ПРИКЛАДНА ЛІНГВІСТИКА

УДК 81`33 + 004.82 + 004.91
DOI https://doi.org/10.52726/as.humanities/2022.2.14

В. В. ПРИХОДНЮК
кандидат технічних наук,
завідувач відділу створення та використання інтелектуальних мережних інструментів,
Національний центр «Мала академія наук України», м. Київ, Україна
Електронна пошта: Prikhodnyuk_Vitaly@nas.gov.ua
https://orcid.org/0000-0002-2108-7091

В. В. ГОРБОРКУОВ
кандидат технічних наук,
науковий співробітник відділу створення та використання інтелектуальних мережних інструментів,
Національний центр «Мала академія наук України», м. Київ, Україна
Електронна пошта: slavon07@gmail.com
https://orcid.org/0000-0002-2758-7724

МЕТОД РЕКУРСИВНОЇ РЕДУКЦІЇ ЯК СКЛАДОВА ЛІНГВІСТИЧНОГО ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНО-АНАЛІТИЧНИХ СИСТЕМ

Наш час характеризується стрімким ростом кількості наявної інформації, що призводить до створення великих за обсягом масивів тематично та просторово розподіленої інформації. Обробка таких масивів вручну є надзвичайно складним процесом, що потребує значних трудозатрат.

Такі масиви можуть бути оброблені з допомогою сучасних інформаційно-аналітичних систем, зокрема – лексикографічних. Однак значна частина інформації в таких масивах може бути слабко і неструктурованою (зокрема – в формі природномовних текстів), що створює потребу в якісному лінгвістичному забезпеченні таких систем, як елементу прикладної лінгвістики.

Існує велика кількість методів та засобів для обробки природномовних текстів – на основі словників, машинного навчання, статистичних показників тощо. Однак результати роботи всіх цих методів потребують подальшої обробки і перетворення в формат, передбачений системою, що використовується. Саме для такої обробки пропонується метод рекурсивної редукції, що може використовуватись як складова лінгвістичного забезпечення інформаційно-аналітичних систем. Метод передбачає створення формалізованого опису потрібного системі формату даних, на основі якого відбувається перетворення результатів аналізу вхідного природномовного тексту, з подальшим формуванням онтології. Результуюча онтологія може бути використана для формування онтологокерованої лексикографічної системи.

Ключові слова: прикладна лінгвістика**,** лінгвістичне забезпечення, інформаційно-аналітична система, онтологія, лексикографічна система.

Постановлення проблеми. Швидкий ріст кількості інформації, що створюється та використовується людством, робить надзвичайно актуальною задачу використання високоефективних інформаційно-аналітичних систем (ІАС). При цьому значна кількість такої інформації є слабко- і неструктурованою, зокрема представленою в формі природномовних текстів. Це, в свою чергу, потребує розвитку лінгвістичного забезпечення таких систем, як елементу прикладної лінгвістики.

Основною функцією такого лінгвістичного забезпечення є структуризація, тобто перетворення неструктурованої інформації, призначеної для сприйняття людиною (природномовні тексти) в структуровану, призначену для обробки з допомогою ІАС. Саме для виконання даної задачі використовується запропонований статтею метод – метод рекурсивної редукції [Stryzhak, Prykhodniuk, Popova, Nadutenko, Haiko, Chepkov].

Поточний стан проблеми. Автоматичний аналіз та класифікація неструктурованої інформації – це складна задача, способи розв’язання якої охоплюються різними підходами та технологіями [Hartmann, Huppertz,

Schamp, Heitmann; Berardi; Humphreys, Wang; Aggarwal, Zhai]. З часом підходи почали спиратися один на одного, хоча все ще існують чіткі розмежувальні лінії і, як правило, покладаються переважно на той чи інший підхід. Широке розгортання штучного інтелекту зараз стає популярним у багатьох сферах, і аналіз тексту та його структуризація на базі різних типів машинного навчання не є винятком [Sebastiani]. Машинне навчання полягає у застосуванні методів до великого об'єму різнотипних даних, і все частіше використовується для визначення закономірностей в тому числі і в режимі реального часу. Такі системи можуть використовувати контрольоване навчання, яке працює після застосування навчального набору даних, або неконтрольоване навчання, яке працює безпосередньо з даними, які йому представлені [Perkins]. Контрольовані методи машинного навчання на основі тренувальних текстових даних індуктивно створюють класифікатори на основі спостережуваних шаблонів, не вимагаючи ручного визначення правил класифікації. Це робить їх гнучкими в розумінні граматичної конструкції, характерної для певних областей [Hartmann, Huppertz, Schamp, Heitmann].

