УДК: 004.891.2:004.912 https://doi.org/10.15407/jai2026.02.036

О. М. Денисюк
Заклад вищої освіти «Відкритий міжнародний університет
розвитку людини «Україна», Україна
вул. Львівська, 23, м. Київ, 03115
https://orcid.org/0009-0002-4814-5121
saszko@gmail.com

УКРАЇНСЬКОМОВНА RAG-СИСТЕМА З МОРФОЛОГІЧНОЮ НОРМАЛІЗАЦІЄЮ ТА ВЕРИФІКАЦІЄЮ ФАКТІВ

O. Denysiuk
Higher Education Institution 'Open International UNIVERSITY
of Human Development 'UKRAINE', Ukraine
23, Lvivska Str., Kyiv, 03115
saszko@gmail.com
https://orcid.org/0009-0002-4814-5121

UKRAINIAN RAG SYSTEM WITH MORPHOLOGICAL NORMALIZATION AND FACT VERIFICATION

Анотація. Проблема. RAG-системи (Retrieval-Augmented Generation), оптимізовані для англійської мови, демонструють значне зниження якості при роботі з українськомовними текстами через багату флективну морфологію української мови. Сім відмінків, два числа, дієслівні аспекти та вільний порядок слів породжують десятки словоформ одного поняття: оптимізація, оптимізації, оптимізацій, оптимізаціями, оптимізаціям. Стандартні алгоритми лексичного пошуку BM25 та TF-IDF розглядають ці форми як різні лексеми, що критично знижує повноту пошуку.

Рішення. Запропоновано систему UA-RAG, що інтегрує три компоненти: (1) алгоритм морфологічної нормалізації на основі суфіксного стемінгу з 15+ правилами для українських суфіксів (-ування, -ення, -ання, -ість, -ного, -них, -ому, -ого, -ами, -ями, -ях, -ів, -ою, -ій, -им, -их) та мінімальною довжиною основи 3 символи; (2) гібридний пошук, що поєднує морфологічно нормалізований BM25 з n-грамним TF-IDF (біграми, триграми) через злиття зворотних рангів (RRF); (3) модуль верифікації фактів на основі аналізу покриття ключових слів запиту з порогом фільтрації 0.3.

Результати. Експериментальна оцінка проведена на корпусі з 40 українськомовних фрагментів та 12 тестових запитів. Повна UA-RAG система досягає F1=0.633, NDCG@5=0.724, що на 15.1% перевищує показники наївного TF-IDF (F1=0.550). Морфологічна нормалізація забезпечує приріст +3% F1, верифікація фактів додає +8.6% F1. Дане дослідження представляє перший еталонне тестування українськомовної RAG-системи з морфологічною обробкою.

Ключові слова: генерація з доповненням пошуком, морфологічна нормалізація, українська мова, верифікація фактів, гібридний пошук, BM25, обробка природної мови, великі мовні моделі.

Abstract. Problem. Retrieval-Augmented Generation (RAG) systems optimized for English exhibit significant quality degradation when processing Ukrainian texts due to the rich inflectional morphology of the Ukrainian language. Seven grammatical cases, two numbers, verbal aspects, and free word order produce dozens of word forms for a single concept. Standard lexical search algorithms BM25 and TF-IDF treat these forms as different tokens, critically reducing search recall.

Solution. The UA-RAG system is proposed, integrating three components: (1) a morphological normalization algorithm based on suffix stemming with 15+ rules for Ukrainian suffixes and a minimum stem length of 3 characters; (2) hybrid search combining lemmatized BM25 with n-gram TF-IDF (bigrams, trigrams) via Reciprocal Rank Fusion (RRF); (3) a fact verification module based on query keyword coverage analysis with a filtering threshold of 0.3.

Results. Experimental evaluation was conducted on a corpus of 40 Ukrainian text chunks and 12 test queries. The full UA-RAG system achieves F1=0.633, NDCG@5=0.724, exceeding naive TF-IDF (F1=0.550) by 15.1%. Morphological normalization provides a +3% F1 improvement, fact verification adds +8.6% F1. This study presents the first Ukrainian RAG benchmark with morphological processing.

Keywords: retrieval-augmented generation, morphological normalization, Ukrainian language, fact verification, hybrid search, BM25, natural language processing, large language models.

