Маргарита Надутенко, Максим Надутенко
Український мовно-інформаційний фонд НАН України

ЦИФРОВІ МЕТОДИ ТА ШТУЧНИЙ ІНТЕЛЕКТ У МОВОЗНАВЧИХ ДОСЛІДЖЕННЯХ

Обґрунтовано тезу про основні методи цифрових лінгвістичних досліджень, які сформовано та застосовано в Українському мовно-інформаційному фонді (УМІФ); наведено приклади використання цифрових методів та штучного інтелекту для створення аналітичних платформ та мультимедійних словників УМІФ НАН України. Презентовано елементи теорії та експерименту у середовищі штучних нейронних мереж в моделі узагальнених семантичних станів.

Ключові слова: *цифрові методи лінгвістичних досліджень, аналітичні платформи, мультимедійний словник, штучний інтелект, машинне навчання, нейронні мережі, теорія семантичних станів.*

The article discusses the main methods of digital linguistic research, which were formed and applied in the Ukrainian Lingua-Information Fund (ULIF); the article presents the examples of use of digital methods and artificial intelligence for creation of analytical platforms and multimedia dictionaries of ULIF NAS of Ukraine. The elements of theory and experiment in the environment of artificial neural networks in the model of generalized semantic states are presented.

Keywords: *digital methods of linguistic research, analytical platforms, multimedia dictionary, artificial intelligence, machine learning, neural networks, theory of semantic states.*

Постійне збільшення обсягів накопиченої лінгвістичної інформації та висока швидкість надходження нових даних зумовлюють потребу в цифрових технологіях високої ефективності.

Актуальність проблеми визначається необхідністю застосування методів цифрових технологій для комплексного аналізу лінгвістичного матеріалу на всіх етапах роботи з текстом: цифровізація контенту, автоматичне оброблення, систематизація, класифікація та відповідна презентація великих масивів даних і забезпечення доступності набутих знань для широкого кола користувачів із застосуванням можливостей штучного інтелекту та нейронних мереж.

Наукова новизна дослідження полягає у розробленні комплексу методів для дослідження великих масивів цифрової лінгвістичної інформації та у подальшому використанні для досліджень можливостей штучного інтелекту.

Мета статті — презентація основних методів дослідження цифрової лінгвістичної інформації та елементів теорії й експерименту у середовищі штучних нейронних мереж з використанням моделі узагальнених семантичних станів.

Українським мовно-інформаційним фондом НАН України розроблено теоретичні та науково-технічні засади цифрових методів дослідження, які довели необхідність та доцільність використання; практичне застосування продемонструвало високу ефективність створених програмних продуктів та лінгвістичних платформ, які активно впроваджуються у загальноукраїнський та європейський простір. На сучасному етапі мовознавчих досліджень **метод штучного інтелекту (ШІ)** виділяємо як один із основних та найбільш перспективних. Водночас інноваційні методи дослідження не суперечать вже існуючим класичним методам та мають бути використані лише у комплексі.

Лінгвістичні платформи для оброблення великих масивів даних, які розроблені в УМІФ на основі цифрових технологій та ШІ: *«Словники України online»*; *Тлумачний словник української мови у 20 томах Система лінгвістичної взаємодії «ВЛЛ»*; *«Український національний лінгвістичний корпус»*; *Leksykon aktywnej*

frazeologii polskiej i ukraińskiej; *Трансдисциплінарний кластер знань про коронавірусну інфекцію*; *Довідник АТО*; *Онтологічне середовище дослідження життя і творчості Тараса Шевченка*; *Онтологічне середовище «Музей НАН України»*; *Педагогічно-меморіальний музей Сухомлинського Василя Олександровича*; *Цифровий портрет Олеся Терентійовича Гончара* та інші.

У 2023–2024 роках фахівці УМІФ НАН України спільно з фахівцями НЦ МАН України за підтримки ЮНЕСКО, а також Міністерства освіти і науки України та Міністерства закордонних справ України здійснювали наукову роботу з оцінювання шкоди науковій інфраструктурі в Україні через війну. В результаті було розроблено спеціальну мовно-інформаційну аналітичну платформу https://polyhedron.ulif.org.ua/en/destroyed-property/ (Стрижак О. Є., Надутенко М. В., Довгий С. О., Широков В. А. та ін. Результати зазначеного аналізу (https://unesdoc.unesco.org/ark:/48223/pf0000388803) були представлені на Симпозіумі «Rebuilding Scientific Ecosystem in Ukraine» у березні 2024 р. План слугуватиме дорожньою картою для майбутніх дій і буде наданий відповідним зацікавленим сторонам для мобілізації підтримки та ресурсів для його впровадження (https://www.unesco.org/en/articles/unesco-outlines-recovery-plan-ukrainian-science).

