УДК 004.82

Ю. І. Кисленко, Д. С. Сергеєв

ПОРІВНЯННЯ СПОСОБІВ ЗБЕРЕЖЕННЯ СЛІВ В ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

Анотація: У даній статті окреслено основні проблеми представлення природномовного слова у вигляді структурованого об’єкту даних. В рамках теоретичного обґрунтування виділено й описано особливості слова як об’єкту даних та визначено загальні вимоги до його загальної структури. На основі отриманих вимог побудовано ієрархічну класифікацію способів представлення слова в ІТ за рівнем деталізації описаного об’єкта. Розглянуто характерні особливості кожного з класів, проведено їх порівняльний аналіз, результатом якого є рекомендації щодо можливих сценаріїв використання для кожного з них.

Ключові слова: база даних, база знань, слово, природна мова, інформаційні природномовні технології.

Вступ

Робота з даними – один з найбільш популярних напрямів досліджень в сучасному ІТ-товаристві. До технологій цього напряму належать, зокрема, розробка та вдосконалення різноманітних баз даних та баз знань, розробки в галузі роботи з великими масивами даних (*Big Data*), автоматичний аналіз та пошук закономірностей у наборов даних (*Data Mining*) тощо [1]. Окремі підгрупи утворюють серед них технології для роботи зі спеціалізованими типами даних, самі алгоритми роботи яких прямо залежать від особливостей відповідних даних. Однією з найбільш помітних серед них є група інформаційних природно-мовних технологій (ІПМТ), об’єктом роботи яких є природно-мовна інформація (ПМІ).

Природно-мовна (текстова) інформація зустрічається у дуже широкому спектрі інформаційних технологій – від інтерфейсів користувача до створення мов програмування, від систем обміну миттєвими повідомленнями до електронних енциклопедій. Але не всі технології, в яких використовуються природномовні дані, можна однозначно ідентифікувати як ІПМТ. Лише частина з цих технологій працює з ПМІ як з окремим, повноцінним типом інформації – таким, що має унікальну структуру, що відрізняє його від як від інших типів інформації, так і від неструктурованих даних взагалі. У випадку ІПМТ такою характерною структурою є, власне, структура природної мови – яка, в свою чергу, багато в чому залежить від структури найменшої своєї значущої одиниці – структури слова.

Хоча слово є наріжним каменем природної мови – найменшим її елементом і водночас ланкою, що пов’язує природно-мовні конструкції з реальним світом, узгодженого розуміння щодо форми його представлення у реальних системах досі немає. Більш того, поняття слова взагалі виноситься за дужки й обмежується дослідженням його характеристик, що використовуються на більш високих рівнях організації мови (речення, текст тощо). Ціль даної статті – вивчення способів використання слова в сучасних ІТ та спроба їх класифікації як з точки зору їх внутрішніх особливостей, так і з точки зору можливостей їх практичного використання.

Перша частина статті присвячена аналізу особливостей слова як об’єкта даних та побудові вертикальної класифікації існуючих підходів до вирішення задачі збереження слів в інформаційних технологіях з урахуванням висвітлених особливостей.

Друга частина статті складається з порівняльного аналізу різних способів в рамках отриманої класифікації та висновків, заснованих на результатах цього аналізу – зокрема, виділення особливостей підходів різних класів та можливості використання цих підходів до різних типів практичних задач.

Слово як об’єкт даних

Перш за все, для формалізації та більш детального визначення об’єкту дослідження, розглянемо особливості слова як об’єкту обробки та збереження даних.

Слово, за своєю природою – це символьне представлення (ярлик) певного феномену реального світу. Таким чином, слова в контексті природної мови можемо розглядати як символи, а саму природну мову – як логічно структуровану систему, утворену з цих символів та правил їх взаємодії. Ця проста модель досить вірно представляє загальний вигляд організації природної мови – але, на жаль, досвід показує, що вона не є достатньо точною для практичного застосування. Виділимо основні особливості природної мови як структури, що формувалася протягом багатьох віків, які заважають її формалізувати та уніфікувати:

– по-перше, кожна мова світу унікальна: кожна мова має великий історичний багаж виключень та притаманних тільки їй одній особливостей – тобто, у різних мов може істотно відрізнятися сама їх структура;

– по-друге, навіть в рамках однієї мови граматичні правила часто бувають досить складними і, при цьому, часто ще й неоднозначними;

– по-третє, при урахуванні словотворення кількість нових слів прямо залежить від кількості слів у мові загалом, що робить фактично неможливою попередню оцінку їх кількості та їх зв’язків з іншими словами.

