DOI https://doi.org/10.15407/csc.2021.04.013
УДК 004.62:004.91

О.О. ЛЕТИЧЕВСЬКИЙ, доктор фіз.-мат. наук, завідувач відділу,
Інститут кібернетики імені В.М. Глушкова НАН України,
03187, м. Київ, просп. Академіка Глушкова, 40, Україна,
oleksandr.letychevskyi@litsoft.com.ua

М.К. МОРОХОВЕЦЬ, кандидат фіз.-мат. наук, старший науковий співробітник, Інститут кібернетики імені В.М. Глушкова НАН України,
03187, м. Київ, просп. Академіка Глушкова, 40, Україна,
marina.morokhovets@gmail.com

Н.М. ЩОГОЛЕВА, науковий співробітник, Інститут кібернетики імені В.М. Глушкова НАН України,
03187, м. Київ, просп. Академіка Глушкова, 40, Україна,
natashch2904@gmail.com

ДЕЯКІ ЗАСОБИ ОБРОБКИ ЕЛЕКТРОННИХ ТЕКСТОВИХ ДОКУМЕНТІВ

Розглянуто проблеми, що виникають при використанні автоматичних засобів лінгвістичного аналізу текстів, поданих природною мовою. Подано параметричну модель структурування довгих речень, що містять переліки. Розглянуто джерела та види синонімії найменувань об'єктів. Запропоновано метод усунення синонімії найменувань об'єктів, про які йдеться у тексті, призначеному для автоматичного аналізу.

Ключові слова: *цифрові юридичні документи, синонімія найменувань об'єктів, структурування речення, надійність системи штучного інтелекту.*

Вступ

Цифровізація законодавства є важливим напрямом сьогодення, що визначено урядом, як пріоритетний. Створення цифрових юридичних документів та перевірка їх на відповідність законам є необхідним завданням в усіх галузях юриспруденції.

У зв'язку з цим виникає задача автоматичної формалізації юридичного документу, створеного як довільний текст природною мовою.

Ця задача еквівалентна відомій задачі «розуміння» тексту та потребує використання і лінгвістичних формальних методів, і методів машинного навчання. Проблема ускладнюється також специфікою української мови, адже з нею працює доволі невелика кількість наявних лінгвістичних систем через особливості морфології, що відрізняється від аналітизму романо-германських мов.

Прикладом систем морфологічного аналізу є такі системи, як *NLTK* (*Natural Language Toolkit*) [1], ОРФО [2], *LingPipe* [3], МетаФраз [4], *Pullenti SDK* [5]. Основні функції таких систем — розбивання тексту на певні складові, аналіз та встановлення зв'язків між ними. Однією із задач обробки документів є класифікація даних за властивостями. Хоча однією з технологій класифікації є машинне навчання на

прикладах з ігноруванням семантики класифікованих сутностей, розпізнання властивостей семантичним аналізом також має значення й іноді комбінується із машинним навчанням.

У задачах класифікації виникає, зокрема, проблема синонімів, що ускладнює розпізнання тексту та класифікацію сутностей.

У цій роботі розглядається метод, що дає змогу подолати таку складність.

Постановка проблеми та мета роботи

Зберігання юридичних документів різних видів у цифровому форматі створює можливості їхнього пошуку та обробки за допомогою ЕОМ. Підготовка документа до зберігання у цифровому форматі з метою подальшої обробки може потребувати попередньої роботи з початковим текстом. Обробка тексту включає його перетворення, видобування даних. Робота з текстами для розв'язання різноманітних задач потребує обробки тексту зокрема як мовного об'єкта лінгвістичного аналізу. Прикладами такої обробки є синтаксичний аналіз тексту, перевірка тексту на наявність у ньому певних мовних конструкцій.

При використанні автоматичних засобів лінгвістичного аналізу поданих природною мовою текстів, зокрема, юридичних, що здійснюють обробку тексту по реченнях (послідовно опрацьовуючи текст речення за реченням), виникають проблеми локального та глобального характеру.

Локальна проблема: у тексті є речення (можливо, кілька), що є складним для обробки як окрема складова тексту, тобто обраний засіб (або наявні засоби) аналізу не можуть його обробити в належний спосіб.

Глобальна проблема: для досягнення мети обробки тексту або його окремих речень потрібно при аналізі речення використовувати результати аналізу інших речень або всього тексту, тобто використовувати дані, що виходять за межі речення.

