УДК 004.822

Ю.І. Кисленко, Д.С. Сергеєв

СТРУКТУРНО-ФУНКЦІОНАЛЬНИЙ РІВЕНЬ ОРГАНІЗАЦІЇ ПРИРОДНО-МОВНОЇ БАЗИ ЗНАНЬ

Анотація: З позицій системної організації мови, сформованої на підґрунті інтегрального підходу до аналізу мовленнєвої діяльності людини, розглядаються проблеми формування архітектури природно-мовної бази знань як складової індивідуальної мовної системи, що постає основою формування всіх сучасних технологій, орієнтованих на опрацювання природно-мовної інформації.

Ключові слова: БССС, природно-мовні технології, ситуація, природномовно база знань.

Серед розмаїття сучасних ІТ все більшу вагу займають технології, орієнтовані на опрацювання природно-мовної (ПМ) інформації. Визначимо ці технології як ***інформаційні природно-мовні технології*** (ІПМТ). На сьогодні ІПМТ складають потужний кластер ІТ, до якого входять всі пошукові системи, інтернет, експертні системи, системи синтезу/аналізу текстів та мовлення, сучасні бази даних та бази знань, системи накопичення знань тощо. Аналізуючи сучасний стан інформаційних технологій зазначеного кластеру, отримуємо не зовсім втішний висновок: всі вони не виправдовують окреслені надії, бо не моделюють у повному обсязі особливості мовленнєвої діяльності людини.

За свідченням видатних лінгвістів (наприклад, Л.В.Щерба [1]) мовленнєва діяльність людини актуалізується ***індивідуальною мовною системою*** (ІМС), де чітко виділяються дві складові: знання, що стосуються мовної організації (умовно позначимо цю складову як ***лінгвістичний процесор*** - ЛП), та вся сукупність накопичених знань щодо довкілля, в якому людина живе (позначимо її як ***база знань*** - БЗ).

Розробникам і користувачам ІТ добре відомо, що будь-які технології функціонують тим ефективніше, чим повніше в них закладені особливості предметної сфери, для роботи з якою вони створені. ІПМТ ж орієнтовані на моделювання однієї з найскладніших форм інтелектуальної діяльності людини – її мовленнєвої здатності, що актуалізується індивідуальною мовною системою. Отже, для більш-менш адекватного якісного моделювання мовленнєвої діяльності людини ми повинні у всі ІПМТ закладати у повному обсязі особливості організації і функціонування ІМС.

Що ж ми маємо на сьогоднішній день? Практично за всіма напрямами реалізованих ІПМТ фахівці стверджують, що до якісного моделювання інтелектуальних процесів мовленнєвої діяльності

ще досить далеко. Інтернет не зовсім задовольняє вимогам користувачів через відсутність семантичного WEB та моделі світу, експертні системи найчастіше працюють в якості радника, автоматичний переклад не використовується кваліфікованими перекладачами, природно-мовні бази знань як і системи автоматичного накопичення знань взагалі відсутні.

На кафедрі технічної кібернетики НТУУ "КПІ" Кисленком Ю.І. [2,3,4], сформовано інтегральний підхід до аналізу мовленнєвої діяльності, що враховує сучасні досягнення аналізу мовленнєвої організації людини у багатьох суміжних напрямах – нейрофізіологія, психологія, філософія, кібернетика тощо. В результаті був сформований новий погляд на структурну організацію мови, що знімає багато суперечностей класичної лінгвістики. Ключовими поняттями такого підходу постає формальне визначення ***ситуації*** та схема її вербалізації на мовному рівні у вигляді ***базової семантично-синтаксичної структури*** (БССС), що визначається основою структурної організації всього мовного матеріалу.

Це дозволяє з нових позицій підійти до моделювання ІМС як основи формування відповідного кластеру ІПМТ. По-новому виглядатимуть у структурному плані основні складові ІМС: ЛП вважається головним модулем ідентифікації структури повідомлення; БЗ постає основою накопичення знань та формування моделі світу, де відбуваються, власне, процеси сприйняття та інтерпретації повідомлення. У структурному плані ЛП та БЗ визначаються діалектичною єдністю, а їх взаємодія постійно реалізується через відповідний інтерфейс. На даний момент вже підходимо до моделювання всіх складових ІМС: започаткована робота стосовно структурно-функціонального рівня організації ЛП [5], змодельована ситуація використання запропонованого підходу для інформаційного пошуку [6], розглядаються можливості формування електронної оболонки під природно-мовну базу знань (ПМБЗ), а також розглядаються особливості та можливості реалізації інтерфейсу між складовими ІМС.

