УДК 811.161.2'282 : 004.85

**Д-р. ф, ст. викладач Жук І. С., асистент Громова В.В.,
магістрант Жмурко К.Р.**

**Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»**

КОНЦЕПТУАЛЬНА МОДЕЛЬ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ ДІАЛЕКТІВ УКРАЇНСЬКОЇ МОВИ

Abstract

**Ivan Zhuk, Ph.D., senior lecturer; assistant Gromova V.V.,
Kseniia Zhmurko, masters-student;**

Conceptual model and system for researching language dialects of the Ukrainian language.
The work presents a conceptual model and a prototype of a system for the dialectal analysis of Ukrainian texts. It combines the business process modeling approach (Eriksson-Penker profile) for architecture formalization with a modern NLP pipeline based on BERT embeddings and clustering/classification methods. The paper highlights the advantages of the proposed component architecture, which standardizes interfaces and links corpus and model versions to ensure full reproducibility. The suitability of the system for flexible and scalable research is analyzed, laying the groundwork for the further development of Ukrainian computational dialectology.

Вступ

Дослідження діалектних особливостей української мови є міждисциплінарним і набуває актуальності завдяки розвитку NLP. Українська, як low-resource, нині отримує цифрові інструменти — корпуси BRUK та UberText 2.0 [1].

Сучасні підходи застосовують трансформерні моделі (BERT), що перевершують традиційні методи, працюючи з текстовими ембеддінгами та подальшою кластеризацією. У роботі аналіз виконується на художніх творах: літературні тексти відображають відносно стабільні стилістичні й регіональні маркери, релевантні для виявлення діалектних рис.

Джерелом даних слугує регіонально анотований корпус української мови ГРАК; формується підкорпус художньої літератури з прив'язкою до метаданих автора та публікації [2].

Постановка задачі

Предметом дослідження є концептуальна модель і прототип багатокомпонентної системи обробки й аналізу діалектних контекстів у великих корпусах українських текстів.

Метою дослідження є визначення структурної архітектури системи, уніфікованих інтерфейсів компонентів, логіки обробки корпусу, методики побудови векторних подань, вибору оптимальних алгоритмів аналізу та встановлення принципів відтворюваності результатів через фіксоване версіонування даних та моделей у рамках корпусних релізів і Model Registry.

Модель NLP-системи на основі бізнес-профілю Еріксона–Пенкера

Нормативне окреслення та зв’язування понять «концептуальна модель» і «комп’ютерне моделювання» визначає основу архітектурного проєктування концептуальної моделі NLP-системи діалектного аналізу.

Для розробки такої моделі застосовується бізнес-профіль Еріксона–Пенкера, що визначає класи сутностей: проблему, мету, ресурси, процеси, бізнес-правила та їх взаємозв'язки.

Класи бізнес-профіля розробляються для вирішення завдань інженерії системи діалектного аналізу:

Проблема: Необхідність створення автоматизованої системи, здатної аналізувати українські тексти та ідентифікувати їхні діалектні ознаки для подальшої класифікації.

Мета: Автоматизація класифікації текстових даних за основними наріччями української мови (Північне, Південно-східне, Південно-західне).

Ресурси: Корпус діалектних текстів, метадані (зокрема місце народження авторів), попередньо навчені NLP-моделі (наприклад, BERT), програмні бібліотеки для NLP та машинного навчання, обчислювальні потужності. [3]

Процеси: Визначені послідовності дій, що забезпечують функціонування системи, такі як збір та попередня обробка текстів, отримання векторних представлень (ембедингів), навчання моделі класифікації (supervised learning) та/або кластеризація текстів.

Бізнес-правило: Формальні інструкції, що регулюють процеси. Наприклад: критерії валідації та включення текстів до корпусу, правила співвіднесення географічних метаданих автора з цільовими діалектними групами, або порогові значення для впевненості класифікатора.

Модель NLP-системи з дослідження діалектів України.

