УДК 355.40

Informatics and Mathematical Methods in Simulation
Vol. 3 (2013), No. 1, pp. 75-81

МОДЕЛЬ ЛІНГВІСТИЧНОЇ БАЗИ ДАНИХ В СИСТЕМАХ АВТОМАТИЧНОЇ ОБРОБКИ ПРИРОДНОМОВНОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

І.В. Замаруєва, В.Б. Толубко, Л.О. Литвиненко, О.Ю. Ніколаєвський

Військовий інститут Київського національного університету імені Тараса Шевченка,
просп. Глушкова, 2, корпус 8, Київ, 03680, Україна

У статті запропоновано модель лінгвістичної бази даних як складової лінгвістичного забезпечення систем автоматичної обробки природномовної текстової інформації. Відмінність запропонованої моделі полягає у відокремленні знань про мову від знань про світ (предметну область). Розроблена модель лінгвістичної бази даних у порівнянні з іншими має значно менший обсяг словників, формат представлення даних дозволяє моделювати закономірності як флективних мов, так й аналітичних, здатна забезпечувати обробку природномовного тексту для різних прикладних задач.

Ключові слова: лінгвістична база даних, лінгвістичне забезпечення, системи автоматичної обробки текстової інформації, мовна система, семіотична система

Вступ

Лінгвістичне забезпечення є невід'ємною складовою систем автоматичної обробки природномовної текстової інформації (ПМТІ). До таких систем сьогодні відносять системи машинного перекладу, інформаційно-пошукові системи, системи автоматичного індексування й класифікації ПМТІ, автоматичного реферування тощо. Ядром лінгвістичного забезпечення є лінгвістична база даних (ЛБД), яка складається із різноманітних словників заданого формату і забезпечує задачі автоматичної обробки ПМТІ.

Досвід побудови ЛБД свідчить, що визначення форматів словникових статей є окремою дослідницькою задачею, оскільки від повноти урахування лексико-граматичних характеристик в кінцевому рахунку залежить якість автоматичної обробки ПМТІ. Аналіз підходів до побудови ЛБД [1–4] показав, що вони орієнтовані здебільшого як на вузьку предметну область, так і на обмежену кількість задач автоматичної обробки (наприклад, автоматична класифікація ПМТІ, машинний переклад). Недоліком такого підходу є те, що розроблені словникові бази не можуть бути замінними як між собою, так і використовуватися в інших задачах обробки ПМТІ, які на сьогодні виникають (наприклад, в інформаційно-аналітичних системах з обробки ПМТІ). В той же час ЛБД моделює закономірності мовної системи, які не залежать від предметної області та розв'язуваних задач автоматичної обробки ПМТІ.

Мета статті і постановка задачі дослідження

Метою статті є розробка моделі ЛБД, яка б задовольняла вимогам незалежності від предметної області, мови представлення тексту, розв'язуваної задачі автоматичної обробки тексту.

При постановці задачі дослідження автори виходили з таких міркувань: архітектура ЛБД визначається об'єктом моделювання і задачами щодо його автоматичного опрацювання.

Об'єктом моделювання є природномовний текст. При цьому текст розглядається як об'єкт різних рівнів опрацювання: як знакова система (тобто правила графемного оформлення тексту), як лінгвістична система (тобто знання про мову, якою представлений текст) і як система знань про світ (предметну область). Кожний рівень має свої особливості, свої засоби виразу і, отже, припускає наявність специфічних методів обробки. Універсальними задачами опрацювання тексту є аналіз, прагматична інтерпретація й синтез.

Автори розрізняють лінгвістичну базу даних, яка представляє формальний запис знань про вхідну мову, та інформаційну базу даних (ІБД), яка представляє формальний запис знань про світ (прикладну область). До останньої можна віднести, наприклад, електронні перекладні словники в системах машинного перекладу, тезауруси в інформаційно-пошукових системах тощо.

Лінгвістична база даних представляється четвіркою:

$$LBD =: \langle G, Lm, Lsyn, Lsam \rangle,$$

де

G – словники для аналізу й синтезу графемної структури тексту;

Lm – словники для аналізу й синтезу морфологічного рівня мовної системи;

$Lsyn$ – словники для аналізу й синтезу синтаксичного рівня мовної системи;

$Lsam$ – словники для аналізу й синтезу семантичного рівня мовної системи.