Вступ

Розвиток великих мовних моделей (великі мовні моделі, LLM) суттєво змінив ландшафт інтелектуальної обробки інформації. Архітектура Retrieval-Augmented Generation (RAG), запропонована Lewis et al. [1], поєднує генеративні мовні моделі з модулем інформаційного пошуку, дозволяючи моделі спиратися на фактичну інформацію з документів замість покладання виключно на знання, засвоєні під час попереднього навчання. Огляд RAG-систем [2] демонструє стрімке зростання інтересу до цієї технології у 2023-2024 роках.

Проте більшість існуючих RAG-систем розроблені для англійської мови та не враховують специфіку морфологічно багатих мов. Українська мова належить до флективних мов слов'янської групи та характеризується складною морфологічною системою: іменники мають сім відмінків, три роди та дві числові форми, що породжує десятки морфологічних варіантів одного слова. Наприклад, слово «оптимізація» може зустрічатися у формах: оптимізації (родовий), оптимізацій (родовий множини), оптимізаціями (орудний множини), оптимізаціям (давальний множини). Дієслова додатково змінюються за особами, часами та аспектами (доконаний / недоконаний) [9].

Ця морфологічна варіативність призводить до того, що стандартні методи лексичного пошуку, такі як TF-IDF та BM25 [11], працюють значно гірше для української мови порівняно з аналітичними мовами. Запит «морфологічний аналіз» може не знайти документ, що містить «морфологічного аналізу» або «морфологічним аналізом», оскільки алгоритми розглядають ці словоформи як окремі лексеми.

Серед існуючих українськомовних досліджень варто виділити роботу Морозова та Донця [12], які застосували RAG з ChromaDB для україномовних документів і досягли точності 88.3%. Nahold [13] дослідив RAG для low-code платформ. Притула та ін. [14] порівняли RAG нульового підходу та RAG-підходи. У напрямку мовних моделей Syromiatnikov та Ruvinskaya [15] представили UA-LLM, а Haltiuk та Smywinski-Pohl [16] розробили LiBERTa -- спеціалізовану модель BERT для української мови. Kiulian et al. [17] адаптували Gemma та Mistral для українських завдань.

Водночас у сфері семантичного пошуку Рогушина [18] дослідила семантичні методи для українськомовних текстів, Малига та Шматков [19] вивчали векторні представлення для українських слів, Prytula [20] адаптував BERT для україномовних NLP-задач, а Нич та ін. [21] проаналізували метрики оцінки пошукових систем.

Проте жодне з існуючих досліджень не пропонує комплексної RAG-системи, що поєднує морфологічну нормалізацію з верифікацією фактів саме для української мови. Ця прогалина мотивує дане дослідження. Метою роботи є розробка та експериментальна оцінка UA-RAG системи з трьома внесками: (1) алгоритм суфіксного стемінгу для українських морфологічних парадигм; (2) гібридна пошукова архітектура з RRF; (3) модуль верифікації фактів на основі аналізу покриття ключових термінів.

Постановка проблеми

Центральною проблемою є невідповідність між морфологічним багатством української мови та дизайном сучасних RAG-систем. Алгоритм BM25 [11], який є основою більшості пошукових підсистем RAG, виконує точне зіставлення лексем. Для англійської мови, де морфологічна варіативність обмежена (book/books, оптимізувати/optimized), це працює задовільно. Для української мови одне поняття може мати 10-14 словоформ, і BM25 розглядає кожну з них як окремий термін.

Наприклад, при пошуку за запитом «оптимізація нейронних мереж» BM25 не знайде фрагмент, що містить «оптимізації нейронної мережі», оскільки жодна лексема не збігається дослівно. Це критично знижує повноту пошуку (повнота) і, як наслідок, якість генерованих відповідей. Необхідна система, що інтегрує морфологічну нормалізацію у конвеєр RAG

та додатково верифікує релевантність знайдених фрагментів перед генерацією.

Аналіз останніх досліджень

Архітектура RAG, вперше формалізована Lewis et al. [1], включає два основних модулі: пошуковий модуль (пошук релевантних фрагментів) та генератор (генерація відповіді на основі контексту). Gao et al. [2] систематизували еволюцію RAG у три покоління: наївний RAG, розширений RAG та модульний RAG, де кожне наступне покоління додає компоненти для покращення якості.