Запропонована стаття саме і пов'язана з аналізом структурно-функціонального рівня організації електронної оболонки природно-мовної бази знань та можливостями її реалізації з урахуванням досягнень сучасних ІТ. Платформа формування ПМБЗ саме і визначається запропонованим інтегральним підходом до аналізу мовленнєвої діяльності людини, який відтворюється ланцюжком***: ситуація*** – як частка довкілля, що існує у триєдності часу, простору та дії, сприймається і опрацьовується зоровим аналізатором; ***БССС*** – як схема трансляції ситуації на мовний рівень, як окремий квант знань, з яких формується наше уявлення про довкілля; ***ПМБЗ*** – як частка (символічна форма) наших знань, головна функція яких – інтерпретація (розуміння) мовного повідомлення.

 ISSN 1560-8956

Отже, ***елементом сприйняття, запам'ятовування, пошуку, зчитування, накопичення та формування нових знань постає БССС*** – як квант (елемент) знань, з яких формується вже конкретне знання з подальшою можливістю формування та поповнення моделі світу як узагальненої структури наших знань про життя та існування конкретного індивіда у цьому світі а також про формалізовані сфери різних напрямів знань та наукових досліджень.

На структурному рівні основою формування БЗ як і ЛП індивідуальної мовної системи постає окрема ***базова семантико-синтаксична структура*** як структурний компонент організації наших знань, що визначений на змістовному, графічному та формальному рівнях. Лінгвістичний процесор повинен виконувати всю роботу по ідентифікації структури повідомлення, включаючи процедуру декомпозиції вхідного тексту за структурами БССС, з подальшою передачею їх послідовності до БЗ. При цьому, формується певний фрагмент нових знань, що доповнює попередню версію моделі світу. Отож, головною особливістю структурної організації запропонованої архітектури бази знань (ПМБЗ) постає необхідність вважати БССС основним структурним компонентом ***(елементом, квантом знань),*** множина яких формує як модель світу, так і модель особистості.

Структура БССС представлена на рис.1; ***БССС – двоскладова монопредикатна схема 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 [4]. Довільний мовний матеріал, у загальному випадку, представляється на монопредикатному або полі предикатному рівнях. ***Монопредикатний*** рівень визначається структурами, що не виходять за межі БССС, тоді як ***поліпредикатний*** – визначається множиною повних або спрощених мовних структур. Як моно, так і поліпредикатні рівні визначаються своїми особливостями організації та поєднання. Окрім того, довільний текст характеризується ще особливостями зв'язності, які теж повинні враховуватися при опрацюванні текстової інформації. Тож, основною інформаційною одиницею організації ПМБЗ постає структура БССС. Дана робота пов'язана, власне, з особливостями формування електронної оболонки для збереження пов'язаної множини структур БССС.

Структура БЗ, таким чином, повинна враховувати особливості структурної організації мовного повідомлення, поданого через множину однотипних БССС-утворень. Тож, на цьому шляху необхідно враховувати певні особливості формування як окремої структури БССС ***(квант знань),*** так і особливості їх поєднання в межах окремого повідомлення (речення) та всього тексту ***(фрагмент знань).*** Можемо вже більш-менш чітко окреслити головні

вимоги до електронної оболонки запропонованої БЗ, орієнтованої на сприйняття та опрацювання природно-мовної інформації. Головні функціональні вимоги, що закладатимуться до БЗ, виглядатимуть наступним чином.

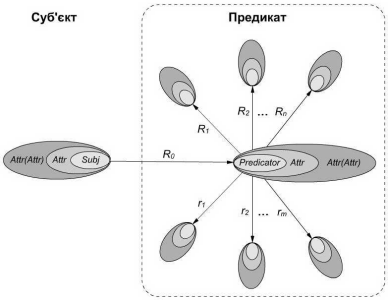

Рис. 1 – Графічний рівень презентації базової семантико-синтаксичної структури

Subj – суб'єкт БССС, Predicator – ядро (дієслово) n-актантного предиката , R_0 *– головне відношення "мати предикат",* $R_1, ..., R_n$ *– предикативні відношення* , $r_1, ..., r_m$ *– ситуаційні відношення, Attr – прикмети складових БССС, Attr(Attr) – міра прикмет*.

