DOI https://doi.org/10.30525/978-9934-26-277-7-172

DIGITAL TECHNOLOGIES IN LINGUISTIC RESEARCH

ЦИФРОВІ ТЕХНОЛОГІЇ У ЛІНГВІСТИЧНИХ ДОСЛІДЖЕННЯХ

Nadutenko M. V.
Candidate of Philological Sciences, Senior Researcher at the Linguistics Department Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine Kyiv, Ukraine

Надутенко М. В.
кандидат філологічних наук, старший науковий співробітник Український мовно-інформаційний фонд НАН України м. Київ, Україна

Nadutenko M. V.
Candidate of Technical Sciences, Head of the Information Department Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine Kyiv, Ukraine

Надутенко М. В.
кандидат технічних наук, завідувач відділу інформатики Український мовно-інформаційний фонд НАН України м. Київ, Україна

Збільшення обсягів накопиченої лінгвістичної інформації та висока швидкість надходження нових даних зумовлюють потребу в високоефективних цифрових технологіях та інтелектуальних системах пошуку нових способів автоматичної обробки інформації та презентації за вимогою користувача.

Актуальність проблеми визначається необхідністю застосування цифрових технологій для комплексного аналізу лінгвістичного матеріалу на всіх етапах роботи з текстом: цифровізація контенту, автоматична обробка, систематизація, класифікація та відповідна презентація великих масивів даних із забезпеченням доступності набутих знань для широкого кола користувачів.

Наукова новизна дослідження полягає у розробці комплексу методів для дослідження великих масивів цифрової лінгвістичної інформації із застосуванням можливостей цифрових технологій та подальшого впровадження.

Мета статті − презентація основних методів дослідження лінгвістичної інформації та програмних продуктів, які використовують цифрові технології.

Українським мовно-інформаційним фондом НАН України (УМІФ НАН України) розроблено теоретичні та науково-технічні

засади цифрових методів дослідження, які довели необхідність та доцільність використання. Їхнє практичне застосування продемонструвало високу ефективність кінцевих мультимедійних словникових продуктів та лінгвістичних платформ, які активно впроваджуються у загальноукраїнський та європейський простір.

З метою збереження цифрових ресурсів УМІФ НАН України та їх подальшого використання керівник відділу інформатики УМІФ НАН України (***Надутенко М. В.***) спільно із зарубіжними колегами (***Eugen Streichert***, Німеччина) забезпечив перенесення потужностей на закордонні сервери.

Лінгвістичні платформи для обробки великих масивів даних − **Лінгвістичні платформи УМІФ НАН України**:

− *«Словники України online»*
− *Тлумачний словник української мови у 20 томах*
− *Система лінгвістичної взаємодії «ВЛЛ»*
− *«Український національний лінгвістичний корпус»*
− *Leksykon aktywnej frazeologii polskiej i ukraińskiej*
− *Кореневий реєстр сайтів-кластерів*
− *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*
− *Науково-теоретичний журнал «Мовознавство»*

Режим доступу: https://svc2.ulif.org.ua/

Виділено основні методи дослідження лінгвістичної інформації, які використовують цифрові технології:

1) **статистичний метод обробки інформації** (математичний);
2) **метод корпусних технологій**;
3) **лексикографічний**;
4) **штучний інтелект**: машинне навчання, глибоке навчання, нейромережі.

На сучасному етапі мовознавчих досліджень **метод штучного інтелекту** виділяємо як один із основних та найбільш перспективних. Водночас інноваційні методи дослідження не суперечать вже існуючим класичним методам та мають бути використані у комплексі.

Нами виділено екстракцію та обробку неструктурованої та слабкоструктурованої концептографічної інформації з розподілених текстових масивів, які виконуються програмною системою. Ідея використання у дослідженнях полягає в **інтеграції методів лінгвістичної онтології та концептографії до технологій корпусної лінгвістики**, проте її програмній реалізації також приділено увагу. **Програмна система** пристосована для представлення інформації, що безпосередньо сприймається людиною, оскільки містить виключно модулі підсистеми структуризації текстів. ***Підсистема попередньої структуризації*** дозволяє розширити множину форматів оброблюваних текстових файлів, отже, зменшує часові та трудові витрати на їх конвертацію. ***Редактор онтологій*** дозволяє виконувати перегляд та редагування результатів структуризації вхідних текстів. ***Веб-орієнтований інтерфейс представлення інформації*** дозволяє представлення концептографічної інформації (для підсистеми трансдисциплінарного представлення інформації доступний модуль перегляду та модулі інтеграції з ArcGIS API for JavaScript, Leaflet.js, Google Maps) [1; 2]. Бібліотека онтологій забезпечує зберігання результатів та правил структуризації. ***Керуючий інтерфейс*** виконує керування підсистемою структуризації текстів. ***Підсистема експорту онтологій*** реалізовує експорт онтологій в потрібні формати представлення даних. ***Аналітичні підсистеми*** забезпечують виконання різноманітних операцій з ідентифікованими даними (прогнозування, багатокритеріальна оптимізація, робота з файлами CSV, OWL та ін.).

Для обробки інформації застосовано ***суфіксні дерева***, ***метод тезаурусу*** з полімовною синонімічною зоною, ***метод шинглів*** (shingles), Bag of Words, N-грамний метод та дистрибутивна семантика. При індексуванні за словами використовуються елементи семантичного індексування та індексування на основі визначення метрики подібності – функції відстані між двома словами, що дозволяють оцінити ступінь їхньої подібності в даному контексті.

Результатом роботи зазначеного набору алгоритмів на основі використання нейромережі прямого розповсюдження та методу негативного семплування (Negative Sampling) є ***векторна модель слів лінгвістичного корпусу***, що використовується для машинного навчання.

Література:

1. ArcGIS API for JavaScript. URL: https://developers.arcgis.com/javascript/ (дата звернення: 24.05.2018).

2. Google Maps APIs Google Developers. URL: https://developers.google.com/maps/?hl=ru (дата звернення: 24.05.2018).

3. Leaflet – an open-source JavaScript library for interactive maps. URL: http://leafletjs.com (дата звернення: 11.11.2019).

4. Mintser O. P., Babintseva L. Yu., Zaliskyi V. M., Nadutenko M. V., Kharchenko N. V., Ladychuk O. K. Theoretical approaches to the creation of systemic biomedicine (based on the materials of the report on SRW «System-Biological And System-Medical Regularities Of Development And Course Of Ischemic Heart Disease») . In: Medical Informatics and Engineering. vol 4. Pp. 16–72. 2021. DOI: https://doi.org/10.11603/mie.1996-1960.2020.4.11889

5. 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

6. 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. 97–114 с. DOI: https://doi.org/10.1007/978-3-030-80126-7_8

7. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков, М. В. Надутенко та ін. Т. 5 : Віртуалізація лінгвістичних технологій. Київ : УМІФ НАН України. 2018. 289 с.

8. Широков В. А., Загнітко А. П., Надутенко Максим, Надутенко Маргарита та ін. Віртуальна лексикографічна лабораторія «Мультимедійний словник з інфомедійної грамотності» [Електронний ресурс] // Український мовно-інформаційний фонд НАН України, створено в рамках грантової програми «МЕДІА&ВЧИТЕЛЬський кампус» проєкту «Вивчай та розрізняй: інфомедійна грамотність», IREX (Рада наукових досліджень та обмінів) за підтримки Посольств Великої Британії та США у партнерстві з Міністерством освіти і науки України та Академією української преси, 2020–2021. URL: https://lcorp.ulif.org.ua/InfoMediaVLL/ (дата звернення: 10.11.2022).