

УДК 004.4

Фоменко В. А., здобувач
Савицький Р. С., аспірант, ст. викладач
Державний університет «Житомирська політехніка»

АВТОМАТИЗОВАНИЙ ПОШУК ПОМИЛОК В УКРАЇНСЬКОМУ ТЕКСТІ З ВИКОРИСТАННЯМ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ

Якість текстового контенту українською мовою стає все важливішою в умовах цифровізації суспільства. Дослідження 2023 року показало, що близько 67% україномовного контенту в інтернеті містить орфографічні, пунктуаційні або граматичні помилки [1]. Українська мова зі своєю складною морфологією, системою відмінків та узгодження слів створює багато можливостей для виникнення помилок, які традиційні системи перевірки не можуть ефективно виявляти.

Словникові методи є найпростішим підходом до перевірки текстів. Вони порівнюють кожне слово зі списком правильних слів, але мають критичні обмеження для української мови.

Основна проблема використання словників у валідації тексту це те, що словники не розуміють контексту речення. Візьмемо для прикладу речення "Він одів пальто". Слово "одів" існує у словнику і є правильною формою дієслова "одіти". Словникова система визнає таке речення правильним. Але правила української мови вимагають тут використання "одягнув" (коли хтось одягає щось на себе), а не "одів" (коли хтось одягає когось іншого). Словник не може цього побачити, тому що аналізує слова окремо одне від одного [2].

Слова "адресат" і "адресант", "абонент" і "абонемент" всі є правильними. Словникова система не може визначити, яке з них підходить у конкретному контексті. Слово "замок" може означати споруду або пристрій для зачинення дверей, але словник не знає, який варіант потрібен у конкретному реченні.

Словоформи створюють додаткові труднощі десятки різних форм. Іменник змінюється за семи відмінками і числами, дієслово змінюється за особами, числами, часами та способами. Для української мови це означає мільйони словоформ, які практично неможливо підтримувати в актуальному стані.

Словникові методи можуть виявити тільки найпримітивніші помилки, коли слово взагалі не існує. Вони безсилі перед контекстними помилками, проблемами узгодження та паронімами, які становлять більшість помилок у реальних текстах.

У роботі використовується модель Pravopysnyk/best-unlp, яку розробили спеціально для української мови. Модель базується на архітектурі Transformers і її навчили на великих обсягах українських текстів. На відміну від словникових методів, модель аналізує весь контекст речення та розуміє зв'язки між словами.

Механізм уваги дозволяє моделі одночасно враховувати всі слова у реченні та їхні взаємозв'язки. Коли модель аналізує слово "одів" у реченні "Він одів пальто", вона бачить, що людина одягає щось на себе, а значить потрібне дієслово "одягнув" [3].

Модель Pravopysnyk/best-unlp навчили розпізнавати типові помилки в українській мові. Це неправильне вживання дієслів одіти та одягнути, помилки в узгодженні прикметників з іменниками, неправильне відмінювання, порушення керування дієслів та пунктуаційні помилки. Модель не просто шукає слова у словнику, вона розуміє граматичну структуру та семантику тексту.

Система виявляє орфографічні помилки (друкарські помилки та порушення правил написання), граматичні помилки (неправильне узгодження слів за родом, числом і відмінком), пунктуаційні помилки (неправильне використання розділових знаків) та лексичні помилки (невідповідне використання слів і плутанина паронімів).

Ефективний пошук помилок в українському тексті неможливий без використання штучного інтелекту. Словникові методи застаріли і не відповідають вимогам сучасної обробки природної мови. Тільки моделі на основі Transformers можуть розуміти контекст та складні граматичні залежності української мови, що забезпечує якісне виявлення помилок у текстах.

Список використаних джерел:

1. Romanyshyn M., Chaplynskyi D. Ukrainian Language Corpus and Error Detection: Current State and Perspectives. Proceedings of the 4th Ukrainian NLP Workshop, 2023. DOI: 10.18653/v1/2023.unlp-1.12
2. Sang-Bum Kim, Hee-Soo Hahn, Changki Lee. Context-sensitive spelling correction using Google Web 1T 5-gram information. IEEE Transactions on Audio, Speech, and Language Processing, 2010. DOI: 10.1109/ICIG.2009.39
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.. Attention is All you Need. Neural Information Processing Systems, 30, 5998–6008? 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>