Загальні вимоги до електронної оболонки ПМБЗ

1 – ПМБЗ повинна бути комфортною до сприйняття довільного повідомлення; будь-який текст, його фрагмент або частка повинні відтворюватися, зберігатися та зчитуватися у незмінному вигляді;

2 – лінгвістичний процесор (як лінгвістичний фільтр) виконує процедуру декомпозиції окремого повідомлення через множину стандартних структур БССС, поєднаних відповідним чином;

3 – елементом (квантом знань) постає окрема БССС, з множини яких формуються окремі фрагменти знань;

4 – кожна БССС мовними засобами відтворює певну ситуацію (у триєдності часу, простору та дії), з яких формується спочатку окремий квант знань, а надалі – вже й певний фрагмент;

5 – кожна складова БССС в БЗ представлена лише одним вузлом, що враховує особливості словозміни;

6 – БЗ повинна проектуватися з урахуванням можливості автоматичного формування, накопичення та використання знань;

7 – конфігурація БЗ передбачає можливість її роботи в режимах взаємодії з ЛП, моделюючи певним чином процеси "розуміння" повідомлення.

На шляху проектування бази знань важливо вирішити комплекс питань, пов'язаних зі структурною організацією електронної оболонки ПМБЗ. У загальному випадку, структурно-

функціональні особливості організації БЗ багато в чому визначаються особливостями функціонування ЛП, який, власне, і визначає структурні особливості декомпозованого тексту, що надходить до БЗ. Найбільш важливими вбачаються наступні.

Особливості ідентифікації окремої БССС лінгвістичним процесором

В реальних текстах зв'язки атрибутивного рівня не представлені в явному вигляді або взагалі відсутні. Відповідно, ЛП повинен вирішувати задачу виділення таких зв'язків у вихідному тексті для врахування особливостей реалізації конкретних БССС при накопиченні в БЗ з наступним їх можливим відтворенням на шляху синтезу повідомлення. В залежності від складності зв'язків ця задача може вимагати збереження додаткової інформації в базі даних, що обслуговує ЛП.

Зв'язки Obj/Subj – Attr(Obj/Subj) з урахуванням можливості пре/пост позиції,

зв'язки Attr(Obj/Subj) – Attr(Attr) з урахуванням можливості пре/пост позиції,

зв'язки Mov – Attr (Mov) з урахуванням можливості пре/пост позиції,

зв'язки Attr (Mov) – Attr(Attr) з урахуванням можливості пре/пост позиції,

Ідентифікація головного відношення R_0,

Ідентифікація предикативних відношень$R_1, \ldots, R_n$,

Ідентифікація сірконстант $r_1, \ldots, r_m$,

Ідентифікація кількісних характеристик.

Особливості опрацювання довільного повідомлення (поліпредикатний рівень)

Мінімальним квантом знань в тексті є окрема БССС. Оскільки довільне повідомлення, у загальному випадку, може бути представлене більш ніж однією БССС, то ЛП повинен мати можливість ідентифікувати в межах окремого повідомлення всі БССС-утворення. Важливо враховувати також, що БССС відтворює, власне, лише структуру даних в реченні, але не весь вхідний текст з його особливостями конкретної реалізації (мається на увазі: порядок слів, наявність ідіом, вставних конструкцій тощо). Так, коли структура БССС не враховує порядок слів, а для флективних мов вона може бути розгорнена мільярдами варіантів, то для кожного вхідного речення необхідно формувати свій унікальний маркер, що містить інформацію про порядок слів та додаткові службові елементи. Всі ці особливості повинні знайти своє адекватне втілення в організації бази знань.

Вище згадувалося, що коли БЗ та ЛП у функціональному плані діалектично пов'язані між собою (ЛП працює на БЗ, а остання,

в свою чергу, допомагає ЛП аналізувати текст), то часто ЛП і БЗ звертаються за допоміжною інформацією, що зберігається в БД лінгвістичного процесора. Одне з головних завдань ЛП — аналіз тексту, і для цього він повинен мати власний словник; цей словник має містити всі словоформи або таблиці флексій, як можливе доповнення — синоніми і однокореневі слова. Сучасні електронні словники містять всі необхідні для цього дані, і в якості словника ЛП може бути використаний фактично будь-який з них. Слід також враховувати, що як для ЛП, так і для БЗ важливо мати інформацію не лише про особливості словозміни але й відносно словотворення (наприклад, можливості трансформування дієслова за схемами дієприкметника, дієприслівника та субстантива). Окрім того, є певні мовні засоби, що не входять явним чином до БССС, але необхідні для аналізу та синтезу тексту. До них зокрема належать вставні слова і конструкції, ознаки часу, простору, причини, комунікативні та ідіоматичні засоби. Кількість таких часто вживаних мовних засобів для флективних мов порівняно невелика, тому має сенс зберігати їх у БД ЛП.