Виділено основні методи дослідження, які використовують цифрові технології: статистичний метод оброблення інформації; метод корпусних технологій; лексикографічний; **штучний інтелект**: машинне навчання, глибоке навчання, нейромережі. Останніми роками розвиток технологій уможливив значний прогрес у сфері оброблення природної мови (NLP). **Основою сучасного NLP є машинне та глибоке навчання.** Ці методи дозволяють комп'ютерам навчатися на основі великих наборів даних і виконувати такі завдання: моделювання мови; класифікація тексту та аналіз тональності текстів (sentiment analysis)

із високою точністю. Алгоритми машинного навчання, зокрема **глибокі нейронні мережі,** показали надзвичайний успіх у завданнях NLP, перевершуючи традиційні підходи, основані на правилах. На нашу думку, останні події в галузі лінгвістичних технологій, штучного інтелекту, що базується на великих лінгвістичних моделях (Large Language Models — LLM), спонукають до побудови формалізованих моделей якомога ширшого класу мовних одиниць задля ефективної рецепції мовних феноменів засобами штучних нейронних мереж. В УМІФ НАН України на основі аналізу значного за обсягом текстового матеріалу було виявлено понад 170 граматичних моделей колокацій довжиною до дев'яти. На сьогодні продовжується робота із розбудови алгоритмічно-програмної моделі, налаштованої на моделювання автоматичного пошуку та ідентифікації колокацій у тексті. Зазначена робота виконується із застосуванням технології і даних Українського національного лінгвістичного корпусу (УНЛК), Віртуальної лексикографічної лабораторії (ВЛЛ) СУМ-20 із застосуванням методу лексикографічних середовищ та агентів, реалізованих в GPT-технології методами штучних нейронних мереж. Програмні агенти, які в даному випадку здійснюють розпізнавальні функції, фактично і забезпечують маркування тексту за наявними в ньому колокаціями, приписуючи останнім значення їхніх граматичних станів згідно з викладеною вище загальною теоретичною схемою. Наведемо деякі приклади результатів проведених експериментів.

У результаті ***видобування словосполучень із 774 статей Кримінального кодексу України*** було виокремлено 11 610 словосполучень. В результаті їх семантичного аналізу, порівняння на еквівалентність та об'єднання з використанням когнітивної ІТ-платформи «ПОЛІЕДР» було виокремлено 868 юридичних понять і термінів проєкту Кримінального кодексу України.

Examples of extracted collocations for certain articles of the draft Criminal Code of Ukraine

Original text	Extracted phrases:
Стаття 1.1.3. Презумпція знання та стабільність Кримінального кодексу 1. В Україні діє презумпція знання положень Кримінального кодексу. 2. З метою забезпечення знання положень та стабільності Кримінального кодексу: 1) цей Кодекс та закони про внесення змін до нього офіційно оприлюднюються в порядку, встановленому законом; 2) зміни до цього Кодексу вносяться виключно шляхом прийняття окремого закону про внесення змін до цього Кодексу, який також може містити положення про зміни до Кримінального процесуального кодексу, Кримінально-виконавчого кодексу або Кодексу України про адміністративні	1.презумпція знання положень Кримінального кодексу 2.забезпечення знання положень та стабільності Кримінального кодексу 3.офіційно оприлюднювати в порядку, встановленому законом 4.зміни до Кримінального кодексу 5.прийняття окремого закону про внесення змін до Кримінального кодексу 6.скасування кримінальної відповідальності за діяння 7.поліпшення правового статусу особи 8.кримінальне правопорушення 9.розмір розрахункової одиниці 10.надзвичайний стан 11.особливий період 12.закон про внесення змін до Кримінального кодексу

Стаття 1.2.8. Сумлінне виконання міжнародних зобов'язань

1. Цей Кодекс має відповідати чинним міжнародним договорам, згоду на обов'язковість яких надано Верховною Радою України.