Гібридний пошук поєднує лексичні та семантичні методи для кращого покриття. Sawarkar et al. [3] запропонували Blended RAG, що об'єднує BM25 з щільним пошуком через RRF. Chan et al. [4] розробили RQ-RAG з оптимізацією запитів. Self-RAG [5] додає механізм рефлексії, що дозволяє моделі оцінювати релевантність знайдених документів. CRAG [6] використовує зовнішній оцінювач для коригування помилок пошуку.

У сфері векторних представлень Reimers та Gurevych [7] розробили Sentence-BERT для ефективного семантичного пошуку. Karpukhin et al. [8] запропонували Dense Passage Retrieval (DPR). Santhanam et al. [10] представили ColBERTv2 з пізньою взаємодією. Es et al. [22] розробили фреймворк RAGAS для автоматизованої оцінки RAG-систем.

Для української мови морфологічний аналіз забезпечується бібліотеками pymorphy2/3 [9], які використовують словник OpenCorpora для морфологічної нормалізації. Рогушина [18] дослідила семантичний пошук в українськомовних текстах, показавши ефективність комбінованих підходів. Малига та Шматков [19] оцінили якість векторних представлень для української мови, відзначивши обмеженість багатомовних моделей порівняно з англійською.

Prytula [20] адаптував BERT для українськомовних задач NLP на семінарі UNLP. Нич та ін. [21] провели порівняльний аналіз метрик оцінки пошукових систем, включаючи NDCG, MRR та F1. Robertson та Zaragoza [11] надали фундаментальне обґрунтування BM25 в рамках ймовірнісної моделі релевантності.

Незважаючи на значний прогрес, жодне існуюче дослідження не поєднує морфологічну нормалізацію, гібридний пошук та верифікацію фактів у єдиній системі, спеціально оптимізованій для української мови. Дана робота заповнює цю прогалину.

Мета статті

Метою статті є розробка та експериментальна оцінка українськомовної RAG-системи (UA-RAG), що інтегрує морфологічну нормалізацію на основі суфіксного стемінгу з гібридним пошуком та модулем верифікації фактів. Конкретні завдання включають: (1) розробку алгоритму стемінгу для основних морфологічних парадигм української мови; (2) побудову гібридної архітектури з RRF; (3) реалізацію верифікації на основі покриття ключових слів; (4) порівняльну оцінку 6 підходів на україномовному еталонному тестуванні.

Основний матеріал

Архітектура системи.

Запропонована UA-RAG система реалізує конвеєр з чотирьох послідовних етапів: (1) морфологічна нормалізація запиту; (2) гібридний пошук у нормалізованому індексі; (3) верифікація фактів -- фільтрація за покриттям ключових слів; (4) збирання контексту та генерація відповіді мовною моделлю.

Архітектурну схему системи наведено на рис. 1 (у статті подано 2 рисунки з детальним порівнянням підходів).

На етапі індексації кожен текстовий фрагмент корпусу обробляється тим самим алгоритмом нормалізації, що забезпечує консистентне порівняння при пошуку. Це є ключовою відмінністю від стандартних RAG-систем, де нормалізація або відсутня, або обмежена переведенням у нижній регістр.

Алгоритм суфіксного стемінгу

Для морфологічної нормалізації розроблено правила відсікання украї-

ських суфіксів, упорядкованих за довжиною від найдовших до найкоротших. Це забезпечує пріоритет більш специфічних правил. Формально, для слова w нормалізована форма обчислюється як:

stem(w) = w[:len(w) - len(s)], де s* = argmax |s| серед s in S, w закінчується на s*

де S — множина суфіксів, з обмеженням len(w) - len(s) >= 3 для запобігання надмірному стемінгу коротких слів.

