УДК 81'322.2:004.89

ОКСАНА КОМАРНИЦЬКА,
кандидат філологічних наук
(м.Хмельницький)

Методи автоматизованого семантичного аналізу природномовної інформації

У статті розглянуто особливості автоматизованого семантичного аналізу тексту, досліджено проблеми створення автоматизованих лінгвістично-програмних засобів, придатних для застосування в системах екстракції семантики з тексту. Автором систематизовано та здійснено порівняльний аналіз результатів наукових досліджень у галузі розробки моделей і методів семантичного аналізу природномовної інформації. Визначено два основні підходи в напрямку комп'ютерної обробки природномовних текстів: лінгвоаналітичний і статистичний. Автором аргументовано, що найбільш перспективними та ефективними з них є, відповідно, експліцитні методи семантичного аналізу текстової інформації (алгоритми онтологічного семантичного аналізу) та методи латентно-семантичного аналізу. Окреслено можливі шляхи удосконалення існуючих комп'ютерних засобів діагностування релевантності природномовної інформації; обгрунтовано, що найпопулярнішими методами обробки природномовної інформації з метою екстракції та репрезентації семантики мають бути системи, що ґрунтуються на ефективному поєднанні лінгвістичних технологій аналізу (графематичного, морфологічного, синтаксичного, семантичного), зокрема із застосуванням онтологій, та методу латентно-семантичного аналізу. Доведено, що інтеграція технологій експліцитного семантичного аналізу, латентно-семантичного аналізу, методів теорії нечіткої логіки, штучного інтелекту та ін. є перспективним шляхом розв'язання проблеми автоматизованого семантичного аналізу природномовної інформації.

Ключові слова: *автоматизований семантичний аналіз, природномовна інформація, лінгвістичний аналіз, латентно-семантичний аналіз, метод, онтологія.*

Постановка проблеми в загальному вигляді... Одним з найголовніших завдань прикладної лінгвістики є розв'язання проблеми якісної автоматизованої обробки природної мови. Це вимагає залучення цілої низки наукових дисциплін та їхніх методів, зокрема,

методів комп'ютерної, когнітивної, математичної лінгвістики, теорії штучного інтелекту, семантичних мереж, нейрокібернетики, логіки тощо. Комплексне застосування засобів, моделей і методів цих наукових галузей у принципі дозволяє створити ефективний інструментарій опрацювання природномовної текстової інформації. Відомо, що процес лінгвістичного аналізу тексту включає певну послідовність дій, а саме: графематичний аналіз, морфологічний аналіз, синтаксичний аналіз, семантичний аналіз. Найбільш складним та важливим етапом обробки природномовної інформації є саме семантичний аналіз, отже на перший план висувається семантична складова лінгвістичного аналізу та екстракція знань із текстової інформації. Отже, розглянемо сучасний рівень розвитку моделей та методів семантичного аналізу природномовних текстів, щоби запропонувати власну методологію побудови таких моделей.

Аналіз досліджень і публікацій... Дослідження з питань обробки природної мови розпочались ще в кінці 40-х років 20 ст., і з того часу пройшли чотири стадії свого розвитку, які характеризуються, відповідно, акцентуванням уваги на створенні систем машинного перекладу, домінуванням теорій штучного інтелекту, запровадженням логіко-граматичного стилю і застосуванням статистичного підходу [1, с. 53]. Після понад 60-ти років розвідок у галузі автоматичного опрацювання природномовної інформації це питання залишається актуальним і в наш час, коли спостерігається бум у розвитку інформаційних технологій, що у свою чергу стимулює необхідність створення високоефективних лінгвістичних технологій комп'ютерного аналізу тексту.

Слід, однак, визнати, що на сьогодні не існує моделей та засобів, які б достатньою мірою враховували особливості природної мови в процесі інтелектуального опрацювання текстової інформації. Це пояснюється труднощами, що виникають під час формального опису системи природної мови, обумовленими її сутністю, оскільки особливістю природної мови є її принципова нечіткість, проблему якої досліджували автори А. Н. Аверкин, Лотфи Аскер Заде, А. П. Рижов, М. Беррі. Свідомість людини здатна сприймати нечіткі судження та з контексту робити цілком певні висновки про змісти, актуалізовані в природномовних конструкціях. Але машина здатна сприймати лише те, що чітко задано в описах відповідних моделей. Багатозначність та непрогнозованість контекстної семантики мовних конструкцій не просто знижує якість роботи систем автоматичного опрацювання текстів, але й часто робить їх функціонування неможливим [2, с. 38].