Крім методів на базі машинного навчання для обробки неструктурованої інформації можуть застосовуватись інші методи, зокрема на основі лексики (*lexicon-based methods*). Такі методи вимагають створених експертами словників, що складаються зі складних списків слів і відповідних міток для класифікації текстових даних [Pennebaker, Boyd, Jordan, Blackburn]. Такі словники часто є загальними та при потребі можуть доповнюватись додатковими користувацькими словами. Однак необхідно створювати власний словник, якщо для роботи з певною областю не має відповідного базового словника.

Важливим аспектом при застосуванні алгоритмів для обробки певної інформації полягає в необхідності опрацьовувати також і не лінгвістичні дані, так звані гіперпараметри (*позначки часу, геодані та інше*) [Perkins].

Метод рекурсивної редукції. Всі вищевказані методи можуть використовуватись для аналізу природномовних текстів та ідентифікації з них інформації. Однак для використання отриманої інформації в роботі ІАС вона повинна бути приведена до певного, сумісного з даною ІАС формату. Для цього необхідний певний формалізований опис даного формату і спеціалізовані засоби, що виконуватимуть власне перетворення на основі даного опису.

Саме для виконання даної функції і призначений метод рекурсивної редукції. Даний метод може бути застосований до певної предметної галузі (ПГ), що описується певним природномовним (ПМ) документом T. Ціллю застосування методу є формування на його основі формалізованого представлення ПГ [Nadutenko, Prykhodniuk, Shyrokov, Stryzhak] у вигляді онтології O.

Процес структуризації документу являє собою послідовність перетворень (1).

$$T \to T_{sn} \to T_{mk} \to \langle X, T_{mk} \rangle \to \\ \to \langle X, R, T_{mk} \rangle \to \langle X, R, F, T_{mk} \rangle \to O, \quad (1)$$

де T – початковий природномовний текст; T_{sn} – лексична структура тексту; T_{mk} – розширена лексична структура; X, R, F – множини об'єктів, зв'язків і функцій інтерпретації вихідної онтології O.

Перетворення $T \to T_{sn}$ є результатом попереднього лексичного або семантичного аналізу. При цьому ніяких жорстких вимог на результати такого аналізу не накладається – він може бути здійснений як простими лексичними аналізаторами, що виділяють з тексту і класифікують лексеми, так спеціалізованими модулями, призначеними для аналізу семантики тексту та виділення спеціалізованих елементів (одиниці вимірювання, координати тощо).

В загальному випадку очікується, що результатом роботи попереднього аналізу є певна лінійно впорядкована послідовність об'єктів – зокрема, речень (2) і лексем в реченнях (3). В такому випадку лінійна впорядкованість на множинах речень та лексем визначається природнім чином заданим відношенням передування $\prec$. Більш складні модулі попереднього аналізу можуть давати в якості результату частково впорядковану множину об'єктів, де відношення часткового порядку буде задаватись спеціалізованими зв'язками – наприклад, семантичними.

Кожна об'єкт (4) складається з текстового представлення і певної множини атрибутів. Для лексем в загальному випадку до множини атри-

бутів входить граматичний клас (частина мови), а інші елементи залежать від мови вихідного тексту і можливостей аналізатора.

$$T = \{S_1 \prec S_2 \prec \ldots \prec S_n\} \quad (2)$$

$$S_i = \{l_1 \prec l_2 \prec \ldots \prec l_m\} \quad (3)$$

$$l_{sn} =< t, P_{sn} >, \quad (4)$$

де S_i – речення; l_{sn} – об'єкт (лексема); t, P_{sn} – текстове представлення та множина лексичних атрибутів лексеми відповідно.

Задане на множині об'єктів відношення може бути представлене як певна множина бінарних зв'язків виду (5). Множина ознак, як і у випадку об'єкт, повинна містити тип зв'язку – в найпростішому випадку це «передує в тексті».

$$r_{sn} = \left\langle l_1, l_2, P_{r_{sn}} \right\rangle, \quad (5)$$

де l_1, l_2 – певні лексеми вхідного тексту; $P_{r_{sn}}$ – ознаки зв'язку r_{sn}.

Така структура утворюватиме орієнтований граф (6), що і представлятиме лексичну структуру тексту.

$$T_{sn} = \left\langle L_T, R_{sn} \right\rangle, \quad (6)$$

де $L_T = \bigcup_{S \in T} S$ – множина лексем тексту T; R_{sn} – множина всіх зв'язків виду (5) між лексемами тексту.