До словників висуваються наступні вимоги: формат представлення граматичної інформації кожного рівня представлення словників є входом для аналізу наступного рівня, аналітичні словники будуються із урахуванням потреб синтетичних словників. Ми не розглядаємо задачу прагматичної інтерпретації, оскільки коректність цієї задачі залежить від ІБД.

Запропонована модель експериментально досліджена на текстах, представлених російською, українською та англійською мовами. Для кожної вхідної мови будується окрема ЛБД, але формати представлення даних є уніфікованими, що дозволяє їх реалізацію у багатомовних системах машинного перекладу, оскільки в ЛБД представлена як аналітична, так і синтетична складова. Далі розглянемо детальніше структуру і зміст словників кожного рівня.

Виклад основного матеріалу дослідження

Структура і зміст словників, що забезпечують автоматичний аналіз вхідного тексту на знаковому рівні його представлення. Кінцевою метою розпізнавання на графемному рівні представлення тексту є побудова графемної структури тексту, яка включає виділення на множині рядків і графем вхідного тексту таких семантично самостійних одиниць тексту: фрагментів (дискурсів), речень, синтагм, лексем; визначення типів (класів) перелічених одиниць тексту та встановлення відношень між ними в певному вхідному тексті [5].

В основу класифікатора графем покладені такі ознаки: тип знаку (цифра, буква, синтаксичний знак, службовий знак тощо), належність до алфавіту (латиниця, кирилиця, виключно російська, виключно українська), розмір літер (прописна, заголовна), фонетичні ознаки (голосна, приголосна). Аналітичний словник цього знакового рівня представлення тексту задає правила утворення лексем. Правила утворення лексем для російської мови представлені в таблиці 1.

Таблиця 1.

Правила утворення лексем для російської мови

Код класу	Найменування лексеми	Опис правила	Приклад
L01	Слово з малої літери	([а-я]+)	танк
L02	З великої літери	([А-Я][а-я]*)	Михаил
L03	Іншомовне слово	[a-zA-Z]+	tomahawk
L04	Неповна лексема	([а-я]+)-	полу-
L05	Слово через дефіс	([а-я]+)-([а-я]+)	полу-автоматический
L06	Слово з латинської літери	[A-Z]-([а-я]+)	U-образный
L08	Неповне позначення	-[0-9]+[A-Z]+	-104G
L09	Ціле число	[0-9]+	1945
L07	Позначення	([0-9А-Яа-я]+[-./]+[0-9А-Яа-я]+)|([0-9A-Za-z]+[-./]+[0-9A-Za-z]+)	МиГ-27; F.82
L10	Дробове число	[0-9]+[,.][0-9]+	143.7
L11	Дата	[0-9]+[-./][0-9]+[-./][0-9]+	03.07.1993
L12	Числов. прикметник	[0-9]+-[а-я]+	5-и
L13		"	"
L14	Скорочення	[цкнгшщхфпрлджчсмтб]+[.]	г.
L15	Скорочене ім’я	[А-Я] [цкнгшщзхфвпрлджчсмтб] + [.]| [А-Я] [.]	Дж.
L20	Спец. символ	[№@#$%^&*]	#
L21	Абревіатура	[А-Я]{2,}	США
L23	Повне ім’я	L15+ L02	Дж. Буш
L25	Складне скорочення	[цкнгшщзхфвпрлджчсмтб]+/[цкнгшщзхфвпрлджчсмтб]+	м/с
L30		([А-Я][а-я]*)-([А-Я][а-я]*)	Гаусса-Остроградского
L31		([А-Я][а-я]*)-([а-я]+)	
L32	Електронне посилання	((http|ftp)\:\/\/[a-z0-9./-]+)	http://univ.kiev.ua
L33	Електронна адреса	^(\w+([_.]{1}\w+)*@\w+([_.]{1}\w+)*\.[A-Za-z]{2,3}[;]?)*$	info@univ.kiev.ua

Синтетичний словник знакового рівня задає правила транслітерації, такий словник використовується в системах машинного перекладу для забезпечення перекладу власних назв, які не мають аналогів у мові, якою перекладається текст (наприклад: «Верховна Рада» англ. «*Verchovna Rada*»). У лівому полі такого словника послідовність букв мови-оригіналу, правому полі послідовність букв мови-перекладу. Довжина ланцюга букв визначається однозначністю передачі фонетичного звучання власної назви і коливається від 1 до 5.

