іУДК 004. 4 : 004. 087

Олена ПОРХУН,
мол. наук. співроб.

Перспективні шляхи розвитку програмного забезпечення в роботі з електронними документами

У статті розглядається питання розбробки системи лінгвістичного аналізу, які використовуються для пошуку, класифікації, кластеризації, реферування та автоматичного перекладу текстових документів. Акцентується увага на використанні семантичного аналізу для забезпечення якості процесів обробки текстової інформації, дається характеристика модулів системи розуміння тексту для виконання семантичного аналізу. Розглядається питання автоматичного реферування та тематичної навігації в інформаційних масивах.

Сьогодні актуальність галузі ефективної обробки текстової інформації складно переоцінити. Це зумовлено в першу чергу тим, що експоненційне зростання обсягів текстової інформації потребує постійного вдосконалення технологій доступу до неї.

Галузь обробки текстової інформації (Information Retrieval) з кожним днем породжує нові та дедалі складніші завдання. Можна виокремити цілий ряд класичних проблем: пошук документів, маршрутизація запиту, фільтрація, класифікація, кластеризація текстових документів, реферування та автоматичний переклад текстів тощо.

У США та країнах Західної Європи протягом останніх 20 років активно виконуються науково-дослідницькі проекти, присвячені лінгвістичній тематиці та створенню систем інтелектуальної обробки текстів. Уряди та комерційні установи інвестують мільйони євро в проекти даного напряму. В умовах швидкого зростання обсягів поточної інформації, яку необхідно щоденно обробляти, системи автоматичної обробки текстів стають необхідними для роботи в усіх галузях.

Ціни на розроблене лінгвістичне програмне забезпечення на софтверному ринку вимірюються десятками, а іноді й сотнями тисяч євро за одне робоче місце (Американська система аналізу текстів CONVERA коштує приблизно 100 тис. дол. за одне робоче місце). Саме тому актуальність створення вітчизняних систем інтелектуальної обробки текстів сьогодні неможливо переоцінити.

Центральну роль у системах лінгвістичного аналізу відіграють лінгвістичні глобальні бази знань, на основі яких функціонують алгоритмічні програмні модулі аналізу. Розробка цих баз є найбільш трудомісткою роботою та потребує тисячі людино-годин праці високо-кваліфікованих комп'ютерних лінгвістів. Треба зазначити, що підготовка фахівця такого рівня триває понад 15 років. Саме все це в сукупності обумовлює необхідність залучення великих коштів у процес розробки лінгвістичного програмного забезпечення та закладає, безумовно, недешеву ціну на лінгвістичне програмне забезпечення.

Сьогодні серед систем пошуку та аналізу текстів найбільш відомими є американська система CONVERA, російські інтелектуальні метапошукові системи «Сиріус», Solarix Intellectronix, Exactus.

Сьогодні в інформаційно-пошукових системах процес пошуку документів базується в основному на попередньому процесі автоматичної індексації за словами. Використання для пошуку документів інформаційнопошукових тезаурусів є недостатньо актуальним через велику трудомісткість та низьку швидкість ручного індексування. Альтернативою індексації за словами є автоматична концептуальна індексація за поняттями тезаурусів (дескрипторів), спеціально розроблених як інструмент для автоматичної обробки текстів. У результаті автоматичної концептуальної індексації для кожного тексту будується не послівний індекс, а індекс за дескрипторами тезауруса, можливе розширення запиту за синонімами.

При цьому необхідно вирішувати питання, пов'язані з представленням у тезаурусі багатозначних термінів, а саме: як та наскільки детально мають бути описані різні значення багатозначних термінів, щоб такий опис виявився базою для ефективного вирішення багатозначності термінів у процесі автоматичного індексування.

Треба зазначити, що основним недоліком сучасних пошукових систем і програм текстового моніторингу та аналізу є відсутність врахування семантичної структури текстів. Хоча навіть при поверхневому семантичному аналізі структур речень тексту якість пошуку стає набагато кращою при несуттєвих втратах часу.

Застосування семантичного аналізу обумовлено прагненням поліпшити якість процесів обробки текстової інформації. Оперуючи з формальним змістом текстів на природній мові, можна досягти більшої повноти та точності результатів при вирішенні завдань пошуку, класифікації, реферування, фільтрації текстових документів.

