3. Коваленко, Е.О. Машиностроение в контексте активизации глобализационных процессов в мировой экономике [Текст] / Е.О. Коваленко. // Проблемы развития внешнеэкономических связей и привлечения иностранных инвестиций: региональный аспект: сб. науч. тр. – Донецк: ДонНУ, 2009. –Ч.3. –с.1404 – 1409. http://archive.nbuv.gov.ua/portal/soc_gum/prvs/2009_3/1404.pdf.
4. Степанова, Е.В. О формировании кластерной модели развития подъемно-транспортного машиностроения в Украине [Текст] / Е.В. Степанова. // Економічні інновації. - 2010. – Вип. 41. – С.260 – 266. http://archive.nbuv.gov.ua/portal/Soc_Gum/Ei/2010_41/PDFFiles/31_Step.pdf.
5. Шпилевский, В.В. Проблемы инновационного развития машиностроения Украины [Текст] / В. В. Шпилевский, А. Д. Олейник. // Бизнес информ. -2010. -№11. –с.186 – 189. http://archive.nbuv.gov.ua/portal/soc_gum/bi/2010_11/186-189.pdf.
6. Рубченко, М. Три полезных буквы – ЭКА. / М. Рубченко. [электронный ресурс]. // http://expert.ru/expert/2005/19/19extishk1_39036/.
7. Продукты торгового финансирования. УкрЭксимБанк. Официальный сайт. // http://www.eximb.com/rus/corporate/trade/scheme/.
8. Hiroyuki O. Technology and industrial development in Japan: building capabilities by learning, innovation and public policy. http://ideas.repec.org/b/oxp/obooks/9780198288022.html.
9. Cimoli M. Industrial policy and development: the economy of capabilities accumulation. http://econpapers.repec.org/bookchap/oxpobooks/9780199235278.htm.
10. Braunerhjelm P.B. Cluster genesis: technology – based industrial development. http://econpapers.repec.org/bookchap/oxpobooks/9780199232208.htm
11. Henderson V. Externalities and industrial development. http://www.sciencedirect.com/science/article/pii/S0094119097920362.
12. Venables A.J. Trade policy, cumulative causation and industrial development. http://www.sciencedirect.com/science/article/pii/0304387895000585.
13. Dahlman C., Westphal L. Technological effort in industrial development. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2005/ 10/14/000178830_98101903361590/Rendered/PDF/REP263000.pdf.

Стаття присвячена питанням машинного розуміння текстів на природній мові. При цьому були розглянуті особливості проектування та практичної реалізації синтаксичного текстового аналізатору російської мови та дано короткі відомості про методи, покладені в основу його функціонування. До особливостей реалізації можна віднести використання граматики залежностей для визначення категорій елементів мовних структур

Ключові слова: розуміння текстів на природній мові, первинний аналіз природної мови

Статья посвящена вопросам машинного понимания текстов на естественном языке. При этом были рассмотрены особенности проектирования и практической реализации синтаксического текстового анализатора русского языка и даны краткие сведения о методах, положенные в основу его функционирования. К особенностям реализации можно отнести использование грамматики зависимостей для определения категорий элементов языковых структур

Ключевые слова: понимание текстов на естественном языке, первичный анализ естественного языка

УДК 004.89, 004.048, 004.912

СИНТАКСИЧНИЙ АНАЛІЗАТОР ЯК ЗАСІБ МАШИННОГО РОЗУМІННЯ ПРИРОДНОЇ МОВИ

І. А. Жирякова
Кандидат технічних наук, доцент*
E-mail: irena_zh@ukr.net
М. С. Симоненко*
E-mail: mikesimons@mail.ru
*Кафедра інтелектуальних систем прийняття рішень
Черкаський національний університет ім. Б. Хмельницького
бул. Шевченка, 81, м. Черкаси, Україна, 18031