Структура і зміст словників морфологічного рівня мовної системи. На цьому рівні аналізуються лише лексеми класів *L01***,** *L02***,** *L05* (див. табл. 1), тобто ті класи, для яких можна визначити лексико-граматичну інформацію. Даний рівень представлений словозмінною та словотвірною моделями мови.

Словозмінна модель визначається задачами автоматичного опрацювання текстів на морфологічному рівні:

1) Автоматичний морфологічний аналіз (АМА) – приписування граматичної інформації словоформам вхідного тексту.

2) Лематизація словоформ – приведення словоформи з тексту до початкової форми.

3) Синтез словоформ (синтезувати словоформу з початкової форми у відповідну необхідну форму для вживання у перекладеному тексті).

Для забезпечення першої задачі укладається аналітичний морфологічний словник, для 2 і 3 синтетичний морфологічний словник.

У розроблюваній ЛБД використано флективний підхід за словником квазізакінчень, який будується автоматично. Вхідними даними для процедури автоматичної побудови словника квазіфлексій є обернений словник словоформ певної вхідної мови з відповідною лексико-граматичною інформацією. До словника словоформ висуваються такі вимоги:

- лексика словника не містить службових частин мови і числівника (у тому числі й порядкового), оскільки дана лексика не визначає предметну галузь, має досить обмежений обсяг (близько 1 тис. словоформ), а тому автори вважають, що такий словник доцільно подавати у вигляді словника словоформ);
- лексико-граматична інформація повинна мати єдину граматичну параметризацію для російської, української та англійської мов та максимально враховувати функціональні особливості словоформи в тексті, формальний запис (код) лексико-граматичної інформації має забезпечувати незалежний доступ до кожної окремої граматичної категорії.

З метою забезпечення другої вимоги розроблено позиційно-цифрове кодування лексико-граматичної інформації [6]. Першим кодом (до символу «*») йде зазначення лексико-граматичного класу, а надалі граматичні характеристики у фіксованому порядку (рід, число, відмінок, особа, вид, стан, ступінь, час, істотність). Кожна граматична характеристика має певний набір параметрів, яким відповідає цифра від 0 до 9. При чому, для всіх характеристик значення 0 та 9 є загальними, а саме 0 – не визначається для даної словоформи взагалі, 9 – невизначена характеристика. Цифри від 1 до 8 мають конкретні значення для кожної характеристики. Наприклад, категорія «*число*» має такі характеристики: 1 – однина, 2 – множина, 3 – тільки однина, 4 – тільки множина. Якщо словоформа має декілька морфологічних значень (присутня омонімія) – то усі можливі комбінації задаються через символ «/». Наприклад, російській словоформі «*стекло*» буде відповідати така словникова стаття:

стекло 1*311000002/1*314000009/8*313021001/8*314021009/8*316021002/

Процес побудови словника квазіфлексій полягає в тому, що розглядаються словоформи з кінця слова в алфавітному порядку. Квазіфлексією в даному разі виступає набір літер з кінця слова. Починається процес з визначення найбільш часто вживаного коду (набору кодів) для однієї останньої літери – наприклад, літери «*а*». Надалі визначається найбільш вживаний код для квазіфлексії «*ба*», якщо він співпадає з кодом для «*а*», то така квазіфлексія не вноситься до словника, інакше – закінчення вноситься до словника та процес продовжується. Якщо на якомусь етапі визначиться, що для більш коротка квазіфлексія відповідає такому самому набору кодів, що і інше, яке включає таку квазіфлексію, то більша за розміром квазіфлексія видаляється зі словника. Такий прохід робиться по всім літерам алфавіту. На виході ми отримуємо словник квазіфлексій у вигляді власне квазіфлексій та відповідних їм визначених наборів кодів.

Лематизаційний словник являє собою словник квазіфлексій, ліва частина представляє мінімальну послідовність букв, яку необхідно відкинути від слова, а права частина – необхідно мінімальний фрагмент коду, на підставі якого розпізнана лексема

приводиться до початкової форми та послідовність букв, яку необхідно додати. Наприклад, російська словоформа *«полка»* має два набори граматичної інформації:

1) іменник жіночого роду, називного відмінку, однини;
2) іменник чоловічого роду, родового відмінку однини.

Стаття лематизаційного словника квазіфлексій для словоформи *«полка»* буде мати такий вигляд: *лка 1*1:лк/1*2:лка*.

