УДК 81’322.4
Литвиненко Л.О. (КНУ)

ОСОБЛИВОСТІ ПОБУДОВИ ЛІНГВІСТИЧНОГО ПРОЦЕСОРА ПАРАЛЕЛЬНОЇ ОБРОБКИ ПРИРОДНОМОВНОГО ТЕКСТУ

Розглянуто концептуальні положення побудови знання-орієнтованих систем машинного перекладу. Проаналізовано підходи до організації лінгвістичних процесорів автоматичної обробки різномовних природно-мовних текстів. Визначено умови розподіленої автоматичної обробки текстів. Розроблено архітектуру лінгвістичного процесора паралельної обробки тексту.
Ключові слова: знання-орієнтована система машинного перекладу, природно-мовний текст, лінгвістичний процесор, паралельна обробка даних.

Рассмотрены концептуальные положения построения знание-ориентированных систем машинного перевода. Проанализированы подходы к организации лингвистических процессоров автоматической обработки разноязычных естественно-языковых текстов. Определены условия распределенной автоматической обработки текстов. Разработана архитектура лингвистического процессора параллельной обработки текста.
Ключевые слова: знание-ориентированная система машинного перевода, естественно-языковой текст, лингвистический процессор, параллельная обработка данных.

The article considers conceptual principles of building knowledge-oriented machine translation systems. It is analyzed the approaches to the organization of linguistic processors automatic processing of multilingual natural language text. The conditions for a distributed automatic text processing are defined. It is developed the architecture of the language processor parallel processing text.
Keywords: knowledge-oriented machine translation system, natural language text, the linguistic processor, parallel processing.

Постановки проблеми. Проблема комп’ютерного моделювання процесу “розуміння” природно-мовному тексту (ПМТ) відноситься до задач штучного інтелекту і потребує розвинутих засобів програмного забезпечення. Відповідно лінгвістичний процесор (ЛП) має забезпечувати обробку ПМТ на всіх рівнях організації тексту:

- знаковому (тобто семіотичної системи);
- на рівні мовної організації (цей рівень включає морфологічний, синтаксичний і семантичний рівні обробки вхідного ПМТ);
- прагматичному, тобто на рівні відбиття знань про навколишній світ у вхідному ПМТ.

В існуючих системах машинного перекладу перший і останній рівень обробки ПМТ практично відсутні.

Призначенням третього рівня є інтегрування понятійної структури тексту до бази знань про предметну галузь. Робота прагматичного інтерпретатора залежить від моделі представлення знань про предметну галузь. Так, наприклад, для аналізу воєнно-політичної обстановки важливо, що слово *Major* в певному контексті означає не просто особу (на відміну від інших перекладних еквівалентів: *головний, майор* тощо), а ***першу*** особу в державі (колишній прем’єр-міністр Великобританії) і, відповідно, контекстне супроводження цього слова набуває іншого прагматичного значення. Тезаурус прагматичного рівня містить енциклопедичні знання про конкретну предметну галузь (ПрГ). Такі знання, як правило, в тексті розпізнати неможливо, оскільки вони не ідентифікуються мовними засобами чи супроводжувальним контекстом. Так, наприклад, російське словосполучення "*беличья клетка"* в електротехніці не має нічого спільного із загально прийнятим значенням в мові. Крім того, частина загально прийнятих знань, як правило, в тексті не супроводжуються відповідним контекстом. Так, наприклад, слова *Росія, США* не будуть супроводжуватися словами *держава, країна.* Реально прагматичний аналізатор починає працювати, починаючи

з доморфемного аналізу. Звертання до нього відбувається на всіх рівнях організації тексту. На заключному етапі він виконує інтегруючу функцію.

Сучасні теоретичні дослідження з ідеології розробки ЛП базуються на двох підходах:

1) послідовна обробка тексту;
2) інтегральна обробка тексту.

Перший підхід передбачає послідовну обробку всього тексту спочатку на граматичному рівні (морфологічному), потім реалізується етап синтаксичного аналізу, потім – семантичного. Перший підхід передбачає розподіленого подання даних. Принципи реалізації першого підходу покладені в розробку більшості ЛП з автоматичної обробки ПМТ.

Основою другого підходу є гіпотеза, яка вперше була сформульована Р. Шенком та Л. Бірнаумом [1], сутність якої можна висловити наступним чином:

- синтаксичні й семантичні структури обробляються одночасно;
- синтаксис і семантика реалізуються в ході одного процесу;
- обробка мовних повідомлень за своєю природою тотожна обробці пам‘яті.

Другий підхід потребує сумісного подання всієї інформації до текстової одиниці, обробка тексту при цьому підході здійснюється по реченнях.