1. Вступ

На сьогоднішній день машинне розуміння природної мови є однією з найбільш актуальних задач в галузі комп'ютерних наук, рішення якої дозволило б досягти високого рівня формалізації мовних структур у різноманітних прикладних цілях. Крім традиційної галузі застосування результатів син-

таксичного аналізу, таких як машинний переклад та генерація текстів на природній мові [1], модулі синтаксичного аналізу активно використовуються в системах автоматичного аналізу контенту при моніторинзі блогів і новин в мережі Internet; аналізі вихідного коду мов програмування в процесі трансляції, компіляції або інтерпретації [2, 3]; аналізі коректності побудови та оптимізації математичних та хімічних виразів тощо.

2. Аналіз останніх досліджень і публікацій та постановка проблеми

Сьогодні в зарубіжній літературі [4-6] домінуючою парадигмою в синтаксичному аналізі є розбір керований даними (data-driven parsing). Це обумовлено доступністю вивірених корпусів дерев синтаксичного розбору для мов з жорстким порядком слів, таких як англійська. Тому, на ринку програмного забезпечення домінують синтаксичні аналізатори (парсери) на основі формальних граматик – явно заданих правил синтезу текстів.

Для мов із м'яким порядком слів, слов'янських, застосування цього підходу викликає складність, пов'язану з отриманням громіздких результуючих правил, і, як наслідок, високої трудомісткості розробки і підтримки. Сучасний стан програмних розробок для деяких слов'янських мов, таких як українська та російська, можна оцінити на щорічній міжнародній конференції «Діалог». Згідно [7], існуючі розробки містять ряд значних розбіжностей за принципами встановлення зв'язків при синтаксичному розборі і, як наслідок, не мають єдиного рішення для представлення вихідних даних.

Тому, враховуючи все вище зазначене, актуальним питанням є проектування та розробка парсерів для мов із м'яким порядком слів, зокрема російської мови, орієнтована на використання сучасних підходів та парадигм в синтаксичному аналізі.

3. Результати

Під метою синтаксичного аналізу у цій статті будемо розуміти виділення базових синтаксичних структур і встановлення синтаксичних зв'язків між ними. Тобто, розроблюваний програмний засіб у якості вхідних данних буде отримувати довільний текст, вихідні дані будуть представляти собою його синтаксичну структуру у вигляді дерева залежностей.

На рис. 1 у вигляді UML-діаграми прецедентів представлено зовнішнє оточення парсеру, що характеризує його функціональне призначення відносно виділених діючих осіб: спеціаліста-мовознавця та користувача системи, кожен з яких відіграє ключову роль на певному етапі життєвого циклу системи (перший – на етапі розробки системи, другий – на етапі експлуатації).

Розглянемо більш докладно принципи роботи парсеру. Теоретичною основою синтаксичного аналізу є граматика членів речення. Отже, парсер має містити словник для аналіз граматичних конструкцій та граматичних атрибутів (частин мови, відмінку, числа, роду та інші), а також словники основ та зворотів.

