

РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ В УКРАЇНСЬКИХ ТЕКСТАХ В УМОВАХ ОБМЕЖЕНОЇ РОЗМІТКИ

Кашперова С.В.¹, Шаповал Н.В.²

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

¹ kashperova.sofia@lil.kpi.ua, ² shovgun@gmail.com

У даній роботі досліджено проблему розпізнавання іменованих сутностей (NER) в україномовних текстах в zero-shot та few-shot режимах. Метою дослідження є розробка компактної та ефективної моделі на основі архітектури GliNER, здатної підтримувати zero-shot та few-shot режими розпізнавання. Запропоновано модифікації базової архітектури, включаючи інтеграцію Post-Fusion блоку з Mixture-of-Experts (MoE), використання функції активації GoLU та оптимізатора Sophia-G для стабілізації навчання. Базовим текстовим енкодером обрано Snowflake Arctic-Embed 2.0-L. Створено український корпус few-shot NER на базі публічних джерел з використанням GPT-4o для анотації. Експериментальне порівняння підтвердило, що запропонована архітектура має ефективну продуктивність ($F1 = 0.7891$). Результатом є модель, що демонструє здатність до узагальнення, актуальну для малоресурсних мов.

Ключові слова: NER, векторні подання токенів, zero-shot та few-shot розпізнавання, суміш експертів, GoLU, Sophia-G, українська мова.

1. ВСТУП

Розпізнавання іменованих сутностей (NER) є фундаментальною підзадачею інтелектуального аналізу текстів, що полягає у виявленні фрагментів (сутностей) та їх віднесенні до визначених класів. Ця задача відіграє ключову роль у information extraction, семантичній анотації та побудові knowledge graphs. Сучасні системи NER вийшли за межі традиційних трьох типів сутностей (персони, локації, організації) і повинні працювати з широким спектром категорій, включаючи дати, грошові суми, продукти, події та інші.

Українська мова в задачі NER вважається малоресурсною, оскільки бракує великих анотованих корпусів для якісного навчання моделей. Станом на 2025 рік основним публічно доступним корпусом є NER-UK 2.0, який, незважаючи на якісну розмітку, обмежений у покритті доменів для навчання моделей в режимі few-shot розпізнавання. Це призводить до того, що більшість відкритих моделей для української мови натреновані на обмеженій кількості класів і не здатні узагальнюватися на нові типи сутностей без додаткового навчання [1].

З появою пропріетарних великих мовних моделей (LLM), таких як GPT, Claude та Gemini, стало можливим ефективно zero-shot та few-shot розпізнавання. Проте ці LLM мають мільярди параметрів, що робить їх надзвичайно дорогими та складними для локального розгортання в умовах обмеженої інфраструктури.

Актуальною задачею є розробка компактної моделі NER для української мови, яка забезпечує прийнятний баланс між якістю та ефективністю, підтримує режими zero-shot і few-

shot і дозволяє швидко та дешево розгортання. Для досягнення цієї мети було обрано архітектуру GLiNER як базову, оскільки вона реалізує end-to-end підхід, одночасно навчаючи представлення слів та типів, і природно підтримує zero-shot/few-shot режими через використання текстових описів типів сутностей.

2. АРХІТЕКТУРА ТА МОДИФІКАЦІЇ БАЗОВОЇ МОДЕЛІ

Базова архітектура GliNER була обрана через її end-to-end характер, що уникає проблеми поширення помилок, притаманної декомпозиційним методам. Окрім того, що модель треба було адаптувати для української мови, оскільки оригінальна GliNER навчалась для англійської та іспанської мов, було вирішено внести архітектурні модифікації для підвищення продуктивності та стабільності навчання. В якості базового енкодера використано Snowflake Arctic-Embed 2.0-L, який підтримує Matryoshka Representation Learning (MRL). Завдяки MRL перші k компонент ембедінга містять найбільш «важливу» інформацію, тож під час інференсу можна гнучко зменшувати розмірність (наприклад, до 256) зі слабкою втратою якості.