Проблему локального характеру створює, зокрема, наявність у тексті одного чи кількох речень, що через значну довжину виявляються складними для обробки (за допомогою того чи іншого засобу аналізу тексту). Проблема глобального характеру виникає, коли при автоматичній обробці тексту має враховуватися змістовий зв'язок між складовими різних речень.

У разі виникнення подібних проблем при використанні засобів аналізу текстів по реченнях з'являється потреба у додаткових засобах попередньої обробки тексту.

У цій роботі пропонуються такі засоби:

- модель структурування довгих речень, що містять переліки,
- метод усунення синонімії найменувань об'єктів, про які йдеться у тексті, призначеному для автоматичного аналізу.

Структурування речень із переліками

У спеціальних текстах, зокрема, юридичних, трапляються речення, що мають значну довжину й через це виявляються складними для обробки. Часто такі речення містять довгі переліки. Переліки у реченні бувають простими або вкладеними. Перелік у реченні назвемо простим, якщо жодна його складова не містить переліку. Перелік у реченні назвемо вкладеним, якщо принаймні одна його складова містить перелік.

Окремі складові переліку в реченні можуть використовуватись як певною мірою автономні одиниці (наприклад, на них можуть бути посилання з того самого тексту або з інших). У цій роботі для тексту, що призначений для зберігання у формі електронного документа, запропоновано розмічати речення, що містять переліки. Оскільки всі випадки використання та обробки електронного документа не фіксуються наперед, пропонований засіб побудовано так, щоб зберігався його загальний характер.

Для розмічання потрібні позначки, для яких попередньо мають бути визначені місця та зміст. Доцільно позначати початок та кінець переліку, вказувати рівень переліку, а також зазначати, як складові переліку розділяються у

реченні. Пропонується такий набір позначок для розмічання переліку.

Позначення початку переліку:

<знак переліку> <позначка початку переліку> <рівень переліку> <розділовий знак>

Позначення кінця переліку:

<знак переліку> <позначка кінця переліку> <рівень переліку>.

Складова <знак переліку> потрібна для позначення початку (кінця) структурного розділу речення. Складові <позначка початку переліку> та <позначка кінця переліку> дають змогу виокремити структурний розділ з конкретним переліком. Складова <рівень переліку> потрібна для структурування вкладеного переліку. Складова <розділовий знак> потрібна для забезпечення доступу до окремих елементів переліку.

Пропонована модель структурування речень з переліком є параметричною. Позначки доцільно обирати, коли визначено мету та засоби обробки речень, що розмічаються.

Синонімія найменувань

У текстах, поданих натуральною мовою, зокрема, юридичних, той самий об'єкт може мати різні найменування. «Ідентичність значень найменувань» становить проблему аналізу натуральномовних текстів за допомогою машинних програм.

Під «найменуванням» розуміємо ім'я, назву (офіційну або неофіційну, загального чи локального вжитку) об'єкта.

Під «ідентичністю значень найменувань» розуміємо повний збіг значень найменувань об'єкта, зокрема, збіг значень найменувань того самого об'єкта, написаних у тексті порізному. У такому сенсі ідентичність значень найменувань є різновидом синонімії (у даному разі — синонімії найменувань). Ідентичні за значенням найменування, тобто такі, що позначають той самий об'єкт, називатимемо синонімічними. Синонімія найменувань у тексті становить проблему глобального характеру, оскільки для роботи з синонімічними найменуваннями потрібні засоби, що виходять за межі обробки окремого речення.

Джерелами «ідентичності значень найменувань» у тексті є:

- зміна форми найменування,
- заміна одного найменування іншим,
- інші.

Розглянемо види та наведемо приклади зміни форми найменування об'єкта.

Використання найменування в різних відмінках. Це поширений вид зміни форми найменування об'єкта в тексті, поданому природною мовою (наприклад, *Міністерство юстиції, Міністерства юстиції*).

Використання різних варіантів написання найменування. Незрідка автори тексту вдаються до різних написань найменування того самого об'єкта. До цього виду зміни форми найменування належать такі:

- використання абревіатур (наприклад, написання *ТОВ Ххх* замість (або також) *Товариство з обмеженою відповідальністю Ххх*),
- використання великих літер замість малих (наприклад, написання *ТОВ ХХХ* замість (або також) *ТОВ Ххх*),
- внесення додаткових допоміжних символів (наприклад, написання *ТзОВ Ххх* замість (або також) *ТОВ Ххх*).