Формулювання цілей статті… Метою статті є дослідження та аналіз існуючих моделей та методів автоматичного семантичного аналізу природномовної інформації, а також визначення можливих шляхів їх удосконалення.

Виклад основного матеріалу… На сучасному етапі розвитку науки і техніки в напрямку комп'ютерної обробки природномовних текстів виділяють два основні підходи: лінгвоаналітичний і статистичний. Найбільш перспективними та ефективними з них визнано, відповідно, експліцитні методи семантичного аналізу текстової інформації (алгоритми онтологічного семантичного аналізу) та методи латентно-семантичного аналізу. Саме ці методи дозволяють визначити та побудувати смислову структуру природномовного тексту у формалізованому вигляді (наприклад, у вигляді фрейму, семантичної мережі, онтологічного запису представлення запису тексту тощо) [3, с.169].

На думку Н. М. Леонтьєвої [4], автоматичне розуміння тексту, як необхідна складова різноманітних прикладних завдань, передбачає побудову таких типів семантичної структури:

- лінгвістичні структури речень тексту;
- семантичні мережі цілого тексту;
- інформаційні структури цілого тексту;
- структури баз даних та баз знань [6, с. 22].

Лінгвістичні структури речень тексту ґрунтуються на семантико-синтаксичному представленні у вигляді синтаксичних дерев речення із семантичними вузлами, прикладом такої роботи є багаторівнева динамічна модель «Смисл⇔Текст», розроблена І. О. Мельчуком, О. К. Жолковським та Ю. Д. Апресяном [5]. Основою роботи автомату на основі цієї моделі є побудова правильної синтаксичної структури, вузли якої замінюються на відповідні тлумачення зі словника, в результаті чого отримуємо семантичну структуру. Головними недоліками таких систем є обмеженість виходу за рамки речення, вибіркового сприйняття найважливішої інформації, недостатня кореляція із системами подання знань [6, с. 50-51].

Побудова семантичних мереж цілого тексту передбачає механізм розуміння тексту шляхом встановлення референтних зв'язків між реченнями та зв'язків «тема-рема» у межах самого речення.

Інформаційні структури цілого тексту реалізуються, переважним чином, в інформаційно-пошукових системах і ґрунтуються на опрацюванні різноманітних текстів за допомогою рубрикаторів, тезаурусів, класифікаторів, результатом роботи яких є отримання узагальненого розуміння тексту.

Структури баз даних та баз знань, як правило, застосовуються в системах штучного інтелекту і враховують екстралінгвістичні моделі, що ілюструють залежність розуміння тексту від попередніх знань про предмет [4, с. 26]. Такі структури відображують розуміння цілого тексту і передбачають розпізнавання певного сюжету в тексті на основі денотативного підходу [6, с. 52].

Саме на побудові бази знань предметної галузі ґрунтуються експліцитні методи аналізу природномовної інформації, зокрема на побудові основної її частини – онтології, що є чітко структурованою моделлю предметної галузі із систематизованим набором термінів, які описують можливі відношення між об'єктами предметної сфери [7, с. 571].

Побудова онтолого-керованих систем передбачає розробку теоретичних основ і методології проектування, що містять узагальнену архітектуру і структуру системи, формальну модель і методологію проектування онтології предметної області, формальну модель представлення знань, узагальнені алгоритми процедур обробки знань та ін. Кожна з перелічених складових загальної методології проектування являє собою складну інформаційно-алгоритмічну структуру [8, с. 64].