При цьому, вирішення багатьох практичних задач (наприклад автоматичний переклад, пошукова оптимізація, автоматичне наповнення баз знань тощо) потребує ефективних механізмів роботи не лише з суцільними текстами, але й з окремими словами і їх поєднаннями.

Розділимо усю сукупність способів представлення слова за ступенем деталізації слова як окремого об’єкту даних в рамках системи, від простого до складного.

Елементарні способи

Клас *елементарних способів* охоплює ті випадки, де слова на рівні системи виділяються у окремі сутності, відмінні від інших даних, але не мають власних унікальних особливостей. Іншими словами, будь-який рядок, складений з певного набору символів (наприклад, символів одного з природно-мовних алфавітів), є словом. При цьому смислове навантаження слова повністю втрачається.

Так, в електронній книзі або системі обміну текстовими повідомленнями може виникнути потреба підрахувати кількість слів, знайти слово в тексті тощо – але саме слово визначається виключно як певних сукупність символів у заданому порядку.

Яскравим прикладом використання такого способу є обробка тексту при архівації даних. На перший погляд, робота архіватора подібна до роботи більш високорівневого алгоритму, такого як стемінг: текст розбивається на невеликі фрагменти, які мають деякі спільні характеристики. Більш того, як і у високорівневих алгоритмах, під час архівації зі слів часто виділяється найбільша спільна частина. Але, незважаючи на зовнішню подібність, ці підходи кардинально відрізняються: при архівації смислове навантаження слів при цьому повністю втрачається. Так, текст, який було зашифрований з використанням шифрування підстановкою [2], буде архівований абсолютно ідентично до оригінального тексту, хоча сам по собі він не несе жодного змісту. Наприклад, цілком можливе виділення спільної частини «*мент*» зі слів «*мент*альність» та «аргу*мент*», які жодним чином не пов’язані за смислом.

Структурні способи

До *структурних способів* належать ті системи, де слово не лише є окремою сутністю, але також обмежене певними правилами, що відрізняє його від інших типів даних. Слово в рамках цього підходу все ще розглядається як сукупність символів і його смислове навантаження не враховується при обробці. Але, на відміну від елементарних способів, система розрахована на роботу саме з словами природної мови – тобто, правильна робота системи при заміні слів на згенеровані випадковим чином рядки або інші типи даних не очікується і, відповідно, не гарантується.

Одним з класичних структурних способів є алгоритм перевірки правопису в електронних текстових редакторах на основі методу найближчого сусіда за спільними символами: в системі задано певні правила правопису і певні правила подібності (для виправлення помилок вводу), що побудовані для роботи з даною мовою. Таким чином, усі слова даної мови обробляються достатньо точно; при цьому цілком можливо внести в базу абсолютно штучне слово (наприклад, ім’я власне або абревіатуру), яке буде оброблятись за тими самими правилами. Таким чином, наприклад, реалізоване збереження текстових даних у вигляді «словників» у базі даних *PostgreSQL* [3].

Мережеві способи

Мережеві способи – ті способи, де слово розглядається як повноцінний об’єкт в рамках системи. На відміну від інших способів, сама структура слова тут вже носить вторинний характер – адже слово визначається як сукупність різних словоформ і зв’язків між ними.

Системи цього класу майже не дають приросту швидкодії, але мають великий потенціал для вирішення аналітичних задач високого рівня абстракції, в тому числі в галузях когнітивної лінгвістики, психолінгвістики, штучного інтелекту. Це стає можливим завдяки надлишковості бази в мережевих способах (що дозволяє відстежувати зв’язки між окремими словоформами) та простоті накопичення семантичної інформації про словоформи – смислові зв’язки, контекст, джерело тексту тощо.

Найпростішим прикладом є загальний випадок системи пошуку в мережі Інтернет або іншому великому корпусі природно-мовних даних: при аналізі тексту на веб-сторінці необхідно ідентифікувати слова не тільки за структурою, але й за смисловим навантаженням, чого неможливо добитись, розглядаючи одне слово як окрему сутність. Так, у системі *WordNet* навіть різні смислові значення однієї й тієї ж словоформи представлені як окремі сутності: наприклад, сутність «*moving*» з ключем 01 має сенс «*зворушливий*», а «*moving*» з ключем 02 – «*той, що рухається*».

Більш того, мережеві способи дозволяють на полі словоформ і їх зв’язків визначити лексему – сукупність словоформ, які можна отримати з одного кореня за правилами словозміни. Це дозволяє одночасно розглядати будь-яку словоформу як частину лексеми (або відображення смислового навантаження відпо-

відного слова) та як окрему сутність, що пов’язана смисловими зв’язками з іншими словами. Зокрема, представлена таким чином лексема є потужним ядром для використання у задачах, тісно пов’язаних з семантикою – семантичних мережах, базах знань тощо.