Слід зазначити, що перевагою першого підходу є:

можливість зупинитися на будь якому етапі обробки, що сприяє більш ефективній “відладці” програмних модулів;

окреме подання даних робить систему в технологічному плані більш гнучкою до нових прикладних задач.

Недоліком таких систем є те, що виникнення відмов в роботі ЛП на будь-якому етапі призводить до невиконання задачі в цілому.

Цікаво, що другий підхід є дзеркальним відображенням першого, тобто недоліки першого підходу є перевагою другого, переваги першого – недоліками другого.

Формулювання цілей статті. Метою даної статті є дослідження умов паралельної автоматичної обробки природно-мовного тексту в багатомовній знання-орієнтованій системі.

Виклад основного матеріалу. В основі запропонованої концепції багатомовного машинного перекладу лежить знання-орієнтована технологія, сутність якої полягає в комплексному розв‘язанні завдань автоматизації вилучення, подання і обробки знань з ПрГ, які містяться в різномовних текстових джерел. Особливості аналізу ПМТ визначаються спрямованістю на формування поняттєвої структури, тобто на автоматичний витяг знань з різномовних текстів та їх прагматичну інтерпретацію в термінах прикладної задачі. При цьому текст розглядається як об'єкт різних рівнів аналізу: як знакова система, як граматична система і як система знань про світ (проблемну область) [2,3]. Кожний рівень має свої особливості, свої засоби виразу і, отже, припускає наявність специфічних методів обробки. В основу пропонованого ЛП покладена гіпотеза розподіленої обробки ПМТ, її сутність зводиться до наступних положень:

- ПМТ являє собою єдність трьох різних систем: семіотичної системи, лінгвістичної системи та системи знань про світ (предметну галузь);
- процес обробки ПМТ (людиною) проходить в трьох системах одночасно;
- в рамках кожної системи обробка здійснюється по спіралі.

Ядром ЛП є алгоритми обробки ПМТ, до складу яких входять алгоритми доступу й обробки до лінгвістичної бази даних (ЛБД) та бази знань з ПрГ. Традиційно ЛП в системах машинного перекладу передбачає два блоки: аналізу та синтезу. Робота зі знаннями з ПрГ, які містяться в ПМТ, потребує розвинутих засобів інтерпретації мовних одиниць в термінах саме знань (понять) як в ПрГ, так і відносно цільових настанов про прикладну задачу (знання про прикладну задачу). Це обумовило розробку ЛП, який би включав три самостійних блоки: аналізу, інтерпретації та синтезу. Структурно-логічна схема лінгвістичного процесора, яка відбиває сутність обробки ПМТ, наведена на рис. 1. Теоретичний і практичний базис створення таких ЛП закладені в системах штучного інтелекту для підтримки природно-

мовного інтерфейсу користувача [4,5]. Але, слід зазначити, що задачі інтерпретації в системах підтримка діалогової взаємодії людини й ЕОМ значною мірою відрізняється від задач обробки ПМТ.

Відмітною особливістю обробки ПМТ є залучення різних видів знань на кожному з її етапів. Змістова сутність задач, які вирішуються лінгвістичним процесором на кожному з

