Микола Вертман
Магістрант спеціальності комп'ютерні системи та мережі, Національний університет «Запорізька політехніка» м. Запоріжжя, Україна
0009-0009-4810-6270
wertmanick1997@gmail.com

Степан Скрупський
к.т.н., доцент, доцент кафедри комп'ютерних систем та мереж
Місце роботи: Національний університет «Запорізька політехніка», м. Запоріжжя, Україна
0000-0002-9437-9095
sskrupsky@gmail.com

ПІДХІД ДО ПЕРЕВІРКИ ТА КОРИГУВАННЯ ПОМИЛОК В ТЕКСТІ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. Зростання обсягу цифрового текстового контенту підвищує вимоги до якості письма українською. Існуючі системи перевірки граматики (наприклад, Grammarly, LanguageTool) не повною мірою враховують морфологічні та синтаксичні особливості української мови, що призводить до неточних виправлень і пропуску специфічних помилок. У роботі представлено дослідження комп'ютерної системи граматичної корекції українського тексту з використанням технологій штучного інтелекту, завдяки якому перевіряється можливість до підвищення ефективності виявлення і виправлення помилок шляхом поєднання спеціалізованого україномовного корпусу, сучасних архітектур та багатокритеріального оцінювання.

Ключові слова: граматична корекція, українська мова, штучний інтелект, великі мовні моделі, точність, повнота, $F_{0.5}$-міра, GPT-3, GPT-4.

1. ВСТУП

Якість письмової комунікації українською мовою набуває критичного значення в умовах стрімкої цифровізації суспільства. Ручне вичитування текстів на наявність орфографічних, граматичних та стилістичних помилок є трудомістким процесом, що вимагає значного часу і високої кваліфікації. Це обумовлює актуальність розроблення інтелектуальних систем автоматизованої перевірки тексту. Глобальні онлайн-сервіси, як-от Grammarly чи LanguageTool, хоча й успішно застосовують методи штучного інтелекту, не завжди точно обробляють українську мову через її унікальні лінгвістичні особливості. Унаслідок цього виникають випадки некоректних виправлень або пропуску специфічних помилок, характерних для української граматики. Нещодавній прогрес у галузі Natural Language Processing (NLP) і поява великих мовних моделей (LLM) відкрили нові можливості для розв'язання задачі граматичної корекції тексту. Такі моделі, навчені на великих корпусах даних, продемонстрували вражаючу здатність генерувати текст і виправляти помилки в різних мовах. Однак, для максимального ефекту в українському контексті необхідно спеціалізоване навчання цих моделей на україномовних даних.

Мета публікації. Підвищити ефективність перевірки та коригування помилок в тексті за рахунок запропонованого підходу з використанням ШІ.

2. АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

Проблематика автоматичного виправлення граматичних помилок (Grammatical Error Correction, GEC) інтенсивно досліджується для англійської та інших мов світу. Традиційно застосовувались статистичні підходи та правилові методи, але наразі

домінують нейромережеві sequence-to-sequence моделі, які можуть прямо «переписувати» речення без помилок [1]. У той же час з'являються нові парадигми sequence-to-edit, які зосереджуються на точковому виправленні помилок із мінімальними змінами [1]. В англомовному GEC-аналізі накопичено значний досвід, проте для української мови кількість досліджень обмежена. Окремі роботи з'явилися останнім часом, наприклад спроби створення систем на основі машинного навчання для українських текстів [2]. Також на ринку існують комерційні рішення (Grammarly), які підтримують українську, але їх алгоритми є закритими; відкриті інструменти (LanguageTool) мають базову підтримку української, проте поступаються якістю перевірки.

Важливим аспектом сучасних досліджень є вдосконалення метрик оцінки якості GEC. Стандартна метрика MaxMatch (M^2), що розраховує Precision, Recall та F-міру з ваговим коефіцієнтом 0.5, широко використовується для порівняння систем. Однак в літературі відзначено, що формальні метрики на кшталт M^2 або BLEU/GLEU можуть бути занадто «жорсткими» і не враховувати випадки, коли модель запропонувала граматично правильне виправлення, еквівалентне за змістом еталонному, але не дослівно співпадає з ним [3]. Це спонукало дослідників до пошуку наближених до людської оцінок метрик. Зокрема, у роботі [3] запропоновано переглянути підходи до оцінювання GEC, наголошуючи на важливості семантичної адекватності виправлень. Такі ідеї лягли в основу нашого підходу із використанням моделі-арбітра для оцінки якості виправлень.