Аналіз способів збереження слів

Оскільки в основі визначеної вище класифікації лежить розподіл способів збереження слів за рівнем абстракції (деталізації) представлення одного слова, будь-який спосіб можна за цією ознакою віднести до одного з представлених класів.

Єдиним випадком, коли точна класифікація може бути ускладнена – системи, де використовуються одночасно підходи різних рівнів. Як приклад можна навести збереження форм слова у вигляді стисненого тексту (елементарний спосіб) одночасно з об’єднанням посилань на них у спільний об’єкт (мережевий спосіб). За таких умов спосіб слід класифікувати за найвищим використаним рівнем, доступним для користувача – адже зміна алгоритму його роботи однозначно приведе до змін в роботі системи, в той час як використання інших низькорівневих способів вплине лише на кількісні її показники.

Слід також зазначити, що елементарні способи в контексті даної статті не становлять істотного інтересу для дослідження, оскільки при їх використанні слово втрачає властивості цілісного об’єкту даних – а отже, стає неможливим застосування до нього спеціалізованих груп алгоритмів і методів.

Враховуючи вищесказане, область дослідження звужується до порівняння структурних та мережевих способів. Окреслимо основні задачі, які вирішуються при роботі зі словами взагалі, і на основі цього визначимо спільні та відмінні вимоги до вибору інструментів для їх вирішення.

Першою з типових проблем, для вирішення яких використовується робота зі словами на рівні даних, є оптимізація роботи системи на низькому рівні. Оскільки слова можна представити як окремі сутності, що мають певні спільні характеристики (структуру, правила змін тощо), виділення яких дозволяє досягти кращої швидкодії системи та/або меншого споживання ресурсів у порівнянні зі стандартною схемою сховища даних «ключ–значення».

Другою класичною задачею є якісне розширення можливостей системи – точніше, використання нових способів представлення тих самих даних, які, в свою чергу, дозволяють використовувати більш високотехнологічні засоби обробки та аналізу даних (в першу чергу тут слід звернути увагу на семантичний

аналіз природно-мовних текстів). Спектр задач, які належать до цього напряму, досить широкий – сюди належать, серед іншого, і визначення смислових зв’язків між словами, і нормалізація слів, і пошук плагіату. Очевидно, що підходи, розроблені для вирішення спеціалізованої групи задач, більш ефективні, ніж загальні способи обробки даних [4].

Враховуючи вищезазначене, можемо тепер визначити спільні риси різних способів збереження слів, які дозволяють однозначно відокремити їх від методів обробки інших типів даних.

З одного боку, усі ці способи обробляють дані, в основі яких лежить слово – об’єкт, що має досить велику кількість елементів зі складною системою зв’язків між ними. Відповідно, комп’ютерна модель повинна мати чітко визначену структуру, яка б відтворювала модель слова хоча б на рівні інтерфейсів доступу (вхідні дані, методи тощо), в тому числі типові операції над даними та формат результатів їх роботи. Дещо меншим, але теж досить важливим, є вплив моделі реальної структури слова на внутрішню структуру моделі даних: хоча ці обмеження менш жорсткі, але вони так само можуть внести позитивні зміни в систему за рахунок невеликих накладних витрат.

З іншого боку, складність структури даних для збереження сукупності слів є динамічною, а відповідні набори логічних правил – досить гнучкими. Отже, вже на етапі моделювання система повинна підтримувати складну сукупність правил обробки форм слова – або ж мати можливість формувати і змінювати їх під час роботи.

Обмеживши таким чином поле подальшого дослідження, можемо досить просто визначити класи задач, в яких кожен з цих способів може показати себе кращим чином.

В основі структурних способів лежить представлення слова як певного статичного об’єкту (кореня), від якого є похідними інші форми слова, що утворюються з кореня згідно відомим наборам правил [5]. Така модель має дві вагомі переваги, що природно зумовлені самою її структурою: значна економія дискового простору і великі можливості для додаткової оптимізації на рівні сховища даних. Так, для збереження слова необхідно фактично вказати лише корінь слова та ідентифікатор відповідного набору правил; отже, при наповненні бази зростання використаної пам’яті фактично лінійне. Водночас, за рахунок уніфікованого формату об’єктів даних, внесення змін у певну групу правил автоматично буде розповсюджено на усі слова, що підпорядковані даній групі – тобто, правила можуть бути уточнені або оптимізовані в будь-який момент часу.