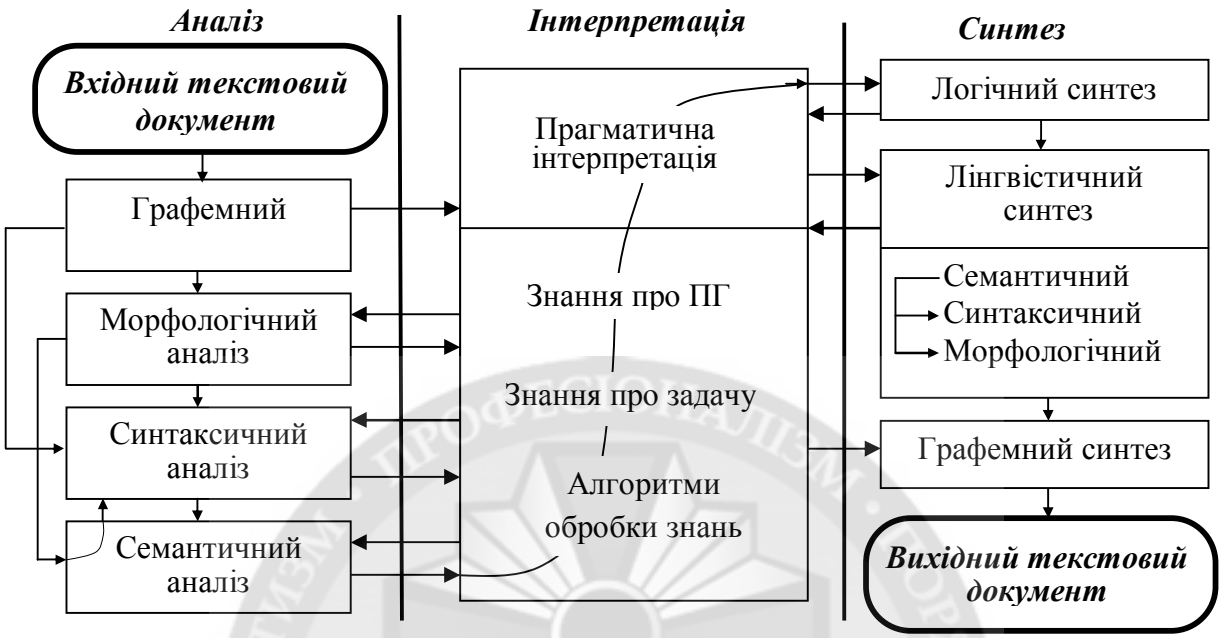

Рис. 1. Структурно-логічна схема лінгвістичного процесора

рівнів організації тексту, представлено на рис. 2. Включення блоку інтерпретації як самостійного на кожному рівні аналізу вхідного тексту дозволяє, з одного боку, розподілити процес обробки як на рівні подання лінгвістичних даних, так і на етапах їх обробки. З іншого боку, введення інтерпретатора дозволяє зробити незалежною логіко-семантичну (засобами формальної логіки) обробку знань від певної вхідної мови. "Скачковість" обробки полягає в тому, що на кожному з етапів обробки текстової інформації здійснюється занурення отриманих результатів обробки в знання про світ. Такий підхід дозволяє обернути на переваги недоліки в рамках перших двох підходів. Це проявляється в тому, що кожний етап обробки моделюється як незалежний модуль, який дозволяє отримати точнішу інформацію. Так, на етапі морфологічної обробки занурення в базу знань про ПрГ дозволяє точніше визначити граматичні характеристики таких лексичних одиниць, як *ім‘я, назва* тощо. Наприклад, в англійській мові для повного ім‘я: *Martha Browner* визначити категорію роду можливо лише за рахунок інтерпретації лексеми *Martha* на базі знань з ПрГ, де для першої словоформи *Martha* буде визначено: жіночий рід, істота (людина). За рахунок роботи інтерпретатора зменшується кількість помилок як в процесі аналізу вхідного тексту, так і в процесі синтезу. Крім того, імітація об‘ємної (тобто в трьох вимірах) обробки інформації дозволяє значно скоротити час обробки тексту.

Висновки. Отже, реалізація знання-орієнтованої технології побудови багатомовної системи машинного перекладу передбачає розробку ЛП, який має забезпечувати процеси автоматичного розпізнавання, формалізації та обробки знань з ПрГ, що містяться в ПМТ. Концепція побудови ЛП базується на таких принципових положеннях:

- природно-мовний текст є відображенням трьох взаємопов‘язаних об‘єктів аналізу: семіотичної системи, граматичної структури певної мови та системи знань про навколишній світ;

▪ фрагменти знань, які описуються в природно-мовних текстах, відбивають стан фахового (або, в загальному випадку, логіко-семантичного) проникнення в ПрГ, а не певної природної мови.