Розгляньмо види та приклади заміни одного найменування іншим. Ідентичні за значенням, але різні за написанням, найменування утворюються не лише внаслідок видозміни форми деякого найменування. Часто в текстах спостерігається таке явище, як заміна одного найменування іншим зі збереженням змісту первісного найменування. До видів заміни одного найменування іншим належать:

- використання псевдо (наприклад, ПП *Ххх* (далі — Замовник)),
- використання займенників,
- використання скорочених найменувань замість повних (наприклад, «*Хх*» замість (або також) *ТЦ* «*Хх*»),
- використання загальних найменувань замість спеціальних (наприклад, підприємство замість (або також) *ПП Ххх*).

До категорії «інші» зараховуємо ті випадки ідентичності значень найменувань, що, можливо, не увійшли у наведений перелік (наприклад, орфографічні помилки в написанні найменування) або можуть з'явитися згодом.

Зазначимо, що визначення того, з яким саме об'єктом пов'язане те чи інше найменування в тексті, може виходити за межі можливостей засобу, що використовується для обробки тексту. Зарадити в таких ситуаціях може Оракул (спеціальний компетентний суб'єкт, знавець, гідний довіри). Також Оракул стає у пригоді, коли потрібно гарантувати правильність виконання тієї чи іншої дії або проконтролювати здійснення перевірки певної умови.

Далі пропонується метод обробки ідентичних за значенням найменувань у натуральномовному тексті, призначений для часткового усунення синонімії найменувань. Ідея методу полягає в тому, щоб, по-перше, замінити найменування об'єктів з початкового тексту на стандартні так, що найменування того самого об'єкта, ідентичні за значенням, замінюються тим самим стандартним іменем, а, по-друге, зберегти зв'язок між первісними та стандартними найменуваннями.

Вхідні дані методу: юридичний текст.

Результат: перетворений текст, у якому замість первісних найменувань об'єктів використано стандартні імена, та таблиця імен, що описує зв'язок між первісними та стандартними найменуваннями.

Метод усунення синонімії найменувань у документі.

1. За заданим текстом скласти список входжень найменувань юридичних осіб. Повноту списку контролює Оракул.

2. Упорядкувати знайдені входження за абеткою.

3. Розбити входження на групи, кожна з яких містить усі найменування того самого об'єкта. Правильність розбиття контролює Оракул.

4. Зв'язати з кожною групою синонімічних найменувань стандартне найменування.

5. Скласти таблицю відповідності первісних та стандартних найменувань.

6. Замінити кожне входження найменування у тексті стандартним іменем згідно з побудованою таблицею.

Оракул пп. 1, 3 — це суб'єкт, що є надійним та відповідальним, а також уповноважений приймати рішення. Оракулом може бути людина-експерт, машинна програма (що має спеціальний статус), людино-машинна система відповідного ґатунку.

Зазначимо, що складання списку п.1 можна автоматизувати, принаймні частково. Прикладами робіт з текстами українською мовою, що здійснюються у даному напрямку, є [6, 7].

Висновки

Розмічання речень з переліками корисне, зокрема, коли текст призначено для аналізу за допомогою процедури, що здійснює обробку тексту по реченнях. Структурування речення з переліком дає змогу, з одного боку, підготувати речення до обробки частинами, а з іншого боку, не втратити цілісність речення під час обробки частинами. Пропоновану модель структурування планується використати в експериментах з пристосування мовного процесора [8] та подібних до нього, побудованих на основі розширної системи програмування [9], до обробки юридичних документів.

Різні види обробки текстів, поданих природною мовою, потребують не лише синтаксичного, а й семантичного аналізу. Зокрема, виявлення синонімічних найменувань об'єктів у тексті пов'язано з його семантикою.

Автоматизація тих видів роботи з текстом, що потребують його розуміння, складає частину діяльності у межах такої галузі як штучний інтелект (ШІ).