На рівні текстового повідомлення

Оскільки БЗ побудована на основі ПМ-текстів, необхідно щоб вона підтримувала зв'язки між окремими мовними фрагментами (від речення і більше). Для реалізації цього автори пропонують використовувати рекурсивні маркери: маркери по реченнями, що побудовані над БССС, маркери по абзацам, що побудовані над маркерами по реченням тощо. Це дозволяє зберігати мовні структури будь-якої складності у вигляді маркерів, використовуючи єдину базу БССС як основу БЗ. За допомогою маркерів представляється можливим зберігати елементи зв'язності в межах тексту — зберігати зв'язок між відомою частиною і відповідною невідомою у пов'язаних реченнях.

Також маркери можуть мати власні атрибути, що вказують на певні особливості джерела тексту. Так, маркер речення-цитати може вказувати на джерело цитати, маркер вищого порядку — на номер глави, маркер тексту — на книгу, що є його джерелом, на автора, на галузь знань тощо. Такі атрибути маркерів можуть бути представлені як окремий прошарок, а можуть бути пов'язаними з відповідними сутностями в БЗ.

Потребує подальших роздумів можливість виділити мережу маркерів в окремий структурний модуль в межах БЗ, адже це відділяє суто інформаційний рівень (відображення ситуації) від текстового рівня (незмінний вхідний текст).

Істотною перевагою ПМБЗ є, крім іншого, простота її наповнення. Оскільки аналіз тексту виконується в межах ЛП, робота людини-оператора зводиться до вибору джерел інформації і встановлення їх маркерів на рівні об'ємів, що відповідають книзі. Так,

 ISSN 1560-8956

оператор може задати ступінь достовірності джерела, галузі знань, до яких належить подана в джерелі інформація, вказати додаткові зв'язки з певними об'єктами БЗ — і ці маркери зберігаються окремо від власне інформації і можуть бути видалені чи виправлені без втручання в БЗ як таку.

Таким чином підходимо вже безпосередньо до ситуації, коли в одній БЗ знаходяться записи з багатьох галузей знань. При розширенні будь-якої з них виникають нові зв'язки, і в результаті утворюється міждисциплінарна БЗ, в якій складається картина світу за аналогією з картиною світу людини як особистості.

Створена за наведеними вище принципами БЗ вирішує питання багатовимірності ПМ-інформації. Оскільки БССС описує ситуацію як триєдність часу, простору та руху, тобто відображає ситуацію реального світу, можливо за допомогою однієї БССС описати майже нескінченну кількість ситуацій. БЗ містить інформацію про світ, ЛП дозволяє розширювати БЗ новими знаннями, а мережа маркерів дає інструментарій для структурування даних. Таким чином, ПМБЗ дозволяє через складність зв'язків розкривати ситуацію реального світу як ситуацію символічного світу, тобто зберігати і обробляти накопичені знання без втрат, які виникають при введенні додаткового апарату для їх структурування згідно обмежень технічних засобів.

Біблиографічний список

1. Щерба Л.В. Языковая система и речевая деятельность / Л.В. Щерба. – Л.: Наука, 1974.
2. Кисленко Ю. І. Системна організація мови: Монографія. / Ю.І. Кисленко – К.: Український літопис, 1997. - 217 с.
3. Кисленко Ю.І. Архітектура мови (лінгвістичне забезпечення інтелектуальних інтегрованих систем) : Учбовий посібник. / Ю.І. Кисленко – К.: Віпол, 1998. - 343 с.
4. Кисленко Ю. И. От мысли к знанию (нейрофизиологические основания) - монография / Ю.И. Кисленко – К.: "Український літопис", 2008 – 102 стр.
5. Структурно-функціональний рівень організації лінгвістичного процессора: збірник праць конференції "Штучний інтелект - 2012" / Ю.І. Кисленко, О.С. Черевко – Кацивелі, 2012. – Т.3. – С. 43–51.
6. Проблемы и перспективы развития поисковых систем: сборник трудов конференции "Искусственный интеллект - 2011" / Ю.И. Кисленко, А.В. Терентьев. – Кацивели, 2011. – Т.3. – С. 55–66.

Отримано 07.10.2013 р.