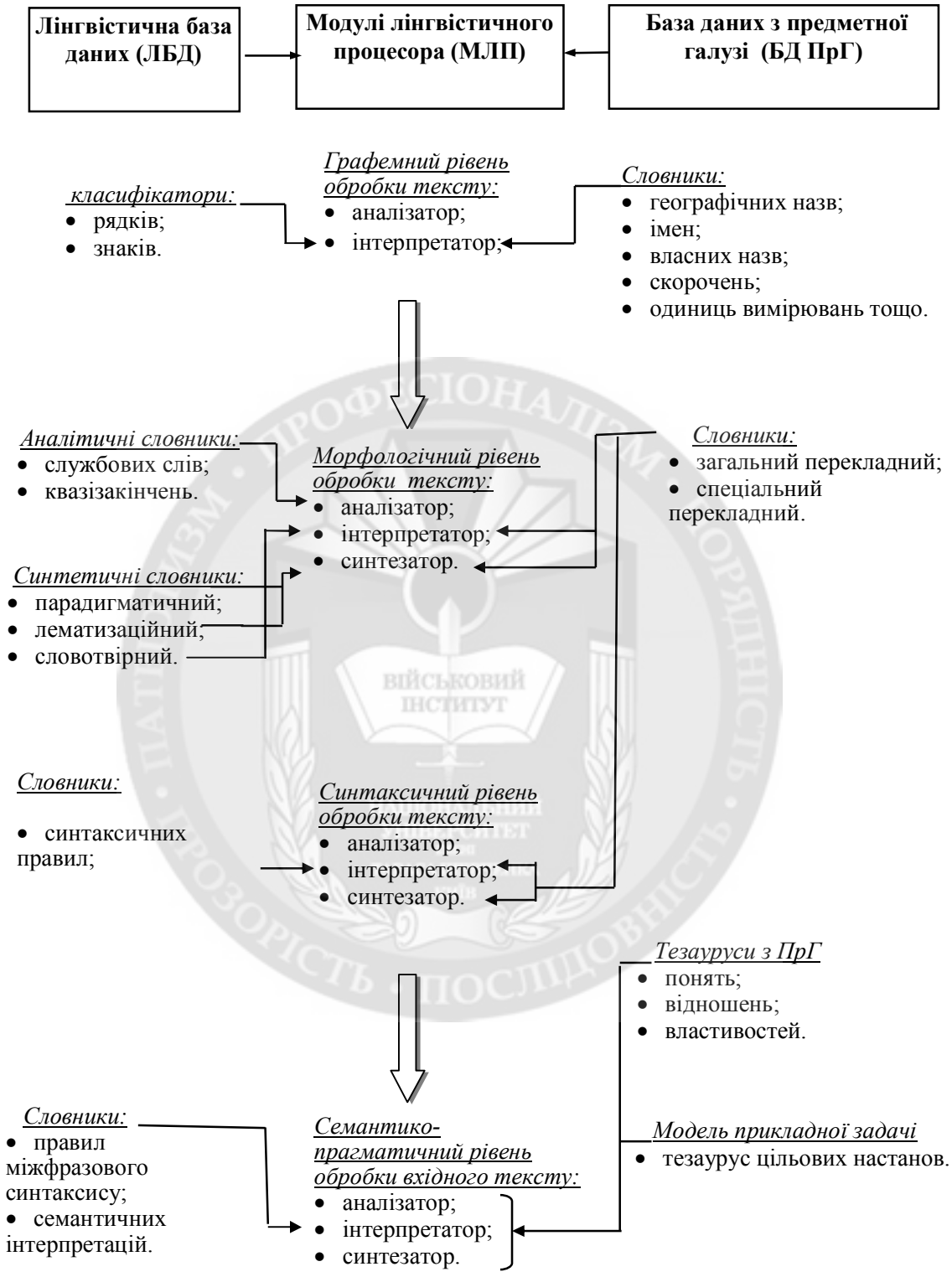

Рис. 2. Структурно-логічна схема паралельної обробки даних

Запропонована архітектура лінгвістичного процесора введенні передбачає крім традиційних модулів обробки: аналізу і синтезу, включення інтерпретатора як самостійного

модуля, виконуючого функції відображення знань про ПрГ мовними засобами в процесі побудови поняттєвої структури за ПМТ та синтезу природно-мовних текстів за їх поняттєвою структурою. Це дозволило, з одного боку, підвищити якість лінгвістичного аналізу на морфологічному, синтаксичному і семантичному етапах, з іншого – скоротити час програмної обробки тексту, за рахунок розпаралелювання процесу обробки даних.

ЛІТЕРАТУРА:

1. Дерецкий В.А. Об одном подходе к обработке естественно-языковых данных на основе анализа семантических сетей // Праці Першої міжнародної конференції з програмування УкрПРОГ‘98. – К.: Кібцентр НАНУ, 2-4 вересня 1998. – С. 405-411.

2. Балабін В.В., Замаруєва І.В. Побудова систем машинного перекладу на основі знання-орієнтованого підходу // К.: ВІ КНУ. – Збірник наукових праць ВІ КНУ. – 2006. – №2. – С.68-74.

3. Балабін В.В. Автоматизація когнітивного розпізнавання текстових об’єктів в умовах багатозначності і невизначеності / В.В. Балабін, І.В. Замаруєва // Збірник наукових праць ВІ КНУ. – К.: ВІ КНУ, 2007. – №6. – С.76-84.

4. Кузин Е.С. Представление знаний в системе интеллектуального интерфейса / Учен. записки Тартуского ун-та, Тарту, 1984. – № 688. – С. 13-22.

5. Широков В.А. Технолингвистика: модели данных и проблемы программирования // Перша міжнародна наук.-практ. конференція. УкрПРОГ‘98.- Київ: Кібцентр НАНУ, 1998. – С.451-461.

Рецензент: д.т.н., проф. Замаруєва І.В.