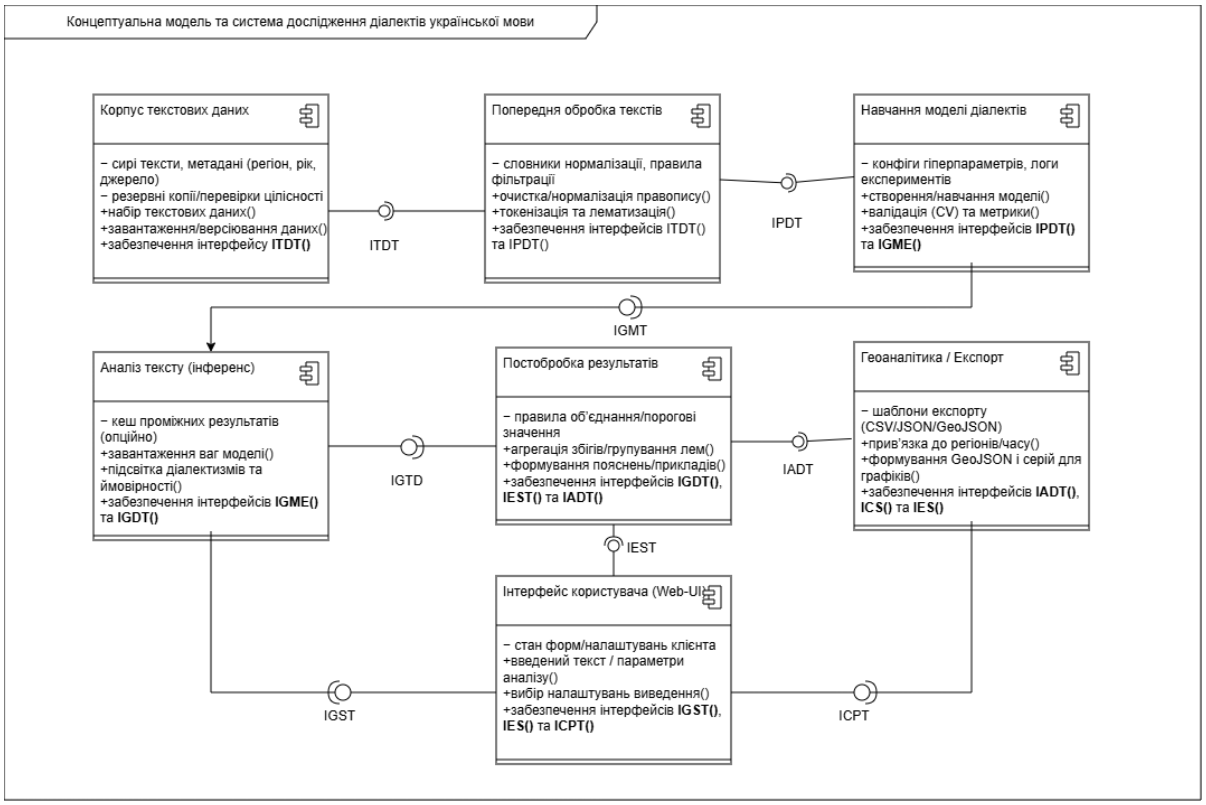

Рис.1. Модель системи дослідження мовних діалектів української мови. Діаграма компонентів в нотації UML. [6]

NLP-ядро системи реалізує чітку послідовність операцій: підготовка тексту, побудова векторного подання, проведення групування через кластеризацію або класифікацію, інтерпретація результатів з виділенням діалектизмів та геовізуалізація отриманих даних. Кожна операція виконується через стандартизовані контракти інтерфейсів між компонентами системи.

Компонент корпус текстових даних (ITDT) зберігає сирі тексти та метадані з версіонуванням і контролем цілісності, надає метод get_corpus, який повертає тексти і метадані для наступних етапів.

Компонент попередня обробка (IPDT) виконує нормалізацію, очистку, токенізацію та лематизацію, приймає тексти з корпусу і повертає підготовлені тексти для навчання.