Але, на жаль, способам цього класу так само притаманні й два суттєві недоліки: висока обчислювальна складність та загальна жорсткість. Жорсткість полягає у тому, що будь-які виключення з правил необхідно або додавати до загальної системи, що ускладнює алгоритми та створює додаткове обчислювальне навантаження, або на рівні об’єкта, що погіршує швидкодію й ускладнює оптимізацію. Висока обчислювальна складність зумовлена тим, що при обробці кожного слова повністю виконується синтез або аналіз його відповідної форми – а отже, при обробці великих корпусів тексту для отримання адекватної швидкодії доводиться додатково використовувати кешування або базу попередньо підрахованих результатів [6].

Також слід зазначити, що створена в результаті система правил зазвичай велика за обсягом і потребує побудови додаткового рівня абстракції, або, як мінімум, чіткої і повної документації – адже в чистому вигляді вона складна для розуміння як для операторів, так і для самих розробників.

Отже, структурні способи збереження даних за ідеологією досить близькі до баз даних; на відміну від них, способи другого класу – мережеві способи – скоріше можна віднести до семантичних мереж.

Оскільки ключовою особливістю мережевих способів є збереження й обробка окремих словоформ, а не їх класів, вони не можуть змагатися зі структурними способами у швидкодії. Більш того, розмір їх бази з великою вірогідністю не буде значно відрізнятися від стандартної схеми “ключ-значення”. Втім, наслідком цього є й певний позитивний ефект – задачі синтезу та аналізу форми слова зводяться до пошуку запису в базі даних, тобто можуть бути оптимізовані на рівні ядра бази даних.

Але основна перевага способів цього класу лежить дещо в іншій площині: оскільки кожне слово й кожна словоформа зберігаються як окремі сутності, отримана система відкриває цілий спектр нових можливостей. Це не лише значно спрощує інтуїтивне розуміння структури системи користувачами, але й надає нові способи взаємодії з даними.

По-перше, відзначимо доступність усіх словоформ для обробки: стає можливим пошук подібних за написанням слів, виправлення помилок вводу, представлення мережі пов’язаних слів у структурованому вигляді, наприклад у вигляді графу.

По-друге, формування лексеми над певним набором пов’язаних словоформ дозволяє фактично представити розуміння тексту не як сукупності пов’язаних слів, а як відображення сукупності ситуацій реального світу – адже лексема в та-

кому представленні є об’єднанням усіх можливих варіантів текстового представлення одного й того ж об’єкту або явища. Це відкриває великі можливості щодо формування семантичної мережі на основі бази слів, в тому числі з використанням різних видів інформації.

По-третє, таким чином отримуємо ще один шар даних у базі – зв’язки між об’єктами. Ці дані суттєво відрізняються від власне словоформ і дозволяють використовувати абсолютно нові підходи при роботі з усією базою.

Висновки

Проблема формалізації способів збереження слів наразі досить слабко досліджена і має певний потенціал для покращення як кількісних, так і якісних показників роботи інформаційних природно-мовних технологій. Серед задач, в рамках яких слово може бути ефективно виділене в окрему сутність, слід особливо відзначити покращення швидкодії сховищ природно-мовних даних та створення нових способів аналізу семантичних зв’язків між окремими словами.

Основними способами оптимізації системи збереження слів є структурний спосіб (побудова системи правил словотворення і словозміни) та мережевий спосіб (збереження всіх форм слів та системи зв’язків між ними). Ці способи є протилежними полюсами компромісу часу та пам’яті: в структурних способах усі обчислення проводяться при кожному зверненні до бази, що дозволяє зменшити її розмір; в мережевих способах усі слова зберігаються в розкритому вигляді, за рахунок чого обчислення зводяться до пошуку запису в базі за ключем.

Структурні способи мають наступні переваги: невеликий дисковий розмір бази даних, малі витрати часу на операцію пошуку, великий простір для подальшої оптимізації на низькому рівні. Отже, їх використання виправдане за таких умов:

– більшість операцій в програмному засобі є операціями синтезу (пошук словоформи за ідентифікаторами слова та відповідної форми), а не аналізу (визначення кореня та форми слова за даним рядком);

– вхідні дані добре структуровані (критерієм якості системи правил в основі словника є точність, а не повнота).

Приклади таких систем: електронні словники та довідники [7], системи синтезу тексту [8].