Структура (6) може використовуватись для аналізу текстових документів, але її використання вимагає достатньо повного формалізованого опису мови, якою написаний вихідний текст. В загальному випадку створення такого опису є дуже трудомісткою задачею, що потребує значних витрат часу. Однак в більшості реальних ситуацій її можна спростити за рахунок використання двох інструментів: використання тезаурусів і аналізу вихідної розмітки документу.

Справа в тому, що більшість форматів вихідних файлів (DOC/DOCX, PDF, HTML тощо) містять в собі елементи внутрішньої структури – абзаци, таблиці, виділені різним шрифтом області та ін. Додавання до структури (6) інформації про таку розмітку дозволить створювати правила аналізу розмітки конкретного типу документів, що більшості задач буде значно ефективніше ніж спроби аналізувати зміст самого вхідного тексту. При цьому загальна структура (6) не потребує особливих змін, однак необхідно значно розширити множини атрибутів об'єктів і зв'язків.

Об'єкт необхідно розширити, додавши до нього додаткову множину атрибутів – атрибутів розмітки (7). Зв'язки ж, задані оригінальною розміткою, будуть в загальному випадку повністю відрізнятись від створених в ході попереднього аналізу, і утворюватимуть окрему множину (8). Це дозволяє виконати перетворення $T_{sn} \to T_{mk}$, сформувавши розширену лексичну структуру (9).

$$l_{mk} =< t, P_{sn}, P_{mk} >, \quad (7)$$

де l_{mk} – розширена лексема; t, P_{sn} – текстове представлення, множина лексичних атрибутів лексеми відповідно; P_{mk} – множина атрибутів розмітки документа, що відносяться до лексеми.

$$r_{mr} = \left\langle l_1, l_2, P_{r_{mr}} \right\rangle, \quad (8)$$

де l_1, l_2 – певні лексеми вхідного тексту; $P_{r_{mr}}$ – ознаки зв'язку r_{mr}.

$$T_{mk} = \left\langle L_{mk}, R_{sn} \bigcup R_{mk} \right\rangle, \quad (9)$$

де L_{mk} – множина розширених лексем (7); R_{sn} – множина лексичних зв'язків (5); R_{mk} – множина зв'язків, сформованих розміткою документу (8).

Структура (9) є вхідними даними для застосування методу рекурсивної редукції. Метод полягає в рекурсивному застосуванні до даної структури спеціалізованих функцій, що здійснюють редукцію, тобто перетворення груп лексем на об'єкти та інші елементи онтології.

Функція редукції призначена для структуризації документів в рамках конкретної задачі, тому її робота базується на певній базі правил перетворення виду (10). На основі такої бази правил формується рекурсивна функція виду (11), що використовується для перебору всіх можливих підмножин вхідної множини лексем доти, доки функція ідентифікації одного з правил не дасть позитивний результат.

$$d = \left\langle f_{id}^d, f_{tr}^d \right\rangle, \quad (10)$$

де f_{id}^d – функція ідентифікації лексем; f_{tr}^d – функція перетворення лексем.

$$F_{rd}(\beta_L, D) = \begin{cases} f_{tr}^d(\beta_0), \exists d \in D, f_{id}^d(\beta_0) \\ F_{rd}(\beta_L \setminus \beta_0, D) \end{cases} \quad (11)$$

де β_L – множина всіх підмножин, сформована для певної множини лексем L; D – база правил виду (10); f_{id}^d, f_{tr}^d – елементи певного правила d.

Функція ідентифікації призначена для перевірки умови виду (12). Дана функція задається двома наборами предикатів – вектором предикатів перевірки лексем і матрицею предикатів перевірки зв'язків.

$$f_{id}(L) = p_i^l(l_i) \wedge ... p_n^l(l_n) \wedge p_{12}^r(l_1, l_2) ... \wedge p_{n-1n}^r(l_{n-1}, l_n), \quad (12)$$

де $L = \{l_1, ... l_n\}$ – вхідна множина лексем; p_i^l – предикат перевірки лексеми; p_{ij}^r – предикат перевірки зв'язку.

