Актуальність дослідження програм семантичного аналізу текстів та огляд методів їх реалізації

Аркадій Анатолійович Зеленський

Криворізький національний університет,
вул. Віталія Матусевича, 11, м. Кривий Ріг, 50027, Україна
arkasha9606@gmail.com

Анотація. Одним із найголовніших завдань прикладної лінгвістики є розв'язання проблеми якісної автоматизованої обробки природної мови. Найпопулярнішими методами обробки природномовних текстових відповідей з метою екстракції та репрезентації семантики мають бути системи, які ґрунтуються на ефективному поєднанні лінгвістичних технологій аналізу та методів аналізу. Серед існуючих методів аналізу текстових даних чинне місце посідає метод з використанням векторної моделі. Іншим ефективним і актуальним засобом вилучення семантики із тексту та її представлення є метод латентно-семантичного аналізу (ЛСА). Метод ЛСА пройшов випробування та підтвердив свою ефективність у таких напрямках обробки природної мови, як моделювання концептуальних знань людини; інформаційний пошук, при реалізації якого ЛСА показує набагато кращі результати порівняно із звичайними векторними методами.

Ключові слова: семантика, прикладна лінгвістика, обробка текстів.

Relevance of research of programs for semantic analysis of texts and review of methods of their realization

Arkadii A. Zelenskyi

Kryvyi Rih National University, 11, Vitalii Matusevych St., Kryvyi Rih, 50027, Ukraine
arkasha9606@gmail.com

Abstract. One of the main tasks of applied linguistics is the solution of the problem of high-quality automated processing of natural language. The most popular methods for processing natural-language text responses for the purpose of extraction and representation of semantics should be systems that are based on the efficient combination of linguistic analysis technologies and analysis methods. Among the existing methods for analyzing text data, a valid method is

used by the method using a vector model. Another effective and relevant means of extracting semantics from the text and its representation is the method of latent semantic analysis (LSA). The LSA method was tested and confirmed its effectiveness in such areas of processing the native language as modeling the conceptual knowledge of the person; information search, the implementation of which LSA shows much better results than conventional vector methods.

Keywords: semantics, applied linguistics, text processing.

Мова – це найважливіший засіб спілкування і пізнання, мова є продовженням думки людини. Це засіб спілкування, який поки що властивий тільки людям.

Програмні продукти ще в другій половині – кінці XX століття почали проникати в широкі маси. Здебільшого це були розважальні програми та відеоігри. А з розвитком інформаційних технологій та телекомунікацій життя становиться все більш та більш мобільним та інформативним [1; 2]. Нові технології міцно входять в різні сфери життя та вносять в них нові стандарти.

Проблема механізмів синтаксису, що раніше відносилася виключно до «чистої науки», стала за останні роки «гостро прикладною». Такі пріоритетні області науки і техніки, як штучний інтелект [3; 4], машинний переклад, створення системи спілкування «людина – Інтернет» тощо потребують від учених максимально конкретних визначень механізмів їх дії на життєвий процес людини.

Інтернет подарував людству надзвичайно великі можливості у спілкуванні, навчанні та праці. Число користувачів мережі Інтернет зростає неймовірними кроками.

Люди щоденно генерують великі обсяги інформації в Інтернет: вони обмінюються повідомленнями, залишають відгуки, оформлюють замовлення, виконують пошук, надсилають листи та ін. Нещодавно для українців стала доступна можливість подачі замовлення на оформлення громадянського паспорту онлайн. 93 % населення України є активними Інтернет-користувачами станом на 30 червня 2017 року.

Компанія Facebook опублікувала звіт, з якого стало відомо, що протягом вересня 2018 року соціальну мережу відвідувало близько 1,49 мільярда активних користувачів щоденно, та загальна кількість користувачів за місяць становила 2,27 мільярда.

Зараз багато компаній розширює свої можливості за рахунок різних ботів, які дозволяють здійснювати найрізноманітніші операції – купувати каву, замовляти піцу, вести моніторинг залізничних квитків, а потім і придбати їх. Ботів лише почали активно впроваджувати в наше життя, в майбутньому з’явиться ще більше ботів різного призначення, і тому питання обробки тексту зараз є актуальним, як ніколи.

Прикладна лінгвістика – розділ мовознавства, який практично розв’язує питання, пов’язані з вивченням мови. З давніх давен люди вели розмову між собою, але завдяки аналізу текстів в людини може з’явитися новий співрозмовник.

