УДК 004.93:159.95

О. В. Бісікало

ОЦІНКА СКЛАДНОСТІ КЛАСУ СЕМАНТИКО-ЗАЛЕЖНИХ ЗАДАЧ ОБРОБЛЕННЯ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Розглянуто формальні ознаки класу семантико-залежних задач обробки тексту, обґрунтовано його NP-повну процедурну складність. Показано, що природні алгоритми мислення людини дозволяють ефективно розв'язувати задачі цього класу.

Ключові слова: процедурна складність, NP-повнота, семантико-залежні задачі, оброблення тексту.

Abstract

Formal features of class of semantic-dependent text processing tasks were considered; NP-complete procedural complexity was proved. It is shown that the natural human thinking algorithms use effectively to solve problems of this class.

Keywords: calculation complexity, NP-completeness, semantic-dependent tasks, text processing.

Вступ

Ряд складних задач комп'ютерної лінгвістики – анотування та реферування тексту, пошук ключових слів, підтримка діалогу тощо – з огляду на деякі спільні формальні ознаки варто виокремити в клас семантико-залежних задач обробки тексту [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 людини швидко та ефективно розуміти смисл нової для неї текстової інформації. Це дозволить визначити доцільність та ефективність додаткових процедур лінгвістичного аналізу тексту, наприклад, вилучення так званих стоп-слів, застосування спеціальних експертних знань тощо.

Мета дослідження полягає в отриманні оцінки процедурної складності класу семантико-залежних задач.

Результати дослідження

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

Для кожного класу існує категорія задач, які являються «найскладнішими». Це означає, що будь-яка задача з деякої множини зводиться до такої задачі, причому сама задача належить до цієї множини. Такі задачі називають повними задачами для даного класу. Найбільш відомими вважають *NP*-повні задачі, що входять до класу *NP* (з англ. *Non-deterministic polynomial*). Прикладами *NP*-повних задач є задача комівояжера, проблема Штейнера, задача про незалежну множину, ігри Сапер та Тетріс, задача про рюкзак тощо. На даний час всі ці задачі потребують експоненційних алгоритмів розв'язку.

Для оцінки складності класу семантико-залежних задач обробки тексту, що пропонується увести, 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

при визначенні 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 вірний з точки зору автора (k – середній коефіцієнт полісемії відповідної мови).

Внаслідок лінгвістичних досліджень було підтверджено гіпотезу: чим вищий ступінь аналітичності мови, тим частіше один и той же 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, менш аналітичної 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 за

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

де, 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 найбільш ймовірних 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 3(4)-х місними 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.

Висновки

У роботі обґрунтовано доцільність виокремлення класу семантико-залежних задач комп'ютерної лінгвістики. Отримано оцінки *NP*-повної процедурної складності таких задач та одночасно акцентовано увагу на наявність природних алгоритмів мислення людини, що дозволяють ефективно розв'язувати задачі цього класу. Отже, перспективним напрямком дослідження є формалізація притаманних людині природно-мовних та когнітивних обмежень, які дозволять зменшити процедурну складність розв'язків семантико-залежних задач до поліноміальної.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Bisikalo O. Solving problems on base of concepts formalization of language image and figurative meaning of the natural-language constructs [Електронний ресурс] / Oleg Bisikalo, Slawomir Cieszczyk, Gulbahar Yussupova // 16th Conference on Optical Fibers and Their Applications. International Society for Optics and Photonics. – December 18, 2015. – Pp. 98161U-98161U-14. – Режим доступу: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=2478661.

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

3. Сопоставительное исследование субстантивной полисемии : На материале немецкого и испанского языков / Тема диссертации и автореферата кандидата филологических наук. – Яковлева, Татьяна Анатольевна, 2001. – Специальность: Сравнительно-историческое, типологическое и сопоставительное языкознание (ВАК 10.02.20).

4. Явище полісемії у номінативних терміносистемах [Електронний ресурс] / Л.Б. Ніколаєва // Культура народов Причерноморья. – 2007. – № 110, Т. 2. – С. 65-67. – Режим доступу: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/55144/24-Nikolaieva.pdf?sequence=1.

5. Основные типы семантических отношений между терминами предметной области [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/osnovnye-tipy-semanticheskih-otnosheniy-mezhdu-terminami-predmetnoy-oblasti. – Назва з екрану.

Олег Володимирович Бісікало — доктор технічних наук, професор, декан факультету комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця.

Oleg V. Bisikalo — Doctor of Engineering, Professor, Dean of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: obisikalo@vntu.edu.ua.