Предикати перевірки лексеми можна поділити на групи:

– Предикати перевірки текстового значення – накладають умову на текстове значення лексеми. Їх можна розділити на:

о Предикати на базі тезаурусів – перевіряють входження лексеми до певного словника ключових слів;

о Предикати на базі регулярних виразів – задають шаблон, якому повинне відповідати текстове значення, з допомогою одної з формальних мову опису регулярних виразів;

– Предикати перевірки лексичних атрибутів – визначають певний атрибут або набір атрибутів, що повинні міститись в множині лексичних атрибутів лексеми (13).

– Предикати перевірки атрибутів оригінальної розмітки – аналогічні, але перевіряють входження атрибутів до множини атрибутів оригінальної розмітки (14).

$$p^l(l) \equiv P_{sn}^l \bigcap P_p \neq \varnothing \quad (13)$$

$$p^l(l) \equiv P_{mr}^l \bigcap P_p \neq \varnothing \quad (14)$$

Функція перевірки зв'язків в загальному випадку має структуру, аналогічну до (13) чи (14) – в залежності від типу зв'язка. Також обидва предикатів можуть мати нульову варіацію (15), яка використовується для відмічення неважливих для конкретного правила елементів вхідної послідовності.

$$p \equiv 1 \quad (15)$$

Функція перетворення, на відміну від функції ідентифікації, базується не на предикатах (тобто функціях, що можуть мати позитивне чи негативне значення), а на власне перетвореннях – тобто операціях, що змінюють певним чином вхідну структуру T_{mr}.

Базові функції перетворення, у відповідності до (1), виконують просту операцію – видалення частини вхідної структури тексту і створення на основі видаленої послідовності нових об'єктів (16), зв'язків (17) чи атрибутів (18).

$$\langle T_{mr}, X \rangle \xrightarrow{f_{tr}(L)} \langle T_{mr} \setminus L', X \cup \{x(L')\} \rangle \quad (16)$$

$$\langle T_{mr}, X, R \rangle \xrightarrow{f_{tr}(L)} \langle T_{mr} \setminus (L_1' \cup L_2'), X, R \cup \{r(x(L_1'), x(L_2'))\} \rangle \quad (17)$$

$$\langle T_{mr}, X, A \rangle \xrightarrow{f_{tr}(L)} \langle T_{mr} \setminus (L_1' \cup L_2'), X, A \cup \{a(x(L_1'), L_2')\} \rangle, \quad (18)$$

де T_{mr} – розширена структура тексту; X, R, A – множини об'єктів, зв'язків і атрибутів онтології, що створюється; $L' \subseteq L$ – задана налаштуваннями перетворення підмножина лексем вхідної послідовності; $x(l)$, $r(x_1, x_2)$, $a(x, l)$ – операції створення об'єктів, зв'язків та атрибутів відповідно.

Функція перетворення в найпростішому випадку може задаватись одним або кількома наборами індексів $\{i_1, ... i_k\}$, таких, що $L' = \{l_{i_1}, ... l_{i_k}\}$. Однак в загальному випадку множина лексем L' не може бути використана в ясному вигляді і потребує попередньої обробки виду

$$L' \to t_L \xrightarrow{f_{norm}} t_{norm}, \quad (19)$$

де L' – вхідна підмножина лексем; t_L – текстове представлення підмножини лексем; t_{norm} – нормалізоване текстове представлення; f_{norm} – функція нормалізації.

Функція нормалізації може виконувати одну або кілька з наступних дій:

1) Узгодження лексем між собою за відмінком, родом та множиною.

2) Нормалізація за відмінком там множиною (як правило – приведення до називного відмінку однини).

3) Синонімічне перетворення, тобто заміна терміну на більш підходящий синонім.

4) Спеціалізовані перетворення типів, такі як зміна одиниць виміру, форматування чисел тощо.

Правильна нормалізація є запорукою коректного створення онтології. Інакше результат міститиме велику кількість об'єктів-синонімів, які ускладнюватимуть роботу з онтологією.

Вищевказаний процес дозволяє структуризувати наявну в природномовних текстах інформацію, представляючи її в потрібній для ІАС формі. За замовчуванням результатом роботи методу рекурсивної редукції є онтологія, що може бути використана для формування онтологокерованих лексикографічних систем [Nadutenko, Prykhodniuk, Shyrokov, Stryzhak]. Однак онтології легко можуть бути перетворені в будь-який інший тип даних, зокрема – в електронні таблиці. Це дозволяє здійснювати інтеграцію з широким спектром вже існуючих ІАС, не залежачи від наявних в них інтерфейсів.