Мова є продовженням думки людини. Розвиток в сфері аналізу тексту дозволить здійснити значний прорив у розробці людино-машинної взаємодії. Використовуючи звичний для людини діалог, вона зможе оперувати складними комп'ютерними операціями. Окрім цього, система, яка буде здатна приймати голосові запити користувача у вигляді звичайного тексту та адекватно на них реагувати, відкриє нові можливості в освоєнні комп'ютерних технологій для людей з обмеженими можливостями.

У XXI столітті змінився механізм передачі і прийняття інформації. Мова, що стала штучною, є наразі основою засобів Інтернет-спілкування. Її лексика насичує інформаційний простір, в цій мові не дотримано ні відповідності правилам, ні допустимого відсотку в тексті слів іншомовного походження, тому проблема семантичного аналізу Інтернет-повідомлень є достатньо складною задачею.

Предметом вивчення семантики є значна кількість мовних об'єктів. До них належить, наприклад, семантична структура слова, семантика мови і мовлення (висловлення), семантика сприймання висловлення, семантика граматичної системи мови та ін.

Прикладна лінгвістика використовується для вирішення низки актуальних питань сучасності таких, як інтелектуальний помічник, що спілкується з людиною, відповідає на питання та виконує команди; аналіз повідомлень, що визначає важливість повідомлення або виявляє інтереси користувача, як частина рекомендаційної системи; аналіз психологічного типу особистості, що ґрунтується на текстах, що набирає користувач та їх емоційному забарвленню; визначення області, до якої відноситься текст, а також пропонування доречної відповіді на проаналізований текст. Наступним етапом прикладної лінгвістики може стати робота з аудіо форматом тексту, де є можливість визначати інтонації та на їх основі виявляти важливі думки в тексті.

Лінгвістична обробка природномовних текстів є однією з центральних проблем інтелектуалізації інформаційних технологій, якій приділяється значна увага в розвинутих країнах Європи, США та Японії, свідченням чого є великі капіталовкладення у розроблення лінгвістичного програмного забезпечення та цілісних лінгвістичних технологій. Велику кількість науково-дослідних програм спрямовано на розвиток лінгвістичних інформаційних систем. У зв'язку з бурхливим розвитком Інтернет ця проблема набуває ще більшої значущості.

Аналіз одиниць лексичного рівня, тобто слів, є одним з найважливіших. До складу цього різновиду аналізу можуть входити різні словники, які містять лексичний репертуар та морфемну структуру лексичних одиниць, словозмінні парадигми тощо. Застосовують і так звані безсловникові методи, які передбачають здійснення розбиття слова на морфеми за заданими алгоритмами, забезпечення віднесення слів до словозмінних парадигматичних класів, виконання процедури лематизації (редукції текстових словоформ до початкових, вихідних форм тощо). На цьому ж етапі частково здійснюється і граматичний аналіз, тобто ідентифікація лексико-граматичних класів та значень граматичних категорій текстових слів. Зрозуміло, що для української мови надзвичайно

важливою є наявність словозмінної парадигми та відповідної словозмінної класифікації.

Роботу в цій сфері можна розподілити на декілька основних етапів. Першим буде очищення вхідного тексту від слів, які не несуть змістовного навантаження, та виокремлення основи слова для більш гнучкого їх порівняння.

Наступним етапом буде виділення ключових слів вхідного тексту. Третім етапом може стати виявлення області, до якої відноситься текст.

Одним з найголовніших завдань прикладної лінгвістики є розв'язання проблеми якісної автоматизованої обробки природної мови. Вирішення такої проблеми може використовуватись для автоматизації перевірки відповіді студента, поданої у текстовому форматі природною мовою. Обробка природної мови буде основою розробки ефективної методики порівняння відповіді студента зі зразками правильної відповіді на запитання.

На думку А. В. Анісімова [5; 6], О. Є. Литвиненка [7] та ін., вищезгадані етапи лінгвістичного аналізу природномовних текстів людиною не є ізольованими один від одного, вони є паралельними процедурами сприйняття змісту тексту. Отже й комп'ютерні моделі семантичного, синтаксичного і морфологічного аналізу природномовних текстів доцільно розробляти як систему паралельних розподілених процесів з заданим відношенням підпорядкування.

Найпопулярнішими методами обробки природномовних текстових відповідей з метою екстракції та репрезентації семантики мають бути системи, які ґрунтуються на ефективному поєднанні лінгвістичних технологій аналізу (морфологічного, синтаксичного, семантичного) та методів аналізу, які дозволили б виявляти приховані асоціативні залежності всередині природномовних текстів.

Серед існуючих методів аналізу текстових даних чинне місце посідає метод з використанням векторної моделі. Основна ідея якого полягає в алгебраїчному представленні колекції текстових масивів векторами одного спільного для всієї колекції векторного простору. Векторна модель є основою для вирішення багатьох завдань, таких як: пошук документа за запитом, класифікація документів та кластеризація документів.