Для глибокої взаємодії між представленнями тексту та типів введено Post-Fusion блок з механізмом Mixture-of-Experts (MoE). Після базового енкодера послідовно застосовуються self-attention для токенів тексту, self-attention для типів і cross-attention «типи $\leftrightarrow$ текст», завдяки чому токени отримують інформацію про семантику доступних типів, а самі типи взаємодіють між собою і зближують споріднені класи. У середині цього блоку впроваджено механізм Mixture-of-Experts (MoE): замість одного фіксованого MLP використовується набір експертів, а операцію top- k (argmax) в роутері замінено на ReLU, щоб це була повністю диференційована операція. Це дозволяє різним експертам спеціалізуватися на різних типах сутностей чи мовних контекстах і при цьому уникати проблеми «нетренованих експертів», оскільки градієнт проходить через усіх активних експертів.

У більшості підмереж замість ReLU/GELU використано функцію активації GoLU (Gompertz Linear Unit), яка має асиметричну S-подібну криву та рідше входить у насичений режим, зберігаючи корисну інформацію у слабких активаціях, рис. 1.

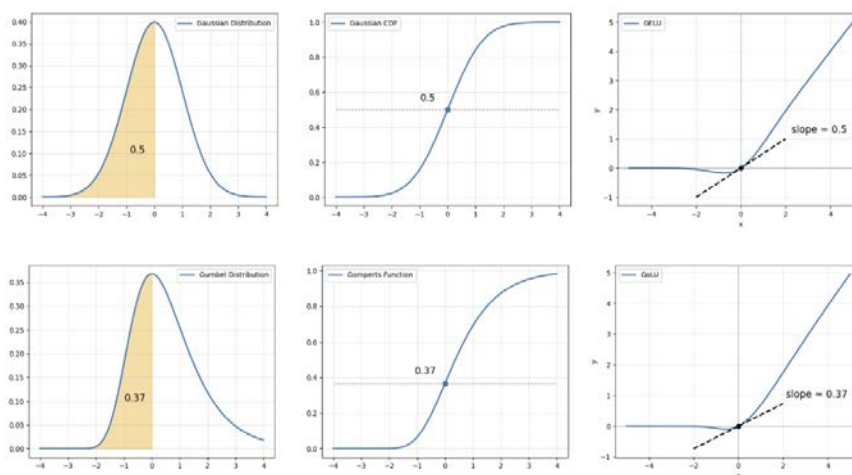


Рисунок 1. Побудова активаційних функцій GELU та GoLU: угорі зліва направо зображено густину нормального розподілу, його функцію розподілу та активаційну функцію GELU; унизу зліва направо – густину розподілу Гумбеля, відповідну функцію Гомперца та активаційну функцію GoLU

Основна функція втрат побудована як фокальна – вона зменшує внесок «легких» негативних прикладів (численних non-entity спанів) і фокусує навчання на складних випадках. Додатково введено контрастивну компоненту відносно прототипів класів: представлення спанів «притягуються» до векторних центрів свого типу та «відштовхуються» від інших прототипів, що структурує латентний простір і покращує роздільність класів. Оптимізація параметрів виконується за допомогою Sophia-G – стохастичного оптимізатора другого порядку, який апроксимує діагональні елементи матриці Гессе й стабілізує оновлення параметрів.

3. СТВОРЕННЯ КОРПУСУ ДАНИХ

Через відсутність відкритих наборів даних для поставленої задачі, було розроблено пайплайн напівавтоматичної анотації українських текстів (рис. 2). Як джерела використано набори `hromadske_corruption` та `HQMS (uk)` з платформи Hugging Face, що сумарно містять близько 36 тисяч прикладів. Процес включав попередню фільтрацію текстів, нормалізацію переліку сутностей, анотацію за допомогою GPT-4o зі strict JSON-схемою виходу та детерміновану поствалідацію меж сутностей з корекцією координат. На фінальному етапі формувався узгоджений JSONL-корпус. Щоб забезпечити відтворюваність, у викликах моделі було фіксовано: температуру, JSON-схему у strict-режимі, політику повторних спроб.