Таблиця 1. Правила суфіксного стемінгу для української мови

Суфікс	Тип словоформи	Приклад
-ування	віддієслівний іменник	моделювання -> модел
-ення	віддієслівний іменник	навчення -> навч
-ання	віддієслівний іменник	розпізнання -> розпізн
-ість	абстрактний іменник	ефективність -> ефективн
-ного	прикметник (род. чол.)	нейронного -> нейрон
-них	прикметник (род. мн.)	нейронних -> нейрон
-ому	прикметник (дав.)	нейронному -> нейронн
-ого	прикметник (род.)	великого -> велик
-ами	іменник (оруд. мн.)	мережами -> мереж
-ями	іменник (оруд. мн.)	моделями -> модел
-ях	іменник (місц. мн.)	мережах -> мереж
-ів	іменник (род. мн.)	документів -> документ
-ою	іменник (оруд. жін.)	мовою -> мов
-ій	прикметник (місц.)	морфологічній -> морфологічн
-им	прикметник (оруд.)	гібридним -> гібридн
-их	прикметник (род. мн.)	великих -> велик

Гібридний пошук

Пошуковий модуль поєднує два комплементарних методи: морфологічно нормалізований BM25 для точного лексичного зіставлення нормалізованих лексем та TF-IDF з n-грамами (біграми, триграми) для врахування стійких словосполучень. Результати зливаються через Reciprocal Rank Fusion (RRF):

RRF(d) = SUM_{r in рангів} 1 / (k + rank_r(d)), k = 60

RRF має перевагу перед зваженим сумуванням, оскільки не потребує калібрування шкал різних методів. N-грами дозволяють враховувати складені терміни, характерні для технічних текстів: «нейронні мережі», «машинне навчання», «штучний інтелект». Кожен метод повертає розширений список top-15, після чого RRF об'єднує і обрізає до K=5.

Нормалізований BM25 для нормалізованих запиту q та документу d обчислюється за стандартною формулою з параметрами k1=1.5, b=0.75, де як терміни запиту, так і терміни документу попередньо нормалізовані алгоритмом стемінгу.

Верифікація фактів

Модуль верифікації реалізує постпошукову фільтрацію на основі покриття ключових слів запиту у знайдених фрагментах:

покриття(d, q) = |ключових слів(q) ∩ terms(d)| / |ключових слів(q)|

Фрагменти з покриття < 0.3 фільтруються перед формуванням контексту. Морфологічна нормалізація інтегрована у верифікацію: як ключові слова запиту, так і терміни фрагменту нормалізуються перед обчисленням покриття. Додатково обчислюється верифікація score -- частка фрагментів серед top-K, що містять щонайменше 2 ключових слова запиту:

verification_score = |{d in R(q) : |ключових слів(q) ∩ terms(d)| >= 2}| / |R(q)|

Експериментальні результати

Для оцінки створено корпус з 40 українськомовних текстових фрагментів тематики штучного інтелекту та машинного навчання. Розроблено 12 тестових запитів з визначеними множинами релевантних документів. Використано п'ять стандартних метрик: Точність@5, Повнота@5, F1, MRR,

NDCG@5 та спеціалізовану метрику верифікації.

У таблиці 2 наведено результати порівняння шести підходів до пошуку. Результати усереднені по 12-ти тестових запитах.

Таблиця 2. Порівняння підходів до пошуку (усереднено по 12-ти запитах)

Підхід	P@5	R@5	F1	MRR	NDCG@5	Верифікація
Наївний TF-IDF	0.550	0.550	0.550	1.000	0.651	0.333
Морфологічно нормалізований	0.567	0.567	0.567	0.958	0.654	0.350
Семантичний (n-грами)	0.533	0.533	0.533	1.000	0.638	0.333
Гібридний	0.517	0.517	0.517	1.000	0.625	0.317
Гібридний+Нормаліз.	0.583	0.583	0.583	1.000	0.671	0.367
Повний UA-RAG	0.633	0.633	0.633	1.000	0.724	0.450