Семантичний аналіз полягає у визначенні інформативності текстової інформації та виділенні інформаційно-логічної основи тексту. Для виконання семантичного аналізу розробляється так звана Система розуміння текстів (СРТ).

Семантичний аналіз тексту припускає розробку та застосування основних процедур «розуміння» або «здобуття знань». Результат виконання цих процедур формалізується у вигляді деякої семантичної структури. Машинне розуміння тексту можна розглядати як процес формування семантичного образу для аналізованого тексту на природній мові, що виконується СРТ.

СРТ базується на спеціальних модулях, що реалізуються лінгвосемантичним та програмним забезпеченням. Перше використовується для опису моделі предметної області та представлене лінгвістичним та семантичним словниками, у термінах яких СРТ формує образ тексту. Програмне забезпечення реалізує методи семантичного аналізу. Роботу СРТ можна поділити на два етапи: лін-

гвістична обробка та семантична інтерпретація, що виконується, відповідно, лінгвістичним та семантичним модулями СРТ.

Лінгвістичний модуль поєднує етапи безпосередньої обробки текстів. На цих етапах відбувається первинна формалізація речень вхідного тексту. Кожен етап використовує словники лінгвістичного забезпечення. На етапі граматичного аналізу виділяються текстові одиниці: слова, речення та абзаци.

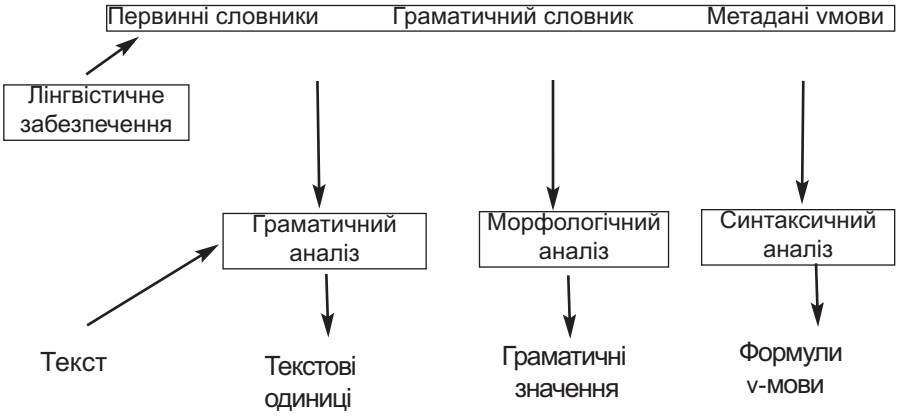

Схема 1.Схема лінгвістичного модуля СРТ

Крім того, на цьому етапі виконується виключення так званих stop-слів та більш складних конструкцій, таких як вступні речення.

На етапі морфологічного аналізу визначаються граматичні значення слів, такі як частина мови, рід, число, відмінок тощо.На етапі синтаксичного аналізу визначається синтаксична структура речення, що описується деякою формулою V-мови. Робота семантичного модуля наведена на рис. 2.

Семантичний модуль виконує смислову обробку тексту, вхідні дані представлені V-формулами, отриманими лінгвістичним модулем. Даний вид обробки називається інтерпретацією, оскільки відповідно до закладеної в словниках семантичного забезпечення моделі предметної області виконується визначення формального

змісту окремих формул V-мови. Ця процедура виконується на етапі семантичного аналізу. На етапі міжфразового семантичного аналізу виконується об'єднання семантичних представлень окремих речень у єдину семантичну мережу, що описує зміст усього тексту.