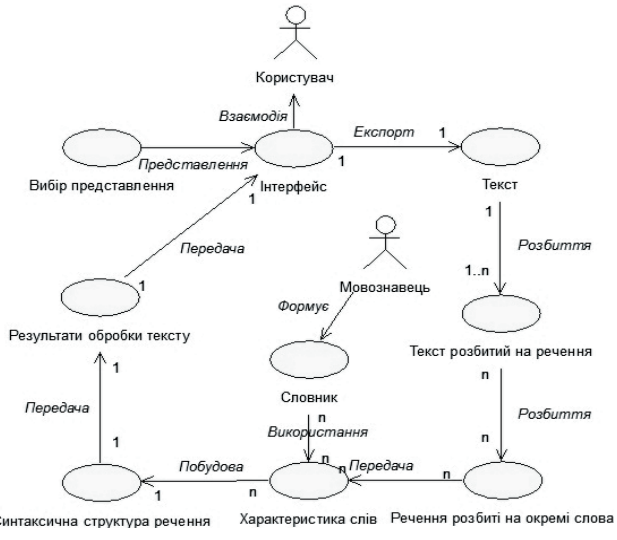

Рис. 1. Функціональне призначення парсеру

Головне завдання синтаксичного аналізу – побудувати всі зв'язки між вузлами в реченні на основі заданого аналізатора. Аналізатор не буде використовувати ніякої семантичної інформації, тобто зв'язки між вузлами матимуть лише функціонально-синтаксичний характер. Для визначення граматичних категорій для кожного слова в аналізаторі буде використовуватись граматика залежностей [8, 9], яка дозволяє встановити від якого слова залежить кожне слово в реченні і тип цих зв'язків не враховуючи порядок слів.

Роботі аналізатора передує процедура розбиття тексту на речення (сегментація) і розбиття складних речень на прості і окремі слова (токенізація), тобто побудова мінімальних одиниць синтаксичної структури. Її виконання проводиться за допомогою введення в синтаксичну структуру нетермінального вузла символу речення.

При цьому підрядні та сурядні сполучники стають частиною граматичної характеристики нетермінальних вузлів.

Упорядкованість повідомлень між зазначеними структурними одиницями системи представимо з допомогою UML-діаграми послідовностей (рис. 2).

В результаті вдалого синтаксичного розбору речення згортається в зв'язне дерево з однією єдиною кореневою вершиною. Але, оскільки одна словоформа може відповідати декільком граматичним формам слова, в тому числі і формам різних слів, в ході аналізу потрібно застосовувати побудову дерева для всіх можливих варіантів граматичних форм. В програмній реалізації дана проблема вирішується перебором всіх граматичних форм та їх синтаксичних дерев. Граматичні форми, які забезпечують максимальну згортку дерева, будемо вважати найбільш достовірними.

Розглянемо особливості логіко-лінгвістичної моделі текстової інформації, яка використовується парсером.

