УДК 004.822/004.652.5

Ю.І. Кисленко, Д.С. Сергеєв

АРХІТЕКТУРА ПРИРОДНО-МОВНОЇ БАЗИ ЗНАНЬ

Анотація: Стаття презентує архітектуру природно-мовної бази знань, що базується на засадах системної організації мови, започаткованої останніми роками на кафедрі технічної кібернетики НТУУ "КПІ". Особливість такого підходу в інтегральному аналізі мовленнєвої діяльності людини з урахуванням сучасних досягнень у багатьох помежованих сферах мовленнєвої діяльності людини. Це, в свою чергу, визначає новий конструктивний підхід до аналізу формування моделі індивідуальної мовленнєвої системи з двома складовими — лінгвістичним процесором та базою знань.

Ключові слова: індивідуальна мовна система, інформаційні природно-мовні технології, лінгвістичний процесор, природно-мовна база знань, квант знань.

Обєкт дослідження

В сучасному світі наукові галузі тісно інтегровані, а міждисциплінарний діалог набув важливого креативного потенціалу. Завдяки цьому в багатьох випадках міжгалузеві проблеми вирішуються досить ефективно та без додаткових ускладнень. Одним з найпомітніших виключень з цього правила є феномен інформаційних природно-мовних технологій (ІПМТ) — сектор ІТ, до якого відносяться пошукові системи, інтернет, експертні системи, системи синтезу/аналізу текстів та мовлення, сучасні бази даних та бази знань, системи накопичення знань тощо.

ІПМТ орієнтовані на моделювання однієї з найскладніших форм інтелектуальної діяльності людини — її мовленнєвої здатності, і лежать на стику технічних та гуманітарних дисциплін. Класичній лінгвістиці притаманна певна нездатність до змін, оскільки її засади були сформовані кілька століть тому і з того часу суттєвих їх змін не відбувалося. Зараз же відбувається активна інтеграція лінгвістики в наукову картину світу, що зазнала за цей час суттєвих змін — і певними результатами цих процесів можемо користуватись вже сьогодні.

Так, ще Л.В. Щерба у 1974 р. [1] засвідчив, що мовленнєва діяльність людини актуалізується індивідуальною мовною системою, яка працює в режимі синтезу та аналізу повідомлення. З цього можемо зробити висновок що для більш-менш адекватного якісного моделювання мовленнєвої діяльності людини ми повинні враховувати у повному обсязі особливості ІМС.

Як було вказано в попередній статті авторів [2], для моделювання мовленнєвої діяльності людини необхідно сформувати обидві складові ІМС: лінгвістичний процесор (ЛП) — як сукупні знання

щодо мовної організації, так і базу знань (БЗ) — де у людини накопичуються всі знання щодо довкілля, в якому людина живе. Саме моделюванню БЗ і призначена дана стаття.

Найкращі БЗ виділяються тим, що містять релевантну інформацію, мають довершені системи пошуку інформації та її накопичення і ретельно пророблену структуру та формат знань. БЗ, що розглядається в рамках статті, орієнтована на опрацювання природно-мовної інформації і, звісно ж, зберігання знання на мовному рівні. Аналіз представлених у наукових статтях та практичних проектах рішень щодо збереження природно-мовної інформації дозволяє зробити висновок про те, що такі рішення в основному знаходяться ще в зародковому стані і не вирішують поставлених задач в повній мірі.

Аналіз існуючих підходів

Можемо умовно розділити наявні рішення на такі основні класи:

Неструктуровані БЗ — тобто ті, в яких найменшим елементом структури є один фрагмент тексту (абзац, речення, але як правило — стаття або подібний за розміром еквівалент). До таких БЗ належать зокрема довідкові системи та системи керування контентом (CMS), як приклад можна навести довідкову систему корпорації *Microsoft*. Слід зазначити, що їх структура на рівнях вище атомарного елемента може бути добре продуманою і загалом якісною, але вона не має прив'язки до характеру інформації: можна замінити статтю на заголовок, ключові слова і фотокопію документа і від цього не матимемо втрати зручності роботи. Це добре ілюструє відкрита база даних "*Меморіал*" — проект по збереженню документів часів II Світової війни: основною формою документа є фотокопія, і хоча розпізнавання частини його тексту полегшує пошук, але ніяк не змінює принципу роботи системи.