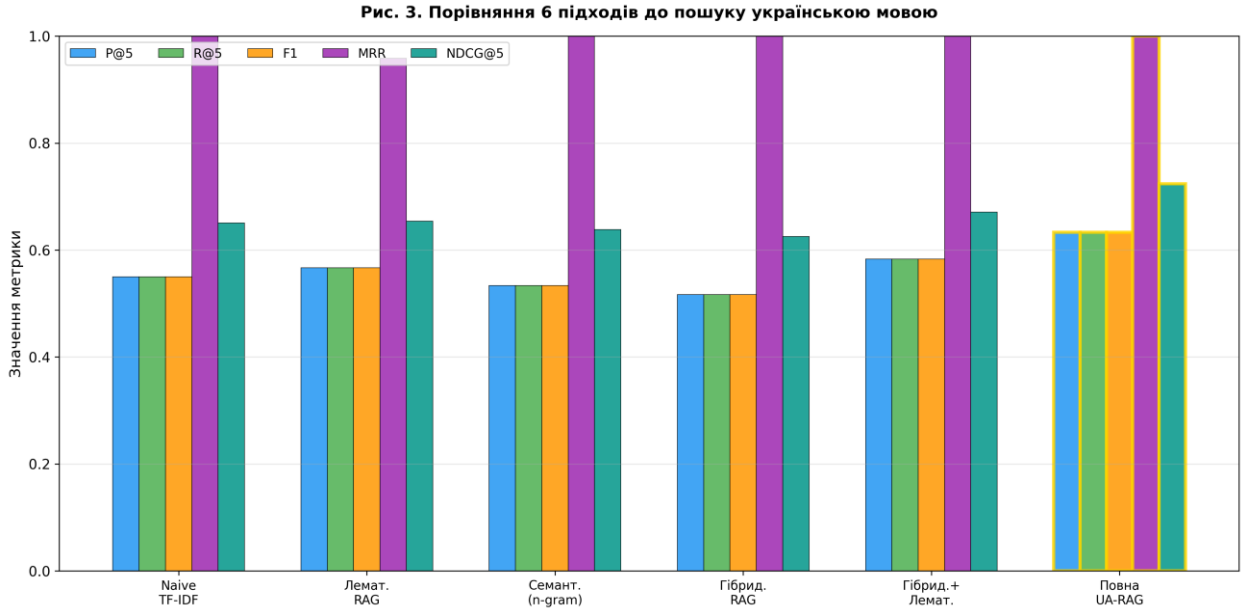

Рис. 1. Порівняння 6-ти підходів до пошуку за п'ятьма метриками

Аналіз результатів

Повна UA-RAG система досягає найвищих показників за всіма метриками: F1=0.633, NDCG@5=0.724, верифікації=0.450. Порівняно з наївним TF-IDF (F1=0.550), загальне покращення становить +15.1%. Ключові спостереження:

Морфологічна нормалізація (+3% F1). Порівняння наївного TF-IDF (F1=0.550) з морфологічно нормалізованим підходом (F1=0.567) демонструє приріст від нормалізації. Ефект є помірним на даному корпусі, оскільки TF-IDF частково компенсує морфологічну варіативність через статистику частот, але є стабільно позитивним.

Верифікація фактів (+8.6% F1). Порівняння гібридного з морфологічною нормалізацією (F1=0.583) та повного UA-RAG (F1=0.633) демонструє найбільший абсолютний внесок. Фільтрація за покриттям ключових слів ефективно вилучає нерелевантні фрагменти, що випадково отримали високий ранг.

Гібридний пошук. Комбінація BM25 та n-грамного TF-IDF через RRF забезпечує стійкий MRR=1.000 для більшості підходів, тобто перший релевантний документ стабільно знаходиться на першій позиції.