Іншим ефективним і актуальним засобом вилучення семантики із тексту та її подання є метод латентно-семантичного аналізу (ЛСА) [8].

Д. В. Ланде розглянув метод ЛСА в рамках технологій глибинного аналізу текстової інформації Text Mining, що розроблені на основі статистичного та лінгвістичного аналізу, і дозволяють не лише здійснювати відбір релевантних документів, а й виділяти їх семантику, яка досить часто буває прихованою [9].

Метод латентно-семантичного аналізу є теорією і методом екстракції і представлення контекстно-залежного змісту слів шляхом статистичної обробки великого корпусу текстів. Головною ідеєю методу є те, що сукупність усіх контекстів, в яких певне слово вживається або, навпаки, не вживається, обумовлює набір обмежень, які визначають подібність значень слів або множини слів. Отже, простежується думка, що між словами і контекстом, в якому вони вживаються, існують приховані (латентні) зв'язки. Метод ЛСА дозволяє

визначити асоціативну і семантичну близькість та вирахувати кореляції між двома термами, двома документами, або між термом і документом.

Ефективність застосування методу ЛСА в сфері знань людини підтверджена різноманітними прикладами його роботи. Зокрема, вперше зазначений метод був застосований з метою автоматичного індексування текстів та виявлення їх асоціативно-семантичної структури. Використання методу ЛСА знайшло своє відображення у системах вилучення, представлення семантичної інформації з тексту.

Метод ЛСА пройшов випробування та підтвердив свою ефективність у таких напрямках обробки природної мови як моделювання концептуальних знань людини; інформаційний пошук, при реалізації якого ЛСА показує набагато кращі результати порівняно зі звичайними векторними методами; процес підбору синонімів, якість якого була перевірена і підтверджена шляхом обробки тестів TOEFL; результат роботи методу ЛСА виявився ідентичним середнім показникам відбору синонімів людьми із неангломовних країн.

За результатами аналізу роботи методу ЛСА для вирішення вищезгаданих та інших завдань, можемо зробити висновок, що розглянутий метод є найкращим засобом для виявлення та представлення прихованих семантичних характеристик окремих слів та текстів загалом.

Отже, доходимо висновку, що найпопулярнішими методами обробки природномовних текстових відповідей мають бути системи, що ґрунтуються на ефективному поєднанні лінгвістичних технологій аналізу та методу ЛСА, що дозволить виявляти приховані асоціативні залежності всередині природномовних текстів. Таке поєднання дає можливість врахувати та частково ліквідувати недоліки, що притаманні обом методам, та удосконалити процес обробки природномовної відповіді шляхом комбінування переваг розглянутих лінгвістичних та статистичних методів.

Список використаних джерел

1. Семеріков С. О. Мобільне навчання: історія, теорія, методика / Сергій Семеріков, Ілля Теплицький, Світлана Шокалюк // Інформатика та інформаційні технології в навчальних закладах. – 2008. – № 6. – С. 72-82 ; 2009. – № 1. – С. 96-104.
2. Семеріков С. О. Мобільне навчання : історико-технологічний вимір / Семеріков С. О., Стрюк М. І., Моісеєнко Н. В. // Теорія і практика організації самостійної роботи студентів вищих навчальних закладів : монографія / кол. авторів; за ред. проф. О. А. Коновала. – Кривий Ріг : Книжкове видавництво Киреєвського, 2012. – С. 188-242.
3. Semerikov S. O. Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot [Electronic resource] / Serhiy O. Semerikov, Illia O. Teplytskyi, Yuliia V. Yechkalo, Arnold E. Kiv // Augmented Reality in Education : Proceedings of the 1st International Workshop (AREdu 2018). Kryvyi Rih, Ukraine, October 2, 2018 / Edited by : Arnold E. Kiv, Vladimir N. Soloviev. – P. 122-147. – (CEUR Workshop Proceedings (CEUR-WS.org), Vol. 2257). – Access mode : http://ceur-ws.org/Vol-2257/paper14.pdf.
4. Markova O. M. CoCalc as a Learning Tool for Neural Network Simulation in the Special Course “Foundations of Mathematic Informatics” [Electronic resource] / Oksana Markova,

Serhiy Semerikov, Maiia Popel // ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer 2018 : Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops (ICTERI, 2018). Kyiv, Ukraine, May 14-17, 2018 / Edited by : Vadim Ermolayev, Mari Carmen Suárez-Figueroa, Vitaliy Yakovyna, Vyacheslav Kharchenko, Vitaliy Kobets, Hennadiy Kravtsov, Vladimir Peschanenko, Yaroslav Prytula, Mykola Nikitchenko, Aleksander Spivakovsky. – P. 388-403. – (CEUR Workshop Proceedings (CEUR-WS.org), Vol. 2104). – Access mode : http://ceur-ws.org/Vol-2104/paper_204.pdf.