Сучасні дослідники (наприклад, [10, 11]) акцентують важливість таких властивостей систем штучного інтелекту як надійність (тобто здатність бути гідним довіри) й пояснюваність (тобто здатність системи ШІ видати разом з результатом своєї роботи пояснення того, на яких засадах й як саме цей результат побудовано) та звертають увагу на брак цих властивостей у наявних систем ШІ. У роботі [10] пропо-

нується доповнювати автоматичні системи прийняття рішень Оракулом для уникнення непередбачених негативних наслідків рішень. У запропонованому в цій роботі методі усунення синонімії найменувань кроки виявлення найменувань об'єктів та ідентичності найменувань потребують семантичного аналізу. Для контролю правильності виконання цих кроків уведений Оракул як засіб підвищення надійності результату роботи.

ЛІТЕРАТУРА

1. *Loper E., Bird S.* NLTK: the natural language toolkit. *arXiv* cs/0205028. Department of Computer and Information Science University of Pennsylvania, Philadelphia, PA 19104-6389, USA, 2002. URL: https://arxiv.org/pdf/cs/0205028.pdf.
2. Многофункциональная система проверки правописания текстов. *ОРФО*. URL: https://orfo.ru/.
3. *Carpenter B.* Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval. 2004. URL: https://trec.nist.gov/pubs/trec13/papers/alias-i.geo.pdf.
4. МетаФраз. URL: http://www.metafraz.ru.
5. Pullenti 4.3. URL: http://www.pullenti.ru/.
6. *Глибовець А. М.* Автоматизований пошук іменованих сутностей у нерозмічених текстах українською мовою. *Штучний інтелект*. 2017. 2. С. 45–51. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/133662/05-Glibovets.pdf?sequence=1.
7. *Погорілий С. Д., Крамов А. А.* Метод виявлення іменних груп в україномовних текстах. *Control Systems and Computers*. 2019. 5 (283). С. 48–59. DOI: https://doi.org/10.15407/csc.2019.05.048.
8. *Мищенко Н. М., Мороховец М. К., Фелижанко О. Д., Штелик Е. В., Щёголева Н. Н.* Новые функциональные возможности системы обработки естественноязыковых спецификаций и среда ее функционирования. *Кибернетика и системный анализ*. 2018. 54 (6). С. 37–46. URL: http://www.kibernetika.org/PDFsE/2018/06/5.pdf.
9. *Мищенко Н. М., Щёголева Н. Н.* О проектировании языковых процессоров на ПЭВМ. *Кибернетика и системный анализ*. 1993. 2. С. 110–117.
10. *Sileno G., Boer A., van Engers T.* The Role of Normware in Trustworthy and Explainable AI. *Proceedings of the XAILA workshop on explainable AI and Law, in conjunction with JURIX 2018, CEUR Workshop Proceedings*. 2018. 2381. P. 9–16. URL: http://ceur-ws.org/Vol-2381/xaila2018_paper_5.pdf.
11. *Sileno G.* Of Duels, Trials and Simplifying Systems. *European Journal of Risk Regulation*. 2020. 11 (3). P. 683–692. DOI: https://doi.org/10.1017/err.2020.38.

Надійшла 28.05.2021

REFERENCES

1. *Loper E., Bird S.,* 2002. "NLTK: the natural language toolkit", arXiv cs/0205028, Department of Computer and Information Science University of Pennsylvania, Philadelphia, PA 19104-6389, USA. [online] Available at: <https://arxiv.org/pdf/cs/0205028.pdf>.
2. "Mnogofunktsionalnaia sistema proverki pravopisaniia tekstov" ["Multifunctional system for checking the spelling of texts"], ORFO. [online] Available at: <https://orfo.ru//>. (In Russian).
3. *Carpenter B.*, 2004. Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval. [online] Available at: <https://trec.nist.gov/pubs/trec13/papers/alias-i.geo.pdf>.
4. MetaFraz. [online] Available at: <http://www.metafraz.ru>. (In Russian).
5. Pullenti 4.3. [online] Available at: <http://www.pullenti.ru/>.
6. *Glybovets A. M.*, 2017. "Avtomatyzovanyi poshuk imenovanyh sutnostei u nerozmichenyh tekstah ukrayinskoiu movoiu" ["Automated search of named entities in unmarked Ukrainian texts"], Shtuchnyi intelekt, 2, pp. 45–51. [online] Available at: <http://dspace.nbuv.gov.ua/bitstream/handle/123456789/133662/05-Glibovets.pdf?sequence=1> (In Ukrainian).
7. *Pogorilyy S. D., Kramov A. A.*, 2019. "Method of Noun Phrase Detection in Ukrainian Texts", Control Systems and Computers, 5 (283), pp. 48–59. DOI: 10.15407/csc.2019.05.048. (In Ukrainian).
8. *Mishchenko N. M., Morokhovets M. K., Felizhanko O. D., Shtelik Y. V., Shchogoleva N. N.*, 2018. "Novyie funktsionalnyie vozmozhnosti sistemy obrabotki iestestvennoiazykovykh spetsifikatsii i sreda ieio funktsionirovaniia" ["New functionalities of the system for natural-language specifications processing and its operating environment"], Cybernetics and Systems Analysis, 54 (6), pp. 37–46. [online] Available at: <http://www.kibernetika.org/PDFsE/2018/06/5.pdf>. (In Russian). (See also: Cybernetics and Systems Analysis, 54 (6), pp. 883–891. (In English)).