3. МЕТОДИ ДОСЛІДЖЕННЯ

Для об'єктивної перевірки моделей створено спеціалізований корпус «Ukr-GEC-Bench» обсягом 500 речень із різними типами помилок. Щоб забезпечити репрезентативність, 60% прикладів згенеровано синтетично: до граматично правильних речень (взятих із новин та літературних джерел) навмисно внесено типові помилки (опечатки, пропуски розділових знаків, неправильні граматичні форми тощо) за допомогою скриптів [2]. Інші 40% набору становлять реальні дані – речення, зібрані з соцмереж, форумів для вивчення мови та студентських есе, які містять помилки, допущені реальними носіями [2]. Кожне речення має вручну виправлений еталон. Для аналізу всі помилки класифіковано за категоріями (орфографічні, морфологічні, синтаксичні, пунктуаційні, лексичні, стилістичні) та за критичністю: від низької (стилістичні огріхи, що не спотворюють розуміння) до високої (грубі граматичні помилки, які змінюють зміст речення). Наявність збалансованого і різноманітного датасету є ключовою передумовою коректної оцінки моделей.

В рамках експерименту протестовано три підходи до автоматичної корекції: - Multilingual T5 (fine-tuned) – нейромережева модель архітектури Encoder-Decoder (google/mt5-base), попередньо навчена на багатомовному корпусі, яку додатково донавчено спеціально на українському GEC-датасеті [4]. Цей варіант представляє підхід адаптації потужної багатомовної моделі під потреби конкретної мови. - Ukrainian-specific GPT (fine-tuned) – модель архітектури Decoder-Only середнього розміру (GPT-подібна). Використано відкриту україномовну модель mGPT-1.3B [5], яка з самого початку тренована на великих обсягах українського тексту, а в нашому експерименті теж була донавчена на тому ж спеціалізованому наборі. Це дозволяє оцінити, наскільки мовно-специфічна базова модель виграє в якості. Commercial GPT-4 (API, zero-shot) – модель GPT-4 від OpenAI, доступ до якої здійснювався через хмарний API. Вона застосована без додаткового навчання (zero-shot), але із спеціальним системним промптом, що інструктує модель діяти як коректор українського тексту. Цей

варіант демонструє можливості сучасної універсальної LLM без адаптації до українського контенту.

Для кількісного вимірювання ефективності моделей застосовано метрики, поширені в задачах GEC: M^2 та GLEU [3]. Метрика M^2 розраховує збалансовану оцінку на основі Precision та Recall з пріоритетом точності ($F_{0.5}$) і порівнює виправлення моделі з еталоном на рівні окремих правок. GLEU – модифікація BLEU, що враховує збіг коректних і некоректних n-грам при порівнянні виправленого речення з еталонним, краще відображаючи специфіку задачі [3]. Проте такі формальні метрики можуть не врахувати коректних перефразувань, якщо вони відрізняються від еталону. Тому додатково проведено якісну семантичну оцінку за допомогою «моделі-арбітра». У ролі арбітра використано GPT-4: йому подавалося оригінальне речення з помилкою, правильний еталон і виправлення, запропоноване тестованою моделлю. GPT-4 незалежно виставляв оцінку від 1 до 5, наскільки виправлений моделлю варіант є граматично правильним і зберігає зміст вихідного речення. Такий підхід наближає автоматичне оцінювання до людського та дозволяє кількісно виміряти семантичну якість виправлень.

4. ЗАПРОПОНОВАНИЙ АЛГОРИТМ

Розроблений підхід реалізований у вигляді веб-застосунку за клієнт-серверною схемою з окремим сервером штучного інтелекту.

Принцип роботи наступний: спочатку зчитується введений текст і ділиться на речення, після чого мовна модель генерує виправлення (за потреби зі стилістичними покращеннями). Зміни порівнюються з оригіналом, типізуються (граматика, орфографія, лексика/стиль), а для кожної помилки формується підказка з поясненням і варіантами, що відображаються у веб-інтерфейсі підсвічуваннями. Наприкінці система обчислює інтегральний показник граматичності й показує його користувачу.