Особливої уваги потребує також знання-орієнтований підхід до аналізу природномовної текстовою інформації, запропонований Замаруєвою І. В. [9], який ґрунтується на побудові процедур автоматизації процесів екстракції, формалізації і логіко-семантичної обробки знань, що містяться у природно мовних текстах. З метою виділення з тексту основних компонентів знань і встановлення логіко-семантичних відношень між ними розроблено логіко-семантичну структуру змісту природномовного тексту, особливістю якої є гібридне представлення, яке об'єднує властивості семантичних мереж і предикатних моделей. Природномовний текст розглядається як сукупність трьох взаємопов'язаних систем: як знакова система, як граматична система та як система знань про світ і відповідно передбачає здійснення графемного (виділення фрагментів тексту, речень, синтагм, лексем) і лінгвістичного (морфологічного, синтаксичного, семантичного) аналізу [9].

Підсумовуючи сказане, доходимо висновку, що алгоритми онтологічного семантичного аналізу широко застосовують лінгвістичні бази знань у поєднанні із процедурами токенізації, лексико-морфологічного (бази знань морфології природної мови, словникові таблиці лексем частин мови), синтаксичного(бази знань синтаксису, таблиці граматик, «банки дерев») та семантичного аналізу (онтологічні

бази знань предметної галузі). Результатом роботи таких систем є отримання семантичної структури тексту, що складається із семантичних графів окремих текстів.

На нашу думку іншим ефективним і актуальним засобом екстракцію семантики із тексту та її представлення є метод латентно-семантичного аналізу.

Д. В. Ланде розглянув метод ЛСА в рамках технологій глибинного аналізу текстової інформації Text Mining, що розроблені на основі статистичного та лінгвістичного аналізу, а також методів ШІ і дозволяють не лише здійснювати відбір релевантних документів, а й виділяти і аналізувати їх семантику, яка досить часто буває прихованою [10, с. 160].

Метод латентно-семантичного аналізу (Latent Semantic Analysis – LSA) або латентно-семантичного індексування (Latent Semantic Indexing – LSI) 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 [11, с. 259]. 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 із тексту, побудови баз знань, у когнітивних моделях розуміння змісту та формування знань. А також широкого застосування метод ЛСА набув у системах навчання та оцінювання знань і дозволив ефективно оцінювати відповіді студентів, подані природною мовою у вигляді, есе, переказу, розгорнутої відповіді на запитання тощо.

ЛСА тісно пов'язаний із моделями нейронних мереж, але ґрунтується на дуже близькому до факторного аналізу методі сингулярного розкладання матриці (Singular Value Decomposition) терми-на-документи (TF-iDF) для відображення терм-векторів і документ-векторів у структурі асоціативних зв'язків меншої розмірності, шляхом її апроксимації до матриці меншого рангу з метою

позбавлення від так званих «шумів», що є присутніми при обробці великих корпусів документів.

Латентно-семантичний аналіз застосовує техніку частотного та імовірнісного аналізу для обробки текстів або текстових корпусів з метою побудови матриць сумісного вживання слів, що можуть інтерпретуватися як певні семантичні мережі у матричному вигляді [3, с. 170], але , на відміну від звичайних частотних методів, дозволяє виявляти глибині, приховані зв'язки і, таким чином, краще моделювати механізм сприйняття інформації людиною. Вважається, що два терми семантично схожі, якщо їх потрапляння у різні документи корелюються. І, відповідно, два документи семантично схожі або, якщо їх терми корелюються. Метод ЛСА демонструє значно кращі результати ніж за стандартна косинусна міра між вектор-документами, яка не враховує випадковість розподілу і семантичну схожість термів у документах [3, с. 172].

Основним підтвердженням цих положень стало використання методу ЛСА для отримання засобів вимірювання подібності значень слів з тексту. Результати показали, що: подібні значення отримані таким чином збігаються з вибором людини, рівень екстракції таких знань з тексту за допомогою методу ЛСА наближається до рівня людини, і ці досягнення значною мірою залежать від розмірності представлення. У цьому та інших випадках, ЛСА є потужним, у порівнянні з людськими стандартами, засобом коректної індукції знань [11, с. 266].

Метод ЛСА пройшов випробування та підтвердив свою ефективність у таких напрямках обробки природної мови (Natural Language Processing) як моделювання концептуальних знань людини; інформаційний пошук, при реалізації якого ЛСА показує набагато кращі результати порівняно із звичайними векторними методами; процес підбору синонімів, якість якого була перевірена і підтверджена шляхом обробки тестів TOEFL (Test of English as a Foreign Language) і результат роботи методу ЛСА виявився ідентичним середнім показникам відбору синонімів людьми із неангломовних країн [11, с. 281]; моделювання семантичного інструктування із подоланням полісемії та омонімії; визначення холістичної оцінки якості розгорнутих письмових відповідей (у вигляді есе), що показує високу кореляцію із оцінками експертів; моделювання розуміння тексту, що фактично залежить від його когерентності тощо.