5. Анисимов А. В. Система обработки текстов на естественном языке / А. В. Анисимов, А. А. Марченко // Искусственный интеллект. – 2002. – № 4. – С. 157-163.
6. Анисимов А. В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык / А. В. Анисимов. – К. : Наук. думка, 1991. – 208 с.
7. Литвиненко О. Є. Інженерно-лінгвістичні принципи аналізу текстів / О. Є. Литвиненко, Д. А. Бурко // Наукоємні технології. – 2009. – Том 3, № 3. – С. 60-62. – DOI : 10.18372/2310-5461.3.5130
8. Заболеева-Зотова А. В. Латентный семантический анализ: новые решения в Internet / А. В. Заболеева-Зотова, А. Ю. Пастухов, П. В. Сердюков, Н. А. Козлова, С. А. Чернов // Информационные технологии. – 2001. – № 6. – С. 67-82.
9. Ландэ Д. В. Поиск знаний в Internet / Д. В. Ландэ. – М. : Вильямс, 2005. – 272 с. – (Профессиональная работа)

References (translated and transliterated)

1. Semerikov, S., Teplytskyi, I., Shokaliuk, S.: Mobilne navchannia: istoriia, teoriia, metodyka (Mobile learning: history, theory, methods). Informatyka ta informatsiini tekhnolohii v navchalnykh zakladakh. 6, 72–82 (2008), 1, 96–104 (2009)
2. Semerikov, S.O., Striuk, M.I., Moiseienko, N.V. Mobilne navchannia: istoryko-tekhnolohichnyi vymir (Mobile learning: historical and technological dimension). In: Konoval, O.A. (ed.) Teoriia i praktyka orhanizatsii samostiinoi roboty studentiv vyshchykh navchalnykh zakladiv, pp. 188-242. Knyzhkove vydavnytstvo Kyreievskoho, Kryvyi Rih (2012)
3. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Yu.V., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot. In: Kiv, A.E., Soloviev, V.N. (eds.) Proceedings of the 1st International Workshop on Augmented Reality in Education (AREdu 2018), Kryvyi Rih, Ukraine, October 2, 2018. CEUR Workshop Proceedings, vol. 2257, pp. 122–147. http://ceur-ws.org/Vol-2257/paper14.pdf. Accessed 30 Nov 2018
4. Markova, O., Semerikov, S., Popel, M.: CoCalc as a Learning Tool for Neural Network Simulation in the Special Course “Foundations of Mathematic Informatics”. In: Ermolayev, V., Suárez-Figueroa, M.C., Yakovyna, V., Kharchenko, V., Kobets, V., Kravtsov, H., Peschanenko, V., Prytula, Y., Nikitchenko, M., Spivakovsky, A. (eds.) Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops (ICTERI, 2018), Kyiv, Ukraine, May 14-17, 2018. CEUR Workshop Proceedings, vol. 2104, pp. 388–403. http://ceur-ws.org/Vol-2104/paper_204.pdf. Accessed 25 Aug 2018
5. Anisimov, A.V., Marchenko, A.A.: Sistema obrabotki tekstov na estestvennom iazyke (Natural Language Processing System). Iskusstvennyi intellekt. 4, 157–163 (2002).

6. Anisimov, A.V.: Kompiuternaia lingvistika dlia vsekh: Mify. Algoritmy. Iazyk (Computational linguistics for all: Myths. Algorithms. Language). Naukova dumka, Kiev (1991)
7. Lytvynenko, O.Ye., Burko, D.A.: Inzhenerno-linhvistychni pryntsypy analizu tekstiv (Engineering-linguistic principles of the analysis of texts). Naukoiemni tekhnolohii. **3**(3), 60–62 (2009). doi:10.18372/2310-5461.3.5130
8. Zaboleeva-Zotova, A.V., Pastukhov, A.Iu., Serdiukov, P.V., Kozlova, N.A., Chernov, S. A.: Latentnyi semanticheskii analiz: novye resheniia v Internet (Latent semantic analysis: new solutions in the Internet). Informatcionnye tekhnologii. 6, 67–82 (2001)
9. Lande, D.V.: Poisk znanii v Internet (Search for knowledge on the Internet). Viliams, Moscow (2005)