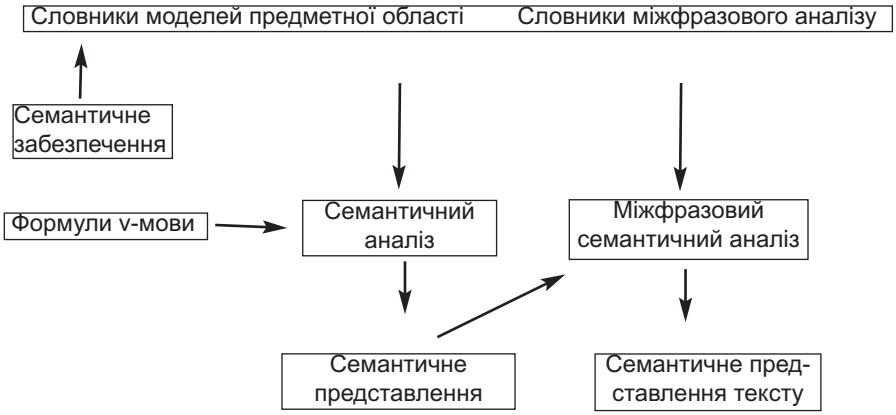

Схема 2. Схема семантичного модуля СРТ

Серед створених програм автоматичного реферування широко відомі такі сучасні системи, як Copernicus summariser, Viewsum summariser, Inxight Summarizer.

Більшість існуючих алгоритмів реферування базуються на виборі в тексті речень, які містять найбільшу кількість часто вживаних понять, іноді з врахуванням зв'язків між поняттями. Таким чином, інформаційний портрет документа відкривається у формі послідовності цитат, відібраних з оригіналу.

Основна проблема в забезпеченні якості роботи існуючих систем автоматичного реферування полягає у використанні виключно частотного аналізу для визначення найбільш важливих сегментів тексту. При цьому не враховується складна об'ємна багатовимірна структура тексту, що зумовлює перетворення вихід-

ного документа на набір невдало пов'язаних речень, який важко вважати текстом реферату.

Використання семантичної мережі понять, пов'язаних з дієсловами, дає змогу формувати основні ідеї тексту, відображені в часто вживаних поняттях та зв'язках, у вигляді простих речень. Окремою проблемою є вибір оптимального порядку фраз. Одним з варіантів її вирішення може бути визначення комунікативної структури тексту – ієрархії тем та рем, яка відображує логіку викладення автором матеріалу. Завдання-тема рематичного аналізу розв'язується у процесі синтаксичного аналізу фрази – поняття з групи підмета є теми, поняття-доповнення до дієслова-реми, які можуть стати темами наступних фраз. Знання синтаксичних ролей слів у реченні також дає змогу ранжувати їх з точки зору важливості для автора фрази. У комбінації з алгоритмами статистичного аналізу ці факти сприяють більш точному ранжуванню понять за їх важливістю в інформаційному портреті документа.

Унаслідок постійного зростання кількості повнотекстових документів в електронному вигляді з'являється дедалі більше нових методів навігації в інформаційних масивах. Сьогодні текстову інформацію звичайно представляють у формі гіпертексту, який надає можливість інтерактивної роботи з матеріалом та багатовимірністю його представлення. При цьому конкретні реалізації гіпертексту відрізняються як способом встановлення зв'язків, так і формою візуального відображення, починаючи з простих систем типу веб-сторінок та довідкових систем, в яких використовується перехід по тексту за допомогою чітко заданих розробниками прийомів, та закінчуючи «інтелектуальними» електронними книгами, де кожне слово супроводжується рядом гіперпосилань, представлених поняттями, пов'язаними за змістом.

Створення пошукових машин в Інтернеті та збільшення обсягів інформації, що публікується, стимулювало розвиток гіпертекстових засобів нового покоління, що називаються тематичними навігаторами. Системи з подібними засобами дають змогу

переміщуватися зв'язаними тематичними категоріями (рубриками), до кожної з яких може належати множина текстів, близьких за змістом. За допомогою тематичних навігаторів можна визначити теми, що об'єднують потрібні тексти (наприклад, тексти, що містять певні слова), а потім переміщуватися цими темами.

Усі відомі тематичні навігатори поділяються на дві категорії. До першої відносяться навігатори, які мають чітко задану структуру з апріорі встановленими темами та зв'язками між ними. У таких навігаторах використовується наперед визначений рубрикатор з ієрархічною структурою категорій, що відображає загальноприйнятий набір галузей знань. Нижні гілки рубрикатора зазвичай містять у собі класи слів мови, що належать до певних тем. Подібні навігатори можуть автоматично розподіляти всі вхідні тексти за відповідними тематичними рубриками та підраховувати, на які з гілок припадає більше слів з тексту.