За результатами аналізу роботи методу латентно-семантичного аналізу 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. Однак, недоліком латентно-семантичного аналізу є відсутність когнітивних вмінь людини, зокрема лінгвістичної складової, а саме, врахування синтаксису та морфології природно мовної текстової інформації, необхідної для автоматизованої обробки.

Висновки і перспективи подальших наукових досліджень… Отже, доходимо висновку, що найпопулярнішими методами обробки природномовної текстової інформації з метою екстракції та репрезентації семантики мають бути системи, що ґрунтуються на ефективному поєднанні лінгвістичних технологій аналізу (графематичного, морфологічного, синтаксичного, семантичного), зокрема із застосуванням онтологій, та методу латентно-семантичного аналізу, що дозволить виявляти приховані асоціативні залежності всередині природномовних текстів. Таке поєднання дає можливість врахувати та частково елімінувати недоліки, що притаманні обом методам, та удосконалити процес обробки природномовної відповіді шляхом комбінування переваг розглянутих лінгвістичних та статистичних методів. Саме інтеграція технологій експліцитного семантичного аналізу, латентно-семантичного аналізу, методів теорії нечіткої логіки, штучного інтелекту та ін. є перспективним шляхом розв'язання проблеми автоматизованого семантичного аналізу природномовної інформації.

Список використаних джерел і літератури:

1. Jones K., Bright W. (ed.) Natural language processing*: a* historical review. *International encyclopedia of linguistics.* New York: Oxford University Press, 1992. Vol. 3. P. 53–59.

2. Комарницька О. І. Моделі та методи лінгвістичного аналізу тексту інтелектуальної системи оцінювання знань: дис. … канд. філол. наук: 10.02.21 / Комарницька Оксана Іванівна. Хмельницький, 2015. 209 с.

3. Марченко О. О. Порівняння методів онтологічного семантичного аналізу та алгоритмів латентного семантичного аналізу. *Вісник Київського національного університету імені Тараса Шевченка.* Сер. : фізико-математичні науки. Київ, 2012. № 2. С. 169−174.

4. Леонтьева Н. Н. Автоматическое понимание текста: системы, модели, ресурсы: учеб. пособ. Москва: Академия, 2006. 304 с.

5. Мельчук И. А. Опыт теории лингвистических моделей «Смысл↔Текст». Москва: Издательство: Школа «Языки русской культуры», 1999. 346 с.

6. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособ / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ и др. Москва: МИЭМ, 2011. 272 с.

7. Даревич Р. Р. Підвищення ефективності інтелектуального аналізу тексту шляхом зважування понять у моделі онтології. *Искусственный интеллект*. 2005. № 3. С. 571–577.

8. Комп'ютерні онтології та їх використання у навчальному процесі. Теорія і практика: [моногр.] / С. О. Довгий, В. Ю. Велічко, Л. С. Глоба та ін. Київ: Інститут обдарованої дитини, 2013. 310 с.

9. Знання-орієнтований підхід до автоматизації інформаційно-аналітичної діяльності / І. В. Замаруєва, А. О. Рось, О. Ю. Губайдулін та ін. *Проблемы программирования*. Киев: ИПС НАНУ, 2000. № 1–2. С. 601–614.

10. Ландэ Д. В. Поиск знаний в Internet. Профессиональная работа. Москва: Издательский дом «Вильямс», 2005. 272 с.

11. Landauer T., Foltz P., Laham D. An introduction to latent semantic analysis. *Discourse Processes*. 1998. №25. P. 259–284.

References:

1. Jones K., Bright W. (ed.) Natural language processing: a historical review. International encyclopedia of linguistics. New York: Oxford University Press, 1992. Vol. 3. P. 53–59.