До цього класу належить ціла низка спеціалізованих рішень, як-то використання технології *Map-Reduce [3]* для збереження текстових даних або повнотекстові індекси в сучасних реляційних БД. Хоча ці рішення мають велике значення з точки зору практичного використання, в їх основі лежить класична парадигма ключових слів і частотного аналізу.

Інший клас — це *БЗ на основі штучної структури*. В більшості своїй його представники це продукти досліджень в галузі комп'ютерної лінгвістики, що є повністю протилежними неструктурованим БЗ. В БЗ цього класу текстова інформація зберігається у формі об'єкта заданої (або безпосередньо, або через набір обмежень) структури. З одного боку це забезпечує відкритість БЗ для автоматичного використання, що було притаманно неструктурованим БЗ: структура може бути однаково зчитана і представлена. Але з іншого боку встановлюється продиктоване форматом структури обме-

ження на вхідну інформацію. Оскільки в переважній більшості прикладів структура заснована на правилах класичної лінгвістики, а в деяких випадках взагалі лише на авторському баченні, що не підтверджене теоретичними засадами [4], опис такої структури містить або відносно велику кількість виключень, що нівелює її переваги для автоматизації, або таку ж кількість обмежень, що дуже обмежує сферу її використання.

Тут слід згадати причини і історію виникнення програмування як такого і, зокрема, баз знань. Програмування було створене для вирішення в першу чергу математичних задач, які були орієнтовані на проблеми реального світу. Саме такі БЗ (на відміну від БД) зберігають знання про світ, де один елемент описує один реальний об'єкт. Логічним виглядає висновок про те, що для адекватного збереження природно-мовної інформації необхідно обрати певну структуру, що відповідає структурі природно-мовного тексту в реальному житті. Очевидно, що така структура має бути теоретично обґрунтована, а довільно обрані конструкції цієї якості не мають.

Власне основний інтерес в контексті природно-мовних баз знань для нас представляють саме ці два класи, оскільки третій клас можна представити як крайні випадки перших двох. До нього відносимо такі проекти як доступні для редагування енциклопедії, семантичні мережі (такі як *WordNet*) — тобто ті, де інформація пов'язана семантично — тобто БЗ містить різносторонню інформацію і дозволяє отримати як певне розуміння поняття через його зв'язки, так і загальну структуру знань про світ. Відзначимо, що акценти в різних проектах зроблені на різних можливостях: у *WordNet* як представника БЗ зі штучною структурою це повне виродження змісту — для кожного терміна надаються тільки словникові визначення різних його значень; у Вікіпедії структура однієї статті є семантично довільною, що фактично унеможливлює використання її як джерела знань у автоматичних системах, але об'єм цих знань є колосальним.

Системний підхід до структурної організації мови

Автори в якості універсальної структури пропонують базову семантико-синтаксичну структуру (БССС) — двоскладову схему опису довільної ситуації реального чи віртуального світу, всі складові якої актуалізовані на атрибутивному рівні. Ця структура постає похідною від структурно-функціонального рівня нейроорганізації зорового тракту [5].

Монопредикатна БССС у загальному випадку описує ситуацію реального світу, тобто є елементарною часткою — квантом — знань. Довільний мовний матеріал, представляється на монопредикатному (в межах однієї структури) або поліпредикатному (на об'єднанні кількох структур) рівнях, але елементарна структура в обох випадках однакова. Також велике значення при моделюван-