Для створення навігаторів другої категорії потрібна участь експертів для формування структури тем на основі аналізу змісту корпусу текстів. У цьому випадку гіпертекстова структура часто являє собою семантичну мережу, зв'язки якої зображують актуальну побудову текстів з семантичної точки зору. Такі навігатори через високі витрати на розробку, що потребує залучення «ручної» праці, як правило, призначаються для невеликих корпусів текстів у вузькій предметній галузі.

У рамках проекту «Образний комп'ютер», що реалізується в Міжнародному науково-навчальному центрі інформаційних технологій та систем, пропонується розробка та програмна реалізація нових ефективних лінгвістичних алгоритмів, які базуються на онтологічному аналізі текстів на природній мові з застосуванням семантичних онтологічних баз знань, та створення алгоритмів семантичного контекстного аналізу текстів на основі цих моделей та методів.

Розробляються системи смислової тематичної рубрикації, кластеризації, якісного моніторингу та семантичного аналізу, які

дають змогу визначати тематику тексту та здійснювати пошук, впорядкування та смислову обробку документів, які належать до тих тем, що цікавлять користувача.

Система смислової тематичної рубрикації текстів дає можливість здійснювати семантичний аналіз текстів та визначати за смисловими структурами їх тематику з подальшим присвоєнням тематичних індексів, відповідним розміщенням, тематичним впорядкуванням, статистичною обробкою тощо.

За допомогою системи семантичної кластеризації з використанням функцій блоку семантичного контекстного аналізу текстів користувач може в автоматичному режимі здійснювати розбиття потоків текстових документів та текстових архівів на тематичні групи/підгрупи на основі глибинного аналізу смислових структур текстів.

Система якісного моніторингу текстових документів та потоків природномовних текстів дає змогу користувачу запитувати теми, які його найбільше цікавлять, та відповідно до цього здійснювати моніторинг текстових документів, які відповідають заданим тематикам та інформаційним об'єктам, а також обчислює негативні та позитивні якісні оцінки даних об'єктів у цих текстах.

При розробці комп'ютерно-лінгвістичних алгоритмів виникають складні наукові проблеми, які пов'язані з необхідністю формалізації нечітких механізмів природної мови. Для подолання цих проблем створюється концептуальна алгоритмічна модель функціонування природної мови, яка включає створення формальної моделі тексту, допоміжні бази даних, словники та евристичні алгоритми розбору слів та речень тексту, а також методи визначення семантичної міри близькості слів та концептів.

Як свідчить практика, статистичні методи аналізу тексту, на яких до цього часу були сконцентровані зусилля розробників інтелектуальних систем обробки текстів, досягли своєї межі. Подальше ускладнення математики без залучення серйозних

лінгвістичних досліджень не дасть змоги помітно поліпшити якість подібних систем.

Навіть при обмеженості більшості сучасних синтаксичних аналізаторів тексту, які працюють без залучення семантики, є всі підстави стверджувати, що їх включення спроможне підвищити точність роботи статистичних аналізаторів, а також відкрити якісно нові можливості, залишаючись у рамках розумних обмежень на обчислювальні ресурси.

Список використаної літератури

1. *Анисимов, А. В., Марченко, А. А.* Система обработки текстов на естественном языке [Текст] / А. В. Анисимов, А. А. Марченко // «Искусственный интеллект» НАНУ и ІПШІ-2002. – Вып. 4.

2. *Анисимов, А. В., Марченко, А. А.* Алгоритмы ассоциативного реферирования естественно-языковых текстов [Текст] / А. В. Анисимов, А. А. Марченко // «Искусственный интеллект» НАНУ и ІПШІ-2006. – Вып. 3.

3. *Ермаков, А. Е., Плешко, В. В.* Тематическая навигация в полнотекстовых базах данных [Текст] / А. Е. Ермаков, В. В. Плешко // Мир ПК. – 2001. – Вып. № 8.

4. *Леонтьева, Н. Н.* К теории автоматического понимания естественных текстов. Ч. 3 [Текст] / Н. Н. Леонтьева : Семантический компонент. Локальный семантический анализ. – М. : Изд-во МГУ, 2002.

5. *Тарануха, В. Ю., Анісімова, О. А., Романік, А. М.* Комп'ютерно-лінгвістичні технології обробки текстів на природній мові в заходах управління документообігом [Текст] / В. Ю. Тарануха, О. А. Анісімова, А. М. Романік : 11-та міжнар. конф. з автоматичного управління. – Київ, 2004.