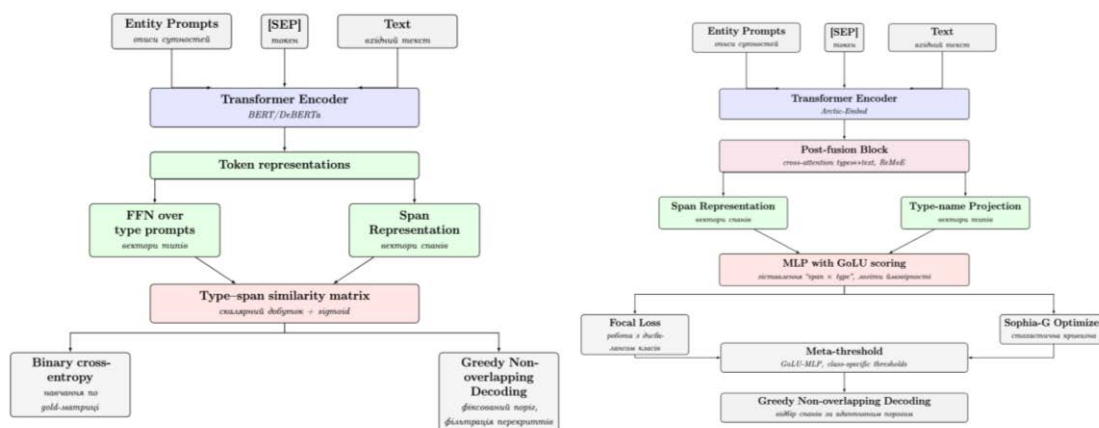


Рисунок 2. Порівняння методу GLiNER (ліворуч) та запропонованого методу (праворуч)

4. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ

Експериментальне дослідження включало порівняння функцій активації, оптимізаторів та конфігурацій скорингового модуля. Показано, що GoLU зменшує частку насичених нейронів з 50–55 % до близько 19 %, що позитивно впливає на стабільність навчання (табл. 1, рис. 3, 4).

Таблиця 1. Результати експериментів з оптимізаторами

Оптимізатор	Precision	Recall	F1-score
Sophia-G	0.7211	0.8713	0.7891
AdaFactor	0.6655	0.8157	0.7330
AdamW	0.6429	0.7931	0.7101

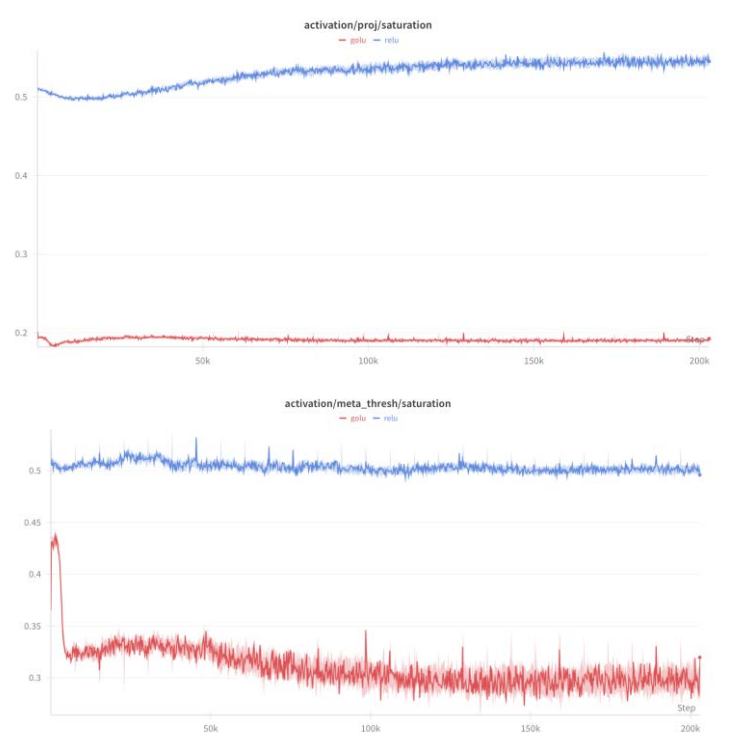
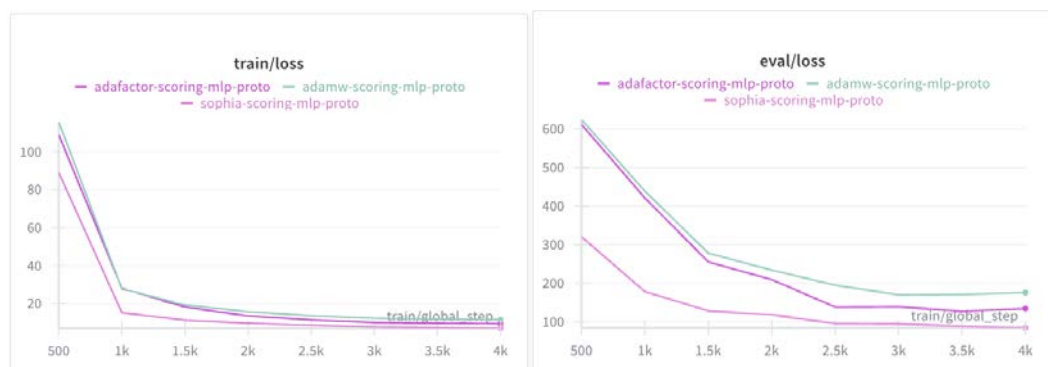