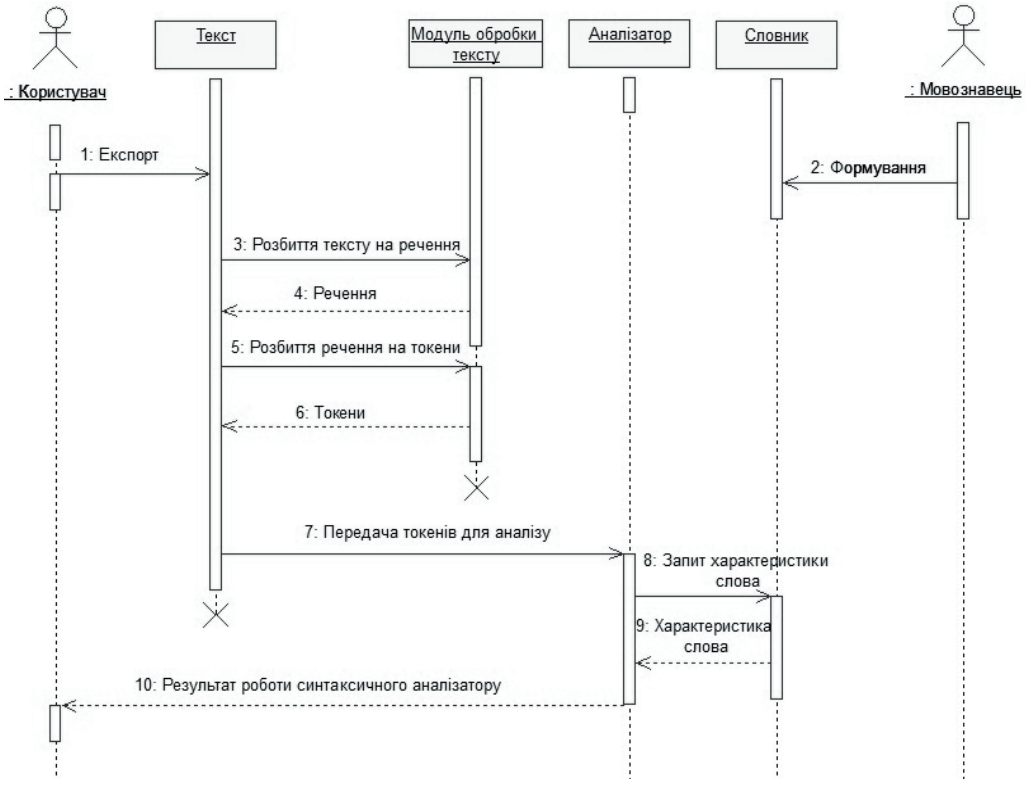

Рис. 2. Взаємодія об'єктів парсеру

Просте речення у формалізмі логіки предикатів – це атомарний предикат; складному реченню зіставляється складне логічне висловлювання, яке є сукупністю атомарних предикатів, поєднаних логічними зв'язками.

Нехай кожне речення S складається з множини слів M та множини простих речень $R_v(S)$. Тоді загальна форма логіко-лінгвістичної моделі набуває вигляду:

$$S=\left\{M,R_v(S)\middle|M\subset R_v(S),v=\overline{1,n}\right\}$$
$$\forall S(B_v\,\&\,C_v)\vee(B_v\to C_v)\vee(B_v\vee C_v)\vee(B_v\sim C_v)\vee A_v, \quad (1)$$
$$v=\overline{1,n}\Leftrightarrow R_v(S),v=\overline{2,n}.$$

Згідно (1), A_v – просте логічне висловлювання, яке описує просте речення. B_v і C_v – складні логічні висловлювання, які описують частину складного речення, що складається з n-тої кількості простих речень, і може набувати вигляду (1), якщо множина простих речень $R_v(S)$ містить більше двох елементів.

Якщо $R_v(S)$ містить два елементи, то вирази B_v і C_v представляють собою атомарні предикати.

Основними складовими моделі (1) є концептуальні відношення, які можуть зустрітися в текстовій інформації, і є відображенням синтаксичної структури будь-якого речення природної мови, а саме:

– $(B_v\,\&\,C_v)$ – опис частини складного речення складові якого рівноправні за змістом;

– $(B_v\to C_v)$ – опис частини речення, в якому залежна частина C_v може уточнювати час, місце, причину, спосіб, про який йдеться в головній частині складнопідрядного речення B_v;

– $(B_v\vee C_v)$ – опис частини складного речення складові якого протиставляються або зіставляються;

– $(B_v\sim C_v)$ – опис частини складного речення складові якого рівнозначні за змістом, тобто тотожні.

Парсер реалізовано за допомогою мови логічного програмування SWI-Prolog.

Словник та правила, які до нього входять, реалізовано у вигляді Prolog бази знань.

Дерево синтаксичного аналізу представляється за допомогою Prolog терма, функтором якого є корінь дерева, а параметри – гілками (піддеревами) дерева [10].

На рис. 3 представлено головне вікно текстового редактору з вбудованим парсером.