2. Komarnytska O. I. Modeli ta metody linhvistychnoho analizu tekstu intelektualnoi systemy otsiniuvannia znan: dys. … kand. filol. nauk: 10.02.21 / Komarnytska Oksana Ivanivna. Khmelnytskyi, 2015. 209 s.

3. Marchenko O. O. Porivniannia metodiv ontolohichnoho semantychnoho analizu ta alhorytmiv latentnoho semantychnoho analizu. Visnyk Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. Ser. : fizyko-matematychni nauky. Kyiv, 2012. № 2. S. 169–174.

4. Leont'eva N. N. Avtomaticheskoe ponimanie teksta: sistemy, modeli, resursy: ucheb. posob. Moskva: Akademiya, 2006. 304 s.

5. Mel'chuk I. A. Opyt teorii lingvisticheskix modelej «Smysl↔Tekst». Moskva: Izdatel'stvo: Shkola «Yazyki russkoj kul'tury», 1999. 346 s.

6. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika: ucheb. posob / E. I. Bol'shakova, E'. S. Klyshinskij, D. V. Lande' i dr. Moskva: MIE'M, 2011. 272 s.

7. Darevych R. R. Pidvyshchennia efektyvnosti intelektualnoho analizu tekstu shliakhom zvazhuvannia poniat u modeli ontolohii. Yskusstvennыi yntellekt. 2005. № 3. S. 571–577.

8. Kompiuterni ontolohii ta yikh vykorystannia u navchalnomu protsesi. Teoriia i praktyka: [monohr.] / S. O. Dovhyi, V. Yu. Velichko, L. S. Hloba ta in. Kyiv: Instytut obdarovanoi dytyny, 2013. 310 s.

9. Znannia-oriientovanyi pidkhid do avtomatyzatsii informatsiino-analitychnoi diialnosti / I. V. Zamaruieva, A. O. Ros, O. Yu. Hubaidulin ta in. Problemы prohrammyrovanyia. Kyev: YPS NANU, 2000. № 1–2. S. 601–614.

10. Lande' D. V. Poisk znanij v Internet. Professional'naya rabota. Moskva: Izdatel'skij dom «Vil'yams», 2005. 272 s.

11. Landauer T., Foltz P., Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998. №25. P. 259–284.

Summary
Oksana Komarnytska
Methods of Automatized Semantic Analysis of Natural-Language Information

In the article the features of automated semantic analysis of the text have been considered; problems of creation of automated linguistic-software tools, which are suitable for application in extraction systems of semantics from the text have been investigated. The author has systematized and carried out a comparative analysis of the results of scientific research in the field of developing models and methods of semantic analysis of natural-language information. Two basic approaches in the field of computer processing of natural-language texts have been determined: lingua-analytical and statistical. The author has argued that the most promising and effective of them are, respectively, explicit methods of semantic analysis of text information (algorithms of ontological semantic analysis) and methods of latent semantic analysis. Possible ways of improvement of existing computer means for diagnosing the relevance of natural-language information have been outlined; it has been substantiated that the most popular methods of processing natural-language information for the purpose of extraction and representation of semantics should be systems based on the efficient combination of linguistic analysis technologies (graphical, morphological, syntactic, semantic), in particular using ontologies, and the method of latent semantic analysis. It has been proved that the integration of technologies of explicit semantic analysis, latent semantic analysis, methods of the theory of fuzzy logic, artificial intelligence, etc. is a promising way of solving the problem of automated semantic analysis of natural-language information.

Key words: *automated semantic analysis, natural-language information, linguistic analysis, latent semantic analysis, method, ontology.*

Дата надходження статті: «03» квітня 2018 р.
Дата прийняття до друку: «30» квітня 2018 р.

УДК 811.161.2:81'367

ВАЛЕНТИНА КРИЩУК,
кандидат філологічних наук, доцент
(м. Хмельницький)

Функції монологічного мовлення в епістолярному жанрі

У статті доведено, що епістолярний стиль мови, сфера його використання необмежена – у листі може йти мова про творчі плани, побут, інтимне життя, політику, релігію, науку, освіту чи виробництво. Основне ж призначення мови – це заочне спілкування у монологічному форматі, писемно оформленій матриці, а тематика залежить від сфери їх впливу, зацікавленості обох сторін (адресанта й адресата), сформованого світогляду, мовної культури. Епістолярій