Алгоритм, що реалізує запропонований підхід, наведено на рис. 1.

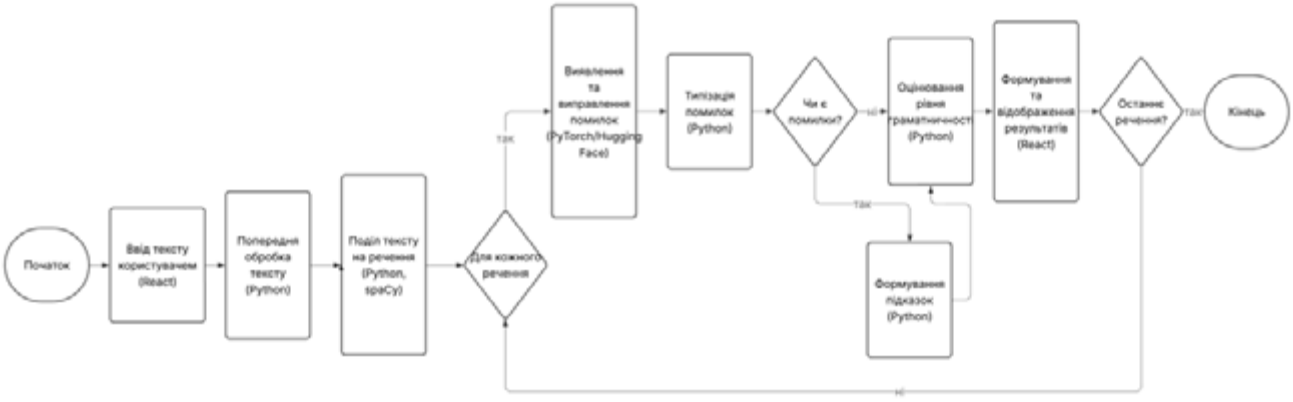

Рисунок 1. Алгоритм, що реалізує запропонований підхід до перевірки та коригування тексту з використанням штучного інтелекту

4. РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

В таблиці 1 наведено узагальнені показники трьох моделей за всіма критеріями.

Таблиця 1

Таблиця результатів моделі GEC

Модель	M^2 ($F_{0.5}$)	GLEU	Оцінка «моделі-арбітра» (середній бал)	Швидкість (мс/речення)	Вартість ($/1M токенів)
Multilingual T5 (fine-tuned)	0.68	0.65	4.5	150	~0 (локально)
Ukrainian-specific GPT (fine-tuned)	0.65	0.63	**4.7**	250	~0 (локально)
Commercial API (zero-shot)	0.59	0.61	4.6	1200	~15.00

Аналіз показників підтверджує суттєву перевагу спеціалізованих моделей, що пройшли навчання на україномовних даних. Модель T5 продемонструвала найвищу формальну точність виправлень: її M^2-оцінка ($F_{0.5}$≈0.68) найкраща серед трьох, що вказує на спроможність робити точні і мінімально необхідні правки. Також вона дещо перевершує інші моделі за GLEU (0.65). Модель mGPT (Український GPT) має ненабагато нижчі формальні метрики (M^2=0.65), але вирізняється найвищою семантичною якістю виправлень: за оцінкою «арбітра» вона отримала середній бал 4.7/5, що є найкращим результатом. Це означає, що виправлення від спеціалізованого GPT найбільш природні та стилістично доречні, майже не відрізняються від еталонних за змістом. Високі бали (понад 4.5) для всіх моделей свідчать про те, що жодна з моделей не спотворює початковий смисл речень – навіть GPT-4 у zero-shot режимі здебільшого пропонував змістовно правильні виправлення. GPT-4 (zero-shot), утім, суттєво поступився за формальними метриками: його $F_{0.5}$ (≈0.59) та GLEU (≈0.61) найнижчі. Це вказує на те, що без спеціалізованого навчання GPT-4 пропускає більше помилок (нижча Recall) і часом вносить зайві або неправильні правки (що знижує Precision), порівняно з меншими моделями, адаптованими до українських помилок.