Мережеві способи, натомість, характеризуються зберіганням усіх можливих словоформ у вигляді рядків. Це забезпечує можливість роботи з різними словоформами як з окремими сутностями, можливість використання у задачах з

неповною інформацією, відкриває великий простір для використання в задачах аналітики. Відповідно, оптимальні умови для використання таких способів наступні:

– великий обсяг бази (наявність неоднозначних слів та виключень);

– необхідність урахування смислового навантаження слова (в тому числі відокремлення омонімів та багатозначних слів);

– необхідність урахування зв’язків між словами (як-то смислові зв’язки, словотворення, структурні зв’язки);

До таких систем належать, зокрема, системи перевірки правопису, деякі семантичні мережі [9], пошукові системи та системи пошукової оптимізації.

Список використаних джерел

1. *Russom P.* Big data analytics, TDWI best practices report / Philip Russom. – Renton, WA: TDWI Research, 2011. – 35 с.
2. *Шнайер Б.* Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си / Брюс Шнайер. – Москва: Триумф, 2002. – 816 с.
3. PostgreSQL Documentation. Full Text Search. Dictionaries [Електронний ресурс] / The PostgreSQL Global Development Group. – 2015. – Режим доступу до ресурсу : http://www.postgresql.org/docs/9.1/static/textsearch-dictionaries.html.
4. *Henning K.* Performance analysis of MySQL’s FULLTEXT indexes and LIKE queries for full text search [Електронний ресурс] / Koch Henning. – 2013. – Режим доступу до ресурсу: http://makandracards.com/makandra/12813-performance-analysis-of-mysql-s-fulltext-indexes-and-like-queries-for-full-text-search.
5. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение / Андрей Анатольевич Зализняк. – Москва : Рус. яз., 1977.
6. *Smiley D.* Solr 1.4 Enterprise Search Server / D. Smiley, E. Pugh. – Birmingham : Packt Publishing Ltd, 2009. – 337 с.
7. Электронный словарь Мультитран [Електронний ресурс] – Режим доступу до ресурсу : http://www.multitran.ru/.
8. Collective Generation of Natural Image Descriptions / [P. Kuznetsova, V. Ordonez, B. Alexander C. et al.]. – Stony Brook, NY : Department of Computer Science, Stony Brook University, 2012.
9. WordNet, a lexical database of English [Електронний ресурс] – Режим доступу до ресурсу: https://wordnet.princeton.edu/.

УДК 004.82

Ю. І. Кисленко, А. В. Хіміч

ІНФОРМАЦІЙНА БАЗА ЛІНГВІСТИЧНОГО ПРОЦЕСОРА

Анотація: Стаття присвячена особливостям формування допоміжної інформаційної бази лінгвістичного процесора, орієнтованої на аналіз текстів. Основою дослідження постає підхід до структурного рівня мовної організації, що спирається на результати сучасних відомостей щодо мовленнєвої діяльності людини у багатьох суміжних напрямах та знімає багато суперечностей класичної лінгвістики.

Ключові слова: лінгвістичний процесор, індивідуальна мовна система, базова семантико-синтаксична структура, інформаційні природно-мовні технології, інформаційна база лінгвістичного процесора.

Вступ

Все більш нагальною постає проблема «спілкування» з комп’ютером природною мовою. Проте, прогрес у цій сфері поки що реалізується ще досить повільними темпами та не носить системного характеру. Звісно, все це зумовлюється складністю об’єкта дослідження та обмеженістю існуючих знань стосовно функціонування нашої нейромережі. Свідченнями такого стану є визнання розробників ІТ, орієнтованих на опрацювання природно-мовної інформації, стосовно двох ключових проблем у цій царині: відсутністю семантичного WEB та неспроможністю на сьогоднішній день формування потужних баз знань (природно-мовних), де б формувалася та відображувалася модель довкілля (а загалом, і модель світу), яка, власне, і зумовлює процес «розуміння» інформації. Таким чином, ці дві проблеми і визначають актуальність заявленої тематики розвитку цілого кластеру сучасних ІТ, орієнтованих на опрацювання природно-мовної інформації (ІПМТ).

На кафедрі технічної кібернетики НТУУ «КПІ» протягом останніх десятиріч формувався комплексний підхід до структурної організації мови, що базується на результатах сучасних досліджень мовленнєвої діяльності людини у багатьох помежованих напрямках, зокрема: біології, нейрофізіології, психології, філософії, кібернетики тощо. В результаті був сформований інтеграційний підхід до структурного рівня мовної організації, який знімає багато суперечностей класичної грамматики, розглянутих у праці [1] з міркуваннями щодо перспектив синтаксичних досліджень. Висновок з використання такого підходу є досить практичним: «Всі ми гарно та впевнено користуємося мовою (часто-густо не