9. *Mishchenko N. M., Shchegoleva N. N.*, 1993. "O proektirovanii iazykovykh protsessorov na PEVM" ["On computer-aided design of language processors"], Cybernetics and Systems Analysis, 2, pp. 110–117 (In Russian). (See also: Cybernetics and Systems Analysis, 29, pp. 242–246. DOI: 10.1007/BF01132785. (In English)).
10. *Sileno G., Boer A., van Engers T.*, 2018. "The Role of Normware in Trustworthy and Explainable AI", Proceedings of the XAILA workshop on explainable AI and Law, in conjunction with JURIX 2018, CEUR Workshop Proceedings, 2381, pp. 9–16. [online] Available at: <http://ceur-ws.org/Vol-2381/xaila2018_paper_5.pdf>.
11. *Sileno G.,* 2020. "Of Duels, Trials and Simplifying Systems", European Journal of Risk Regulation, 11 (3), pp. 683–692. *DOI: 10.1017/err.2020.38.* DOI: 10.1017/err.2020.38.

Received 28.05.2021

O.O. Letychevskyi, Doctor (Phys.-Math.), Head of department,
V.M. Glushkov Institute of Cybernetics of NAS of Ukraine,
Glushkov ave., 40, Kyiv, 03187, Ukraine,
oleksandr.letychevskyi@litsoft.com.ua

M.K. Morokhovets, Ph.D. (Phys.-Math.) Sciences, Senior Research Associate,
V.M. Glushkov Institute of Cybernetics of NAS of Ukraine,
Glushkov ave., 40, Kyiv, 03187, Ukraine,
marina.morokhovets@gmail.com

N.M. Shchogoleva, Researcher Associate,
V.M. Glushkov Institute of Cybernetics of NAS of Ukraine,
Glushkov ave., 40, Kyiv, 03187, Ukraine,
natashch2904@gmail.com

SOME MEANS FOR PROCESSING ELECTRONIC TEXT DOCUMENTS

Introduction. Digitization of legislation is an important area today, which is identified by the government as a priority. Creating digital legal documents and verifying them for compliance with the law is a necessary task in all areas of jurisprudence.

This sets the task of automatic formalizing a legal document created as an arbitrary text in natural language.

Purpose. Preparing a document for storage in digital format for further processing may require prior work with an original text.

When using automatic means of linguistic analysis of the text submitted in natural language, in particular, legal, which process the text in sentences (working up the text sequentially sentence by sentence), problems of local and global nature arise.

The problem of local nature is created, in particular, by the presence in the text the sentences, which due to their considerable length are difficult to process (with the help of one or another tool of text analysis). The problem of a global nature arises when the semantic connection between the components of different sentences should be taken into account during the automatic processing of the text. The purpose of this work is to develop means for overcoming these problems.

Results. A model for structuring long sentences containing enumerations as well as a method for eliminating the synonymy of object names referred to in the text, which is intended for automatic analysis, has been developed.

Conclusion. Marking up sentences containing enumerations is useful, especially when the text is intended for analysis using a procedure that processes the text sentence by sentence. Structuring a sentence with an enumeration enables, on the one hand, to prepare the sentence for processing in parts, and on the other hand, not to lose the integrity of the sentence when processing in parts.

In the method of eliminating the synonymy of names proposed in this paper, both the step of identifying the names of objects and the step of revealing the identity of names require semantic analysis. To control the correctness of these steps, the Oracle was introduced to improve the reliability of the result.

Keywords: *digital legal documents, synonymy of object names, sentence structuring, trustworthiness of artificial intelligence systems.*