Рисунок 3. Порівняння насичення активацій GoLU та ReLU у різних підмережах моделі



а)

б)



в)

Рисунок 4. Порівняння оптимізаторів AdaFactor, AdamW і Sophia-G: (а) зміна валідаційної функції втрат, (б) зміна тренувальної функції втрат, (в) динаміка норм градієнтів

Sophia-G продемонструвала найкращі показники серед розглянутих оптимізаторів: F1 = 0.7891 проти 0.7330 для AdaFactor та 0.7101 для AdamW, завдяки врахуванню кривизни поверхні втрат та більш адекватним оновленням параметрів (на рис. 4 видно норми градієнтів).

Таблиця 2. Результати експериментів з модулем скорінгу

Експеримент	Precision	Recall	F1-score
MLP with proto	0.7211	0.8713	0.7891
Dot product with proto	0.7039	0.8541	0.7718
MLP without proto	0.6843	0.8345	0.7520

Порівняння модулів скорінгу показало перевагу MLP-скорінгу з контрастивною втратою над скалярним добутком та MLP без прототипів: відповідні F1-score становили 0.7891, 0.7718 та 0.7520.

5. ВИСНОВКИ

Розроблена архітектура на основі GLiNER із post-fusion блоком демонструє ефективність у zero-shot та few-shot NER для української мови. Запропоновані модифікації дозволяють досягти F1=0.7891 при збереженні компактності моделі (900М параметрів) та помірних обчислювальних витрат.

Методологія напівавтоматичної анотації корпусу із залученням GPT-4o забезпечує швидку розмітку даних. Використання Arctic-Embed 2.0-L з MRL робить можливим гнучке налаштування розмірності ембедінгів і адаптацію моделі до різних ресурсних обмежень.

Отримані результати створюють підґрунтя для впровадження ефективних NER-систем у доменах, де відсутні великі розмічені корпуси даних, що має практичну цінність для українського NLP.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Chaplinskyi D., Romanyshyn M. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian // Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024. 2024. P. 23–29.
2. Zaratiana U., Tomeh N., Holat P., Charnois T. GLiNER: Generalist Model for Named Entity Recognition Using Bidirectional Transformer // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024. P. 5364–5376.
3. Yu Y., Zhang Y., Ren Z. et al. Arctic-Embed 2.0: Multilingual Retrieval Without Compromise // arXiv preprint arXiv:2412.04506. 2024.
4. Wang Z., Zhu J., Chen J. ReMoE: Fully Differentiable Mixture-of-Experts with ReLU Routing // Proceedings of the International Conference on Learning Representations (ICLR). 2025.
5. Das I., Safari M., Adriaensen S., Hutter F. Gompertz Linear Units: Leveraging Asymmetry for Enhanced Learning Dynamics // arXiv preprint arXiv:2502.03654. 2025.
6. Liu H., Li Z., Hall D. et al. Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training // arXiv preprint arXiv:2305.14342. 2023.
7. Vitaliias. hromadske_corruption. Hugging Face Datasets. URL: https://huggingface.co/datasets/Vitaliias/hromadske_corruption (дата звернення: 20.11.2025).
8. agentlans. high-quality-multilingual-sentences (subset: uk). Hugging Face Datasets. URL: <https://huggingface.co/datasets/agentlans/high-quality-multilingual-sentences> (дата звернення: 20.11.2025).