование субстантивной полисемии : на материале немецкого и испанского языков : автореф. ... канд. филол. наук : спец. 10.02.20 Сравнительно-историческое, типологическое и сопоставительное языкознание / Татьяна Анатольевна Яковлева, 2001.

4. Ніколаєва Л. Б. Явище полісемії у номінативних терміносистемах / Л. Б. Ніколаєва // Культура народов Причерноморья. — 2007. — № 110, Т. 2. — С. 65—67. — Режим доступу до журн. : http://dspace.nbuv.gov.ua/bitstream/handle/123456789/55144/24-Nikolaieva.pdf?sequence=1 . — Назва з екрана.

5. Найханова Л. В.Основные типы семантических отношений между терминами предметной области [Електронний ресурс] / Л. В. Найханова // Известия высших учебных заведений. Поволжский регион. Технические науки. — 2008. — вып. 1. — Режим доступу : http://cyberleninka.ru/article/n/osnovnye-tipy-semanticheskih-otnosheniy-mezhdu-terminami-predmetnoy-oblasti — Название с экрана.

6. Riedhammer K. «Packing the Meeting Summarization Knapsack» // Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH) / K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tur, 2008. — Pp. 2434—2437.

7. Richard M. Karp. Reducibility among combinatorial problems [Electronic resource] / Richard M. Karp. — Access mode : http://www.cs.berkeley.edu/~luca/cs172/karp.pdf .

8. Шаховська Н. Б. Автоматизована система укладання реферату / Н. Б. Шаховська, З. В. Стахів // Вісник Національного університету «Львівська політехніка». — 2012. — № 743— С. 210—218. — (Інформаційні системи та мережі).

9. Каніщева О. В. Використання карт відношень (TRM) для автоматичного реферування / О. В. Каніщева // Вісник Національного університету «Львівська політехніка». — 2013. — № 770. —С. 108—122. — (Інформаційні системи та мережі).

10. Бісікало О. В. Автоматичне анотування текстів на основі мовних образів / О. В. Бісікало, І. О. Назаров // Кібернетичне управління та інформаційні технології. — 2014. — № 1. — С. 46—51.

11. Бісікало О. В. Визначення змістовних ознак тексту на основі аналізу зв'язків між лексичними одиницями / О. В. Бісікало, А. І. Лісовенко, О. В. Яхимович, С. С. Траченко // Вісник НТУ «ХПІ». — 2015. — № 21 (1130). — С. 83—89. — (Механіко-технологічні системи та комплекси). — ISSN 2411-2798. — Бібліогр. : 10 назв.

12. Статистика корпуса [Электронный ресурс] // Национальный корпус русского языка. — Режим доступа : http://www.ruscorpora.ru/corpora-stat.html . — Название с экрана.

13. Ляшевская О. Л. Новый частотный словарь русской лексики : Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) [Электронный ресурс] / О. Л. Ляшевская, С. А. Шаров // Слова-

ри на основе национального корпуса русского языка. — М. : Азбуковник, 2009. — Режим доступа : http://dict.ruslang.ru/freq.php?act=show&dic=freq pos&title=%C4%E0%ED%ED%FB%E5%20%EE%20%F7%E0%F1%F2%EE%F2%ED%EE%F1%F2%E8%20%F7%E0%F1%F2%E5%F0%E5%F7%ED%FB%F5%20%EA%EB%E0%F1%F1%EE%E2%20%28%ED%E0%20%EC%E0%F2%E5%F0%E8%E0%EB%E5%20%EF%EE%E4%EA%EE%F0%EF%F3%F1%E0%20%F1%EE%20%F1%ED%FF%F2%EE%E9%20%E3%F0%E0%EC%EC%E0%F2%E8%F7%E5%F1%EA%EE%E9%20%EE%EC%EE%ED%E8%EC%E8%E5%E9%29 . — Название с экрана.

14. Adam Kilgarriff. BNC database and word frequency lists [Electronic resource] / Adam Kilgarriff. — Access mode: http://www.kilgarriff.co.uk/bnc-readme.html .

Рекомендована кафедрою автоматики та інформаційно-вимірювальної техніки ВНТУ

Стаття надійшла до редакції 27.01.2016

Бісікало Олег Володимирович — д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики, e-mail: obisikalo@gmail.com;

Богач Ілона Віталіївна — канд. техн. наук, доцент кафедри автоматики та інформаційно-вимірювальної техніки, e-mail: ilona.bogach@gmail.com .

Вінницький національний технічний університет, Вінниця

O. V. Bisikalo[1]
I. V. Bogach[1]

Complexity Class Semantic-Dependent Word Processing Tasks

[1]Vinnytsia National Technical University

Consider the formal signs of class-dependent semantic word processing tasks, it proved of Np-complete procedural complexity. On the basis of the analogy between the problems of the backpack and automatic summarization method shows the feasibility of using formal language skills, taking into account the reduction of procedural complexities. A universal approach to the treatment of the text, taking into account the links between entities, obtained its information and assessment of the ways to improve.

Keywords: summarization of text; calculation complexity, NP-completeness, TRM method.

Bisikalo Oleh V. — Dr. Sc. (Eng.), Professor, Dean of the Department of Computer Systems and Automation, e-mail: obisikalo@gmail.com;

Bogach Ilona V. — Cand. Sc. (Eng.), Assistant Professor of the Chair of Computer Systems and Automation, e-mail: ilona.bogach@gmail.com

О. В. Бисикало[1]
И. В. Богач[1]

Сложность класса семантико-зависимых задач обработки текста

[1]Винницкий национальный технический университет

Рассмотрены формальные признаки класса семантико-зависимых задач обработки текста, обоснована его Np-полная процедурная сложность. На основе аналогии между задачами о рюкзаке и автоматического реферирования текста показана целесообразность использования формальных лингвистических знаний, учитывая уменьшение процедурной сложности. Предложен универсальный подход к обработке текста с учетом связей между сущностями, получена его информационная оценка и определены пути усовершенствования.

Ключевые слова: процедурная сложность, Np-полнота, семантико-зависимые задачи, обработка текста, информационная оценка.

Бисикало Олег Владимирович — д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики, e-mail: obisikalo@gmail.com;

Богач Илона Витальевна — канд. техн. наук, доцент кафедры автоматики и информационно-измерительной техники, e-mail: ilona.bogach@gmail.com