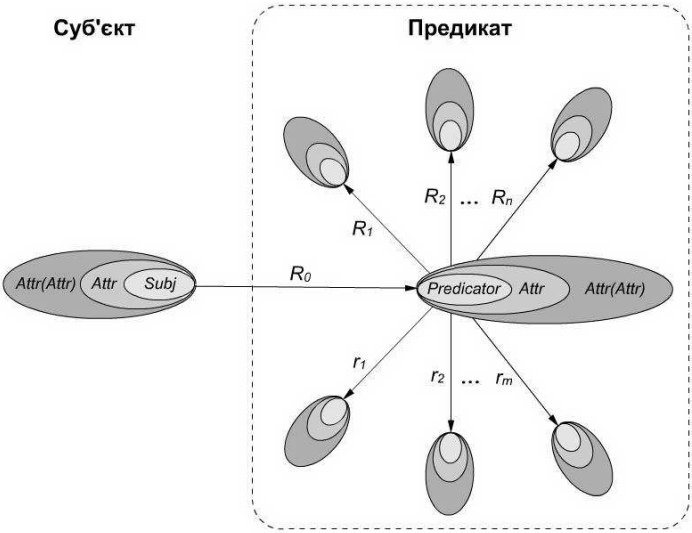

Рис. 1 – Графічний рівень презентації базової семантико-синтаксичної структури Subj – суб'єкт БССС, Predicator – ядро (дієслово) n-актантного предиката, R_0 – головне відношення "мати предикат", $_1, \ldots, R_n$ – предикативні відношення, $r_1, \ldots, r_m$ – ситуаційні відношення, Attr – прикмети складових БССС, Attr(Attr) – міра прикмет

ні БЗ є те, що структура може бути представлена об'єктом в ООП або в нереляційній БД.

За сукупністю цих властивостей можемо стверджувати, що основною інформаційною одиницею організації ПМБЗ постає структура БССС. В даній статті представлено архітектуру розробленої на основі БССС природно-мовної бази знань та опис її окремих елементів.

Особливості запропонованої архітектури БЗ

Таблиці частин мови – представлення кожної з основних частин мови (іменник, прикметник, дієслово та прислівник) та окрема таблиця допоміжних слів (до якої вносяться незмінні слова, наприклад службові частини мови). Особливість архітектури полягає в тому, що кожен об'єкт БЗ враховує особливості словозміни. Це дозволяє використовувати умовний запис посилання на слово і його форму для багатократного звернення до одного елементу таблиці частини мови. В подальшому цей функціонал може бути розширений навіть до реалізації словотворення (наприклад "їхали" — "при-їхали").

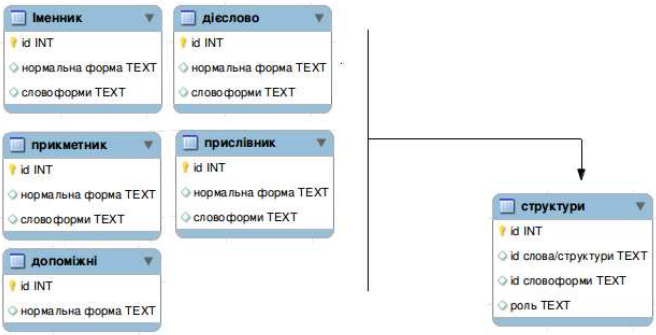

Рис. 2 – Архітектура ПМБЗ В таблиці “структури” елемент має унікальний id, але в його склад входять декілька об’єктів “id слова — id словоформи — роль”

Оскільки ця інформація має переважно незмінний характер, пропонується підключати ці таблиці як окремий модуль або принаймні через інтерфейс для забезпечення можливості використання іншої реалізації. Крім того, в інтерфейсі можливо додавати проміжні функції (наприклад урахування синонімії або словотворення) без внесення змін власне у БЗ.

Таблиця структур — містить текст у вигляді множини БССС. Власне ця таблиця забезпечує перехід від атомарного представлення текстової інформації до структурного. У загальному випадку текст є поліпредикатною структурою, при цьому часто в ролі елементів однієї БССС виступають підрядні структури. До складу ІМС входить ЛП, головною функцією якого є виділення монопредикатних структур з поліпредикатних; отже, власне монопредикатна структура є елементом цієї таблиці. Оскільки просте речення визначається саме монопредикатною структурою, БЗ може використовуватись з простими реченнями шляхом простої емуляції ЛП за умови його відсутності.

Найбільш простим способом представлення речення (як фрагменту тексту) у структурному вигляді є представлення “слово — словоформа — порядковий номер”.