Окрему увагу слід звернути на продуктивність та вартість використання моделей. Локально розгорнуті моделі, що навчалися в процесі (T5 та GPT 1.3B) працюють значно швидше: так, T5 обробляє речення в середньому за ~150 мс, що у 8 разів швидше за GPT-4 через API (~1200 мс на речення). Модель mGPT є трохи повільнішою (≈250 мс), ймовірно через більшу кількість параметрів, але все одно випереджає хмарний сервіс. З точки зору вартості, власні моделі після початкового розгортання не генерують прямих витрат на кожен запит, тоді як використання GPT-4 API потребує оплати (приблизно $15 за обробку 1 млн токенів, згідно з тарифами). Таким чином, спеціалізовані рішення не тільки перевершують GPT-4 за якістю виправлень, але й є вигіднішими в експлуатації (без мережевих затримок та додаткових витрат).

Отримані результати підтверджують висунуту гіпотезу: цілеспрямоване навчання моделей на україномовному корпусі помилок забезпечує вищу точність і повноту корекції, ніж використання навіть дуже потужної, але неадаптованої моделі загального призначення. Хоча GPT-4 демонструє видатні здібності у багатьох NLP-завданнях, у вузькій сфері GEC для української без fine-tuning він поступається меншим за розміром моделям. Це підкреслює критичну роль якісних даних для навчання: моделі, що вчаться на реальних прикладах українських помилок, краще їх розуміють і виправляють. Також результати вказують на доцільність розробки спеціалізованих GEC-рішень для мов зі

складною морфологією (як українська), оскільки універсальні багатомовні системи не досягають у них максимальної ефективності.

ВИСНОВКИ ТА ПЕРЕСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В роботі підвищено ефективність перевірки та коригування помилок в тексті, шляхом реалізації запропонованого підходу, завдяки спеціалізованим моделям, що навчалися на українському корпусі помилок. За результатами роботи можна побачити, що модель T5 забезпечила найкраще співвідношення точності та швидкодії, тоді як mGPT продемонструвала вищу стилістичну якість. Комерційна GPT-4 без додаткового навчання виявилася менш точною та менш економною.

Запропонований підхід формує основу для створення ефективної системи перевірки україномовного тексту.

ПОСИЛАННЯ

[1] P. M. A. Cano, et al., “Natural Language Processing of Grammar Checker Tools for Academic Writing: A Systematic Literature Review,” Journal of Electrical Systems, vol. 20, no. 7s, pp. 1151–1156, 2024.

[2] A. Sinitsa, “Generating Datasets for Grammatical Error Correction,” University of Twente, 2019.

[3] AI-Forever, “mGPT-1.3B-ukrainian”, [Online]. Available: https://huggingface.co/ai-forever/mGPT-1.3B-ukranian.

[4] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer,” in *Proc. NAACL-HLT*, Online, 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.

[5] S. Lin, “Evaluating LLMs’ grammatical error correction performance in learner Chinese: A corpus linguistic approach,” PLOS ONE, vol. 19, no. 10, Article e0312881, Oct. 2024.

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

NATIONAL UNIVERSITY OF LIFE AND ENVIRONMENTAL SCIENCES OF UKRAINE

FACULTY OF INFORMATION TECHNOLOGY

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

PROCEEDINGS

МАТЕРІАЛИ

XIII International scientific and practical conference

XIII Міжнародної науково-практичної конференції

GLOBAL AND REGIONAL PROBLEMS OF INFORMATIZATION IN SOCIETY AND NATURE USING '2025

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2025

13-14 November 2025

Kyiv, NULES of Ukraine

13-14 листопада 2025 року

Київ, НУБіП України

Kyiv 2025

Київ 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ
І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

МАТЕРІАЛИ

XIII Міжнародної науково-практичної конференції

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2025

13-14 листопада 2025 року

Київ, НУБіП України

Київ 2025

УДК 004

Рекомендовано до друку вченою радою факультету інформаційних технологій Національного університету біоресурсів і природокористування України (протокол № 4 від 18.12.2025).

Укладач: д.т.н., доцент Шкарупило В.В.

Збірник матеріалів XIII Міжнародної науково-практичної конференції "Глобальні та регіональні проблеми інформатизації в суспільстві і природокористуванні '2025", 13–14 листопада 2025 року, НУБіП України, Київ. – К.: НУБіП України, 2025. – 206 с.

Відповідальність за зміст публікацій несуть автори.