2. Якщо для надання згоди на обов'язковість для України подається міжнародний договір, виконання якого потребує змін до цього Кодексу, то проєкт закону про внесення таких змін подається на розгляд Верховної Ради України разом з проєктом закону про ратифікацію (приєднання) і приймається безпосередньо перед прийняттям закону про ратифікацію (приєднання).

3. Цей Кодекс застосовується з урахуванням практики Європейського суду з прав людини.

Виокремлені словосполучення:

1. Сумлінне виконання міжнародних зобов'язань
2. Відповідність Кримінального кодексу чинним міжнародним договорам
3. Згода на обов'язковість міжнародних договорів
4. Верховна Рада України
5. Міжнародний договір
6. Виконання міжнародного договору
7. Проєкт закону про внесення змін до Кримінального кодексу
8. Ратифікація міжнародного договору
9. Приєднання до міжнародного договору
10. Практика Європейського суду з прав людини
11. Розгляд Верховною Радою України
12. Прийняття закону про ратифікацію (приєднання)
13. Проєкт закону про ратифікацію (приєднання)
14. Надходження міжнародного договору
15. Обов'язковість міжнародного договору для України
16. Зміни до Кримінального кодексу
17. Безпосереднє прийняття закону про внесення змін до Кримінального кодексу
18. Урахування практики Європейського суду з прав людини при застосуванні Кримінального кодексу.

Як показали результати експерименту, отримана модель має здатність перефразовувати окремі словосполучення в прямому порядку слідування слів.

Семантичний стан такого словосполучення залишається однаковим. Наприклад, у проєкті ККУ немає фрази «Урахування практики Європейського суду з прав людини при застосуванні Кримінального кодексу». Це модифікована колокація з п. 3 статті 1.2.8: «Цей Кодекс застосовується з урахуванням практики Європейського суду з прав людини».

Модель Mixtral-8x7B-UNLC-UA — одна з перших українських великих мовних моделей призначена для внутрішніх експериментів Українського мовно-інформаційного фонду НАН України та демонстрації того, що базову модель можна налаштувати для української мови. Ця модель не має механізмів модерації. Роботу в напрямі розвитку подібних моделей буде продовжено. У подальшому буде досліджено шляхи та засоби ефективного впровадження модерувальних обмежень моделі.

Список використаних джерел

1. Сучасні інформаційні технології в лінгвістиці : навчальний посібник / С. В. Петрасова, Н. Ф. Хайрова. Харків : ФОП Панов А.М., 2020. 124 с.
2. Загнитко А. Теория грамматических и текстовых структур : [монография]. Berlin : Lambert Academic Publishing, 2018. 634 с.
3. Загнітко А. Теорії лінгвістичних учень. Вінниця : ТОВ «ТВОРИ», 2019. 528 с.
4. Семеног О. М., Надутенко М. В. Платформа «Медіа&капсули» як засіб підвищення рівня медіакультури. *Академічні студії. Серія «Гуманітарні науки»*. 2022. № 2. С. 39–50. https://doi.org/10.52726/as.humanities/2022.2.6.
5. Широков В. А., Надутенко М. В., Стрижак О. Є., Ющенко С. С. Технологічні засади логіко-лінгвістичних досліджень законодавства. *Науково-технічний журнал «Біоніка інтелекту»*. 2020. Том 2. № 95. https://doi.org/10.30837/bi.2020.2(95).01.
6. Надутенко М. В. Створення трансдисциплінарних кластерів знань на основі лінгвістичних технологій платформи «ПОЛІЕДР» («POLYHEDRON») з елементами штучного інтелекту. *Матеріали Міжнародної наукової конференції «Актуальні питання сучасної лінгвістики»*. 2021. URL: http://ekmair.ukma.edu.ua/ handle/123456789/19709.

7. Лучик А. А., Надутенко М. В. Класифікація методів наукових досліджень у працях С. І. Дорошенка на тлі сучасної лінгвістичної науки. *Український світ у наукових парадигмах* : збірник наукових праць Харківського національного педагогічного університету імені Г. С. Сковороди. Вип. 11. Харків : ХНПУ; ХІФТ. 2024. С. 20–23.