Таблиця 1. - Елементарне представлення фрагменту текста

1	2	3	4
Іменник “людина” наз.в.одн.	Дієслово “йти” теп.час.одн.2о	Незмінне слово “через”	Іменник “парк” наз.в.одн
Людина	йде	через	парк

Слова і їх форми можуть бути представлені елементами таблиць частин мови; так, в даному прикладі складові слова можуть бути представленими ідентифікатором слова з таблиці відповідної частини мови та ідентифікатором його форми. Лінгвістичний процесор (або його емуляція) визначають ролі слів у фрагменті (у прикладі — "людина" – *Subj*, "йде" - *Pred*, "через парк" – *r*), і таким чином отримуємо відокремлену від тексту структуру ситуації.

Таке представлення дозволяє повністю описати одну монопредикатну БССС (всі слова можуть отримати маркер ролі — *Subj, Pred, R, r, Attr, Attr(Attr)*). Ця структура в загальній БЗ представляє квант знань — одиницю інформації про світ. За умови наявності заповненої БЗ вибірка всіх БССС, що містять в ролі суб'єкта дане слово ("людина"), отримуємо всі знання про суб'єкт "людина"; включення в склад БЗ системи маркерів по галузям знань, джерелам тощо дозволяє обмежувати результат вибірки згідно потребам користувача. Таким самим чином замість посилання на частину мови можемо використовувати посилання на БССС для реалізації рекурсивного включення БССС будь-якого рівня вкладеності.

Крім цих базових елементів окремо окреслимо систему маркерів. Оскільки БССС у БЗ представлена через об'єкт, до нього можна додавати довільні атрибути. В якості таких атрибутів можуть зокрема виступати семантичні маркери — автор оригіналу, галузь знань, рік написання, оцінки або коментарів експертів тощо. Система таких атрибутів може навіть бути реалізована окремо від БЗ або в окремих її модулях. Це дозволяє використовувати всю БЗ як, наприклад, спеціалізовану БЗ для певної галузі знань, а також відкриває великі можливості для створення системи пошуку по БЗ з довільним уточненням результатів.

Окрему увагу також слід приділити ЛП — створення такої системи безсумнівно є першочерговим питанням в контексті даного дослідження, але крім створення власне алгоритмічного апарату, ЛП може виникнути потреба включення в БЗ додаткових таблиць з допоміжною інформацією, як-то таблиці ідентифікаторів просторових і часових відношень [6], фразеологізмів та ідіом, абревіатур тощо.

Висновки

Запропонований вище підхід до системної організації мови дозволяє вже безпосередньо підійти до моделювання індивідуальної мовної системи і її складових — лінгвістичного процесора та бази знань, а отже — і до формування нової ідеології проектування всього кластеру ІТ, орієнтованих на опрацювання природномовної інформації. Розроблена архітектура БЗ дозволяє моделювати сприйняття, збереження та опрацювання мовних структур довільного повідомлення в структурованому вигляді. Запропонова-

но також детальний опис основних складових цієї архітектури та пропозиції щодо розширення її додатковими можливостями.

Список використаних джерел

1. Щерба Л.В. Языковая система и речевая деятельность. – Л.: Наука, 1974.
2. Кисленко Ю.І., Сергеєв Д.С. Структурно-функціональний рівень організації природно-мовної бази знань. — Автоматичні системи автоматичного управління. – 2013. – №2(23).
3. Lin J., Dyer C. Data-intensive Text Processing with MapReduce. - California:Morgan & Claypool Publishers, 2010.
4. Rosenfield, Roni, Zhu, Xiaojin, Chen, Stanley F. Whole-Sentence Exponential Language Models: A Vehicle for Linguistic-Statistical Integration. – *Computers Speech and Language*, 15(1). – Pittsburgh:Carnegie Mellon University, Computer Science Department, 2001. – p. 14.
5. Кисленко Ю. И. От мысли к знанию (нейрофизиологические основания) - монография – Киев, "Український літопис", 2008 – 102 с.
6. Кисленко Ю.І., Черевко О.С. Категорії часу та простору в інформаційних природно-мовних технологіях. – Автоматичні системи автоматичного управління. – 2011. – №13(38).

Отримано 20.03.2014 р.