Компонент навчання моделі (IGME) керує гіпер параметрами, тренує і валідовує моделі, зберігає ваги та ембеддінги і надає до них доступ іншим компонентам. Інференс (IGME, IGDT, IGST) завантажує навчені ваги, обчислює діалектні мітки, ймовірності та виділені фрагменти, підтримує кеш і працює з наборами документів за запитом.

Компонент постобробка результатів (IADT, IEST) агрегує результати, групує леми, формує пояснення і оновлює корпус або сховище ембеддінгів відповідними записами.

Компонент геоаналітика та візуалізація (ICS, ICPT) прив'язує результати до регіонів і часових інтервалів, будує графіки та інтерактивні мапи для інтерфейсу користувача. Експорт і API сервіси (IES) надають інтерфейси для експорту у формати JSON, CSV, GeoJSON та для отримання результатів інференсу клієнтськими застосунками.

Компонент вебінтерфейс і API Gateway (IGST, IEST, ICPT) приймає параметри запуску та постобробки, маршрутизує запити до внутрішніх сервісів і керує варіантами представлення результатів. Відтворюваність і реєстр моделей забезпечують співвіднесення ідентифікаторів релізів корпусу та моделей, зберігають контрольні суми, фіксовані спліти і повну історію навчання для повторення експериментів.

Висновки

1. У статті запропоновано представлення моделі системи дослідження українських діалектів як множини сутностей і відношень між ними (джерело, документ, підготовлений текст, embedding, кластер, реліз корпусу, артефакт моделі), а сам процес оброблення – як RAG-конвеєр над діалектним корпусом з контролем якості даних.
2. Розроблено концептуальну модель NLP-системи, основану на модифікованому бізнес-профілі Еріксона–Пенкера, що забезпечує метарівневе подання архітектури та узгоджує дані, моделі й сервіси інференсу.
3. Побудовано динамічне та структурне представлення у вигляді діаграми компонентів і діаграми діяльності, які відображають внутрішню структуру, реалізацію та інтерфейси модулів (передобробка, ембеддинги, кластеризація/класифікація, геоаналітика, візуалізація).

Перспективи подальших досліджень пов'язані з розробкою пошуково-розміркового алгоритму для виявлення діалектних рис і його імплементацією на множині відкритих текстових ресурсів України, а також із розширенням корпусу, активним навчанням і вдосконаленням геопросторової валідації.

Література

1. Chaplynskyi D. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale // Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). Dubrovnik, Croatia : Association for

Computational Linguistics, 2023. — P. 1–10. — DOI: 10.18653/v1/2023.unlp-1.1. — URL: https://aclanthology.org/2023.unlp-1.1/
2. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко, Т. Ніколаєнко, А. Лукашевський та ін. — Київ ; Львів ; Єна, 2017–2025 : електрон. ресурс. — URL: https://uacorpus.org/
3. Livinska H., Makarevych O. Feasibility of Improving BERT for Linguistic Prediction on Ukrainian Corpus // Computational Linguistics and Intelligent Systems (COLINS 2020): Proceedings of the 4th International Conference. Volume I: Main Conference. Lviv, Ukraine, April 23–24, 2020. CEUR Workshop Proceedings, Vol. 2604. P. 552–561. URL: https://ceur-ws.org/Vol-2604/paper40.pdf
4. Arthur D., Vassilvitskii S. k-means++: The Advantages of Careful Seeding // Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA). 2007. P. 1027–1035. URL: https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf
5. Guo C., Pleiss G., Sun Y., Weinberger K.Q. On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning, ICML 2017. 2017. P. 1321–1330. URL: https://arxiv.org/abs/1706.04599
6. Павловська К. І. Концептуальна модель та система «Текст-зображення» : магістерська дис. : 113 Прикладна математика / Нац. техн. ун-т України «Київ. політехн. ін-т ім. І. Сікорського». – Київ, 2024. – 113 с.