Парадигматичний словник квазіфлексій будується аналогічно. В правій частині послідовність букв, яку необхідно відкинути від слова в початковій формі, а в лівій – парадигма цього слова. Нижче представлений фрагмент лематизаційного парадигматичного словника квазіфлексій для української мови.

а 1*2/а/и/і/у/ою/і/о/и/б/ів/ам/и//ів/ах/и/

б 1*11/б/ба//бу/бу//бові/б//ба/бом/бі/бе/би/бів/бам/би//бів/бами/бах/

йиб
2*11/бий/бого/бому/бий//бого/бим/бому//бім/ба/бої/бій/бу/бою/бій/бе/бого/бому/бе/бим/бому//бім/бі/бих/бим/бі//бих/бими/бих/

аб 1*41/ба/би/бі/бу/бою/бі/бо/би/б/бам/би//б/бами/бах/

Оскільки парадигматичний словник має значно більшу довжину рядка, то власне квазіфлексію виділено жирним шрифтом. Через риску подається парадигма слова у фіксованому порядку відповідно до кожного лексико-граматичного класу.

Структура і зміст словників синтаксичного рівня мовної системи. Словники цього рівня визначають правила синтаксичної сполучуваності в межах речення. Вхідними даними для побудови словників синтаксичної сполучуваності є дані морфологічного аналізу. Декларативне представлення правил синтаксису вхідної мови поділяється на контекстні синтаксичні правила, правила виокремлення присудка і підмета та правила виокремлення другорядних членів речення.

До контекстних синтаксичних правил відносять правила: узгодження, керування, прилягання. Формат опису правил представлений в таблиці 2.

В таблиці 2 знаком «+» показано за якою позицією застосовується правило перетинання значень морфологічного коду, 2 у позиції відмінок означає, що у другої словоформи має бути родовий відмінок, С1 – правило узгодження, У1-У2 – правила керування, П1 – правило прилягання.

Таблиця 2.

Формат опису правил

Який клас	З яким класом	рід	число	відмінок	Порядковий № гол. сл.	Синтаксичний тип	Приклад
2*	1*	+	+	+	(2)	С1	державні органи
23*	1*			2	(2)	У1	від форми
1*	1*			2	(1)	У2	генератор шуму
14*	8*				(2)	П1	швидко біг

Аналогічно задаються інші правила. Декларативне представлення правил синтаксису у вигляді уніфікованих таблиць дає можливість спростити алгоритм обробки на синтаксичному рівні мовної системи, звівши його фактично до обробки даних таблиці. Крім того це дає можливість одним модулем обробляти тексти,

представлені різними мовами, оскільки результат обробки залежить не від алгоритму обробки, а від коректності даних, представлених в таблиці.

Структура і зміст словників семантичного рівня мовної системи. Оскільки семантика тексту сильно корелюється з представленням знань про світ (предметну область), які ми включаємо до ІБД, то ми пропонуємо в ЛБД включати лише обмежений словник – словник, який не увійшов до словника квазізакінчень (це так звані службові частини мови: частка, прийменник, займенник, сполучник; до цього словника ми також відносимо числівник). Такий підхід обумовлений тим, що перелічені класи слів не характеризують предметну область і визначають допоміжну функцію у розумінні тексту. Крім того такий словник є обмежений обсягом – не перевищує 2 тис. словникових статей. Для виконання вимоги узгодженості словників ЛБД та ІБД для представлення семантичної інформації розроблено семантичний позиційно-цифровий код, який за форматом збігається із морфологічним кодом, але має інше змістове наповнення. Словникова стаття має такий вигляд: у левій частині словоформа у правій – код її семантичного класу, який відокремлюється знаком «*» від значень відповідних їй семантичних характеристик. Семантичні характеристики представлені 9-позиційним кодом, де перша позиція – час (як семантична категорія), 2 – простір; 3 – кількість; 4 – якість; 5 – предметність; 6 – явище; 7 – процес; 8 – стан; 9 – властивість. Значення кожної категорії визначаються на інтервалі [0;9], при цьому 0 і 9 мають фіксовані інтерпретації значень: 0 – не визначається; 9 – невизначено. Наприклад, російському слову «*возле*», буде відповідати така словникова стаття:

возле 25*010000000,

де 25* – означає семантичний клас відношення; 1 у другій позиції, що слово характеризується часом з конкретним значенням «*знаходитися в Q-межах*».