Висновок. Запропоновано метод рекурсивної редукції, що може бути складовою лінгвістичного забезпечення інформаційно-аналітичних систем при аналізі великих масивів просторово і тематично розподіленої інформації у форматі природномовних документів. Метод призначений для перетворення результатів аналізу (зокрема, лексичного) природномовних текстів в формат, що використовується інформаційно-аналітичними системами. Створені в результаті роботи методу онтології можуть використовуватись для створення онтологокерованих лексикографічних систем.

ЛІТЕРАТУРА

1. Stryzhak O., Prykhodniuk V., Popova M., Nadutenko M., Haiko S., Chepkov R. Development of an Oceanographic Databank Based on Ontological Interactive Documents. *Lecture Notes in Networks and Systems*. Cham : Springer. 2021. С. 97–114. DOI: https://doi.org/10.1007/978-3-030-80126-7_8

2. Hartmann J., Huppertz J., Schamp C., Heitmann M. Comparing automated text classification methods. *International Journal of Research in Marketing*. 2019. вип. 36. № 1. С. 20–38. DOI: https://doi.org/10.1016/j.ijresmar.2018.09.009

3. Berardi G. Semi-automated text classification. *ACM SIGIR Forum*. 2014. вип. 48. № 1. С. 42–42. DOI: https://doi.org/10.1145/2641383.2641392

4. Humphreys A., Wang R. J.-H. Automated Text Analysis for Consumer Research. *Journal of Consumer Research*. 2018. вип. 44. № 6. С. 1274–1306. DOI: https://doi.org/10.1093/jcr/ucx104

5. Aggarwal C. C., Zhai C. A Survey of Text Classification Algorithms. *Mining Text Data*. Boston, MA : Springer US. 2012. С. 163–222. ISBN: 978-1-4614-3223-4 DOI: https://doi.org/10.1007/978-1-4614-3223-4_6

6. Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*. 2002. вип. 34. № 1. С. 1–47. DOI: https://doi.org/10.1145/505282.505283

7. Perkins M. Approaches to Text Analysis. *Global Language Review*. 2019. вип. IV. № I. С. 1–7. DOI: https://doi.org/10.31703/glr.2019(IV-I).01

8. Pennebaker J. W., Boyd R. L., Jordan K., Blackburn K. The Development and Psychometric Properties of LIWC2015. *University of Texas at Austin*. 2015. С. 1–26. DOI: https://doi.org/10.15781/T29G6Z

9. Nadutenko M., Prykhodniuk V., Shyrokov V., Stryzhak O. Ontology-Driven Lexicographic Systems. *Advances in Information and Communication. FICC 2022. Lecture Notes in Networks and Systems*. Cham : Springer. 2022. С. 204–215. DOI: https://doi.org/10.1007/978-3-030-98012-2_16

V. V. PRYKHODNIUK
PhD in Technical Sciences,
Head of the Department of Creation and Use of Intelligent Network Tools,
National Center "Minor Academy of Sciences of Ukraine", Kyiv, Ukraine
E-mail: Prikhodnyuk_Vitaly@nas.gov.ua
https://orcid.org/0000-0002-2108-7091

V. V. HORBORKUOV
PhD in Technical Sciences,
Researcher at the Department of Creation and Use of Intelligent Network Tools,
National Center "Minor Academy of Sciences of Ukraine", Kyiv, Ukraine
Email: slavon07@gmail.com
https://orcid.org/0000-0002-2758-7724

RECURSIVE REDUCTION METHOD AS A COMPONENT OF LINGUISTIC SUPPORT OF INFORMATION-ANALYTICAL SYSTEMS

Our time is characterized by a rapid increase in the amount of available information, which leads to the creation of large-scale arrays of thematically and spatially distributed information. Manual processing of such arrays is an extremely complex process that requires significant labor costs.

Such arrays can be processed with the help of modern information-analytical systems, in particular – lexicographic ones. However, a significant part of the information in such arrays may be weakly structured or unstructured (in particular, in the form of natural language texts), which creates a need for high-quality linguistic support for such systems as an element of applied linguistics.

There is a large number of methods and tools for processing natural language texts – based on dictionaries, machine learning, statistical indicators, etc. However, the results of all these methods require further processing and conversion into the format required by the system in use. For such processing the method of recursive reduction is proposed, which can be used as a component of the linguistic support of information-analytical systems. The method involves the creation of a formalized description of the data format required by the system, based on which the results of the analysis of the input natural language text are transformed, with a creation of an ontology. The resulting ontology can be used to form an ontology-driven lexicographic system.

Key words: applied linguistics, linguistic support, information-analytical system, ontology, lexicographic system.