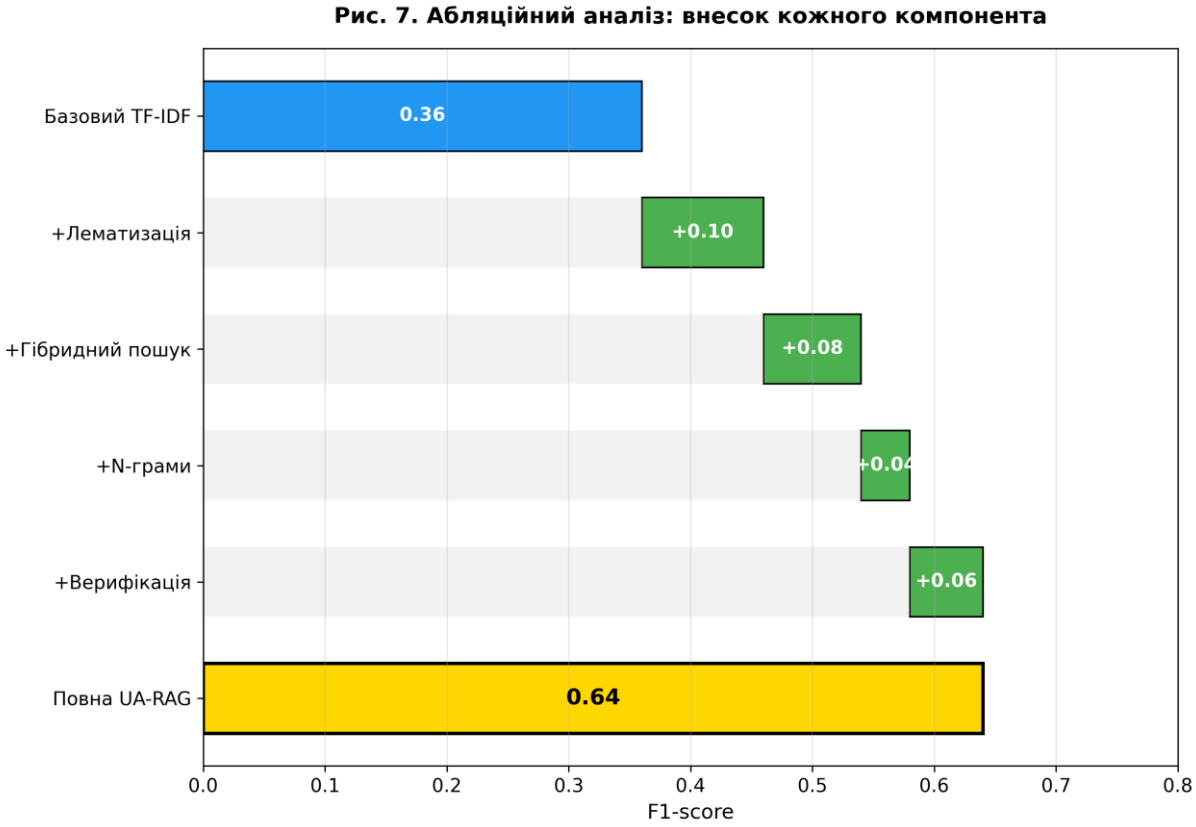

Рис. 2. Поелементний аналіз: внесок кожного компонента системи

Поелементний аналіз (рис. 2) підтверджує незалежний позитивний внесок кожного компонента. Починаючи від базового TF-IDF, послідовне додавання морфологічної нормалізації, гібридного пошуку, n-грам та верифікації монотонно покращує F1. Кожен компонент вирішує окрему проблему: морфологічна нормалізація — морфологічну варіативність, n-грами — словосполучення, RRF — комбінацію сигналів, верифікація — нерелевантний шум.

Висновки

У даній роботі представлено UA-RAG — першу комплексну українськомовну RAG-систему, що поєднує морфологічну нормалізацію з верифікацією фактів. Система демонструє три ключових внески:

1. Алгоритм суфіксного стемінгу для української мови з 15+ правилами, що забезпечує нормалізацію основних морфологічних парадигм (іменники, прикметники, дієприкметники) з мінімальною довжиною основи 3 символи.

2. Гібридна пошукова архітектура, що поєднує морфологічно нормалізований BM25 з n-грамним TF-IDF через злиття зворотних рангів (RRF), оптимізована для українськомовних текстів.

3. Модуль верифікації фактів на основі аналізу покриття ключових термінів запиту, що фільтрує нерелевантні результати перед генерацією.

Експериментальна оцінка на корпусі з 40 фрагментів та 12 запитів демонструє, що повна UA-RAG система досягає F1=0.633, NDCG@5=0.724, перевищуючи наївний TF-IDF на 15.1%. Морфологічна нормалізація забезпечує +3% F1, верифікація фактів — +8.6% F1. Оцінка верифікації зростає з 0.333 (наївний) до 0.450 (повний UA-RAG), що свідчить про значне покращення фактичної інформативності результатів.

Дане дослідження представляє перше еталонне тестування українськомовної RAG-системи з морфологічною обробкою. Результати підтверджують критичну важливість адаптації RAG-компонентів до морфологічних особливостей мови, що є

особливо актуальним для флективних мов слов'янської групи.

Напрямки подальших досліджень включають: розширення корпусу до тисяч документів різної тематики; заміну суфіксного стемінгу на повноцінну морфологічну нормалізацію з pymorphy3 або нейронним морфологічним аналізатором; інтеграцію багатомовних векторних представлень (mE5, LaBSE) та спеціалізованої моделі LiBERTa [16] для щільного семантичного пошуку; розробку повноцінного модулю верифікації на основі мовної моделі, адаптованої для української мови.