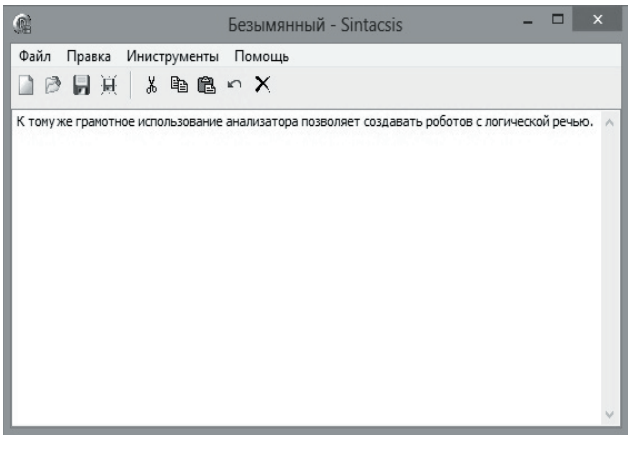

Рис. 3. Головне вікно програми

Окрім проведення синтаксичного аналізу створений програмний продукт дозволяє створювати текстові документи та виконувати команди редагування властиві звичайному текстовому редактору, а також обирати представлення результатів синтаксичного аналізу: псевдографічне, з допомогою графа залежностей, у вигляді семантичного дерева понять, у вигляді xml-файла.

На рис. 4 представлено результат синтаксичного розбору речення.

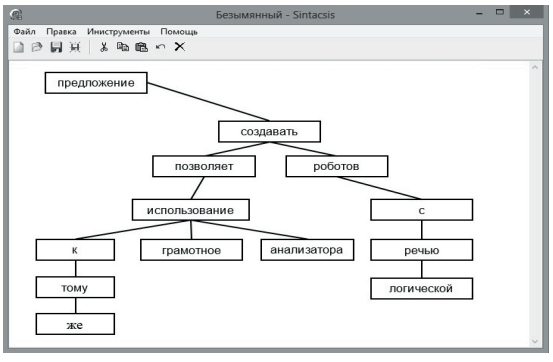

Рис. 4. Результат роботи парсеру

4. Висновки

Як показала практика, незважаючи на обмеженість синтаксичних аналізаторів тексту, працюючих без використання семантики, існують всі підстави стверджувати, що якість їх роботи в рамках обмеженого використання обчислювальних ресурсів є досить задовільною.

В подальшому запропонована розробка буде удосконалена шляхом розширення її функціональності в наслідок застосування семантичного аналізу та використання модуля роботи з іншими природними мовами.

Літуратура

1. Марчук, Ю.Н. Компьютерная лингвистика [Текст] / Ю.Н. Марчук. – М.: Изд-во АСТ, 2007. – 320 с.
2. Компиляторы. Принципы, технологии и инструментарий / [Альфред В. Ахо, Моника С. Лам, Рави Сети, Джеффри Д. Ульман]; пер. с англ. И. Красиков. – М.: Издательство «Вильямс», 2008. – 1184 с.
3. Foster, J.M. Automatic Syntactic Analysis [Текст] / J.M. Foster; general ed. Stanley Gill. – New York: MacDonald, London and American Elsevier Inc., 1970. – 70 p.
4. Dependency Parsing: [Synthesis Lectures on Human Language Technologies] / [Sandra Kubler, Ryan McDonald, Joakim Nivre]; ser. ed. Graeme Hirst. – Morgan & Claypool Publishers, 2009. – 115 p.
5. D. Grune Parsing Techniques – A Practical Guide [Текст] / D. Grune, Ceriel J.H. Jacobs. – [2-ond ed.]. – Amsterdam: Springer, 2008. – 662 p.
6. David R. Dowty Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives [Текст] / David R. Dowty, Lauri Karttunen, Arnold M. Zwicky. – Cambridge University Press, 2005. – 428 p.
7. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка / [Толдова С., Соколова Е., Астафьева И. и др.] // Компьютерная лингвистика и интеллектуальные технологии. – 2012. – Вып. 11 (18). – С. 797–810.
8. Encyclopedia of Linguistics / ed. Philipp Strazny. – [2 vols.]. – New York, Oxon: Fitzroy Dearborn, 2005. – 1304 p.
9. Тестелец, Я. Г. Введение в общий синтаксис [Текст] / Я. Г. Тестелец. – М.: РГГУ, 2001. – 798 с.
10. Братко, И. Алгоритмы искусственного интеллекта на языке PROLOG / И. Братко; пер. с англ. – [3-е изд.]. – М.: Издательский дом “Вильямс”, 2004. – 640 с.