Висновки

Таким чином, запропонована модель ЛБД здатна забезпечувати обробку природномовного тексту для різних прикладних задач. За рахунок розмежування ЛБД по рівням мовної системи є змога використовувати лише окремі словники, так наприклад, для інформаційно-пошукових систем достатньо словників морфологічного рівня. Розмежування ІБД і ЛБД дозволяє значно спростити процедуру ведення електронних перекладних словників, звівши їх до формату слово – перекладний відповідник. Обсяг запропонованої моделі ЛБД значно менший. За рахунок використання словників квазізакінчень (як аналітичних, так і синтезуючих) обсяг всіх словників ЛБД не перевищує 25 тис. словникових одиниць. Декларативне представлення синтаксичних правил дозволяє покращити якість перекладу, оскільки повну синтаксичну структуру на сьогодні не будує жодна система перекладу, а збільшення обсягу словникових статей (близько 4 млн. в системі RETRANS) не дозволяє гарантувати обробку нового тексту.

Список літератури

1. Дарчук, Н.П. Комп’ютерна лінгвістика (автоматичне опрацювання тексту) [Текст] : підруч. для студ. вищ. навч. закл. / Н.П. Дарчук ; Київ. нац. ун-т ім. Т. Шевченка. — К. : Київ. ун-т, 2008. — 351 с.
2. Белоногов, Г.Г. Автоматизированные информационные системы [Текст] / Г.Г. Белоногов, В.И. Богатырев ; под общ. ред. К.С. Тараканова. — М. : Советское радио, 1973. — 328 с.
3. Grishman, R. TIPSTER text phase II architecture design / R. Grishman // TIPSTER’96 Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996. — PP. 249–305.

4. Грязнухина, Т.О. Система автоматичного морфологічного аналізу українського наукового тексту / Т.О. Грязнухина, М.В. Нікула // Проблеми українізації комп'ютерів. Матеріали 2-ої Міжнар. конф. — Київ, 1993. — С. 42–46.
5. Балабін, В.В. Доморфемна обробка текстів в системах машинного перекладу / В.В. Балабін, І.В. Замаруєва // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. — 2008. — Вип. 11. — С. 78–84.
6. Пампуха, І.В. Побудова та використання словника квазізакінчень / І.В. Пампуха, Л.О. Литвиненко, О.Ю. Ніколаєвський та ін. // Збірник наукових праць Військового институту КНУ ім. Тараса Шевченка. — К., 2008. — № 14. — С. 154–158.

МОДЕЛЬ ЛИНГВИСТИЧЕСКОЙ БАЗЫ ДАННЫХ В СИСТЕМАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННО-ЯЗЫКОВОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

И.В. Замаруева, В.Б. Толубко, Л.А. Литвиненко, А.Ю. Николаевский

Военный институт Киевского национального университета имени Тараса Шевченко,
просп. Глушкова, 2, корпус 8, Киев, 03680, Украина

В работе предложена модель организации лингвистической базы данных как составляющей лингвистического обеспечения систем автоматической обработки естественно-языковой текстовой информации. Особенность предложенной модели состоит в отделении знаний о языковой системе от знаний о мире (предметной области). Разработанная модель лингвистической базы данных по сравнению с известными характеризуется значительно меньшими объемами словарей, формат представления данных позволяет моделировать закономерности как флективных языков, так и аналитических. Она способна обеспечивать обработку естественно-языкового текста для различных прикладных задач.

Ключевые слова: лингвистическая база данных, лингвистическое обеспечение, системы автоматической обработки текстовой информации, языковая система, семиотическая система

MODEL OF LINGUISTIC DATA BASE IN AUTOMATIC PROCESSING OF NATURAL LANGUAGE TEXT SYSTEMS

Irina V. Zamaruieva, Vladimir B. Tolubko, Leonid O. Lytvynenko, Oleksandr Yu. Nikolaevsky

Military Institute, Taras Shevchenko National University of Kyiv,
2 Glushkova Ave., build. 8, Kyiv, 03680, Ukraine

The paper presents the model of the organization of the linguistic database as part of linguistic support of the automatic processing of natural language text information. The peculiarity of the proposed model is to separate the knowledge of the language system of knowledge about the world (domain). The developed model of a linguistic database in comparison with known characterized by significantly lower volumes of dictionaries, data format allows you to simulate patterns as inflected languages and analytical. It is able to provide the processing of natural language text for a variety of applications.

Keywords: linguistic database, linguistic software, automatic text processing, language system, semiotic system