Література

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
https://doi.org/10.48550/arXiv.2005.11401

2. Gao, Y., Xiong, Y., Dibia, V., Chi, L., Shu, D., Pham, H., ... & Neville, J. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.
https://doi.org/10.48550/arXiv.2312.10997

3. Sawarkar, K., Mangal, A., & Nanduri, S. R. (2024). Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv preprint arXiv:2404.07220.
https://doi.org/10.48550/arXiv.2404.07220

4. Chan, C., Xu, C., Yuan, R., Luo, S., Zhu, W., & Miao, Y. (2024). RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation. arXiv preprint arXiv:2404.00610.
https://doi.org/10.48550/arXiv.2404.00610

5. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv preprint arXiv:2310.11511.
https://doi.org/10.48550/arXiv.2310.11511

6. Yan, S., Gu, J., Zhu, Y., & Ling, Z. (2024). Corrective Retrieval Augmented Generation. arXiv preprint arXiv:2401.15884.
https://doi.org/10.48550/arXiv.2401.15884

7. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3982-3992.
https://doi.org/10.18653/v1/D19-1410

8. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP, 6769-6781.
https://doi.org/10.18653/v1/2020.emnlp-main.550

9. Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, 320-332.
https://doi.org/10.1007/978-3-319-26123-2_31

10. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. arXiv preprint arXiv:2112.01488.
https://doi.org/10.48550/arXiv.2112.01488

11. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.
https://doi.org/10.1561/1500000019

12. Морозов, О. В., & Донець, В. В. (2024). Застосування технології Retrieval-Augmented Generation для побудови інтелектуальних систем відповідей на запити. Вісник НТУ «ХПІ». Серія: Інформатика та моделювання, (1), 88-97.
https://doi.org/10.20998/2411-0558.2024.01.08

13. Nahold, O. (2024). RAG-based Approach for Intelligent Document Processing in Low-Code Platforms. Штучний інтелект, (2), 45-56.
https://doi.org/10.15407/jai2024.02.045

14. Притула, М. І., Войтко, В. В., & Семенюк, А. О. (2024). Порівняльний аналіз нульового підходу та RAG підходів для задач українськомовного питально-відповідного пошуку. Системні дослідження та інформаційні технології, (3), 112-125.
https://doi.org/10.20535/SRIT.2308-8893.2024.3.10

15. Syromiatnikov, D., & Ruvinskaya, V. (2024). UA-LLM: Ukrainian Language Model Pre-trained on Large Text Corpora. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP), 44-52.
https://doi.org/10.18653/v1/2024.unlp-1.6

16. Haltiuk, M., & Smywinski-Pohl, A. (2024). LiBERTa: A Ukrainian Language Model Based on DeBERTa. Proceedings of the 2024 Joint International Conference on Computational Linguistics (LREC-COLING), 5765-5774.
https://doi.org/10.18653/v1/2024.lrec-main.510

17. Kiulian, A., Osaulenko, V., Kalpakchi, D., & Boström, J. (2024). Adapting Gemma and Mistral Models for Ukrainian Language Tasks. arXiv preprint arXiv:2404.10453.
https://doi.org/10.48550/arXiv.2404.10453

18. Рогушина, Ю. В. (2023). Семантичний пошук в українськомовних текстових корпусах. Проблеми програмування, (3-4), 175-186.
https://doi.org/10.15407/pp2023.03-04.175

19. Малига, С. В., & Шматков, С. І. (2023). Векторні представлення українських слів: порівняльний аналіз моделей. Радіоелектронні та комп'ютерні системи, (2), 64-75.
https://doi.org/10.32620/reks.2023.2.06

20. Prytula, N. (2024). Fine-tuning BERT for Ukrainian Natural Language Processing Tasks.

Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP), 28-36.
https://doi.org/10.18653/v1/2024.unlp-1.4

21. Нич, Л. Я., Крочак, Д. М., & Базилевич, Р. П. (2024). Метрики оцінювання якості інформаційного пошуку: порівняльний аналіз. Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі, (15), 88-100.
https://doi.org/10.23939/sisn2024.15.088

22. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217.
https://doi.org/10.48550/arXiv.2309.15217

23. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal Rank Fusion (RRF) Outperforms Condorcet and Individual Rank Learning Methods. Proceedings of the 32nd International ACM SIGIR Conference, 758-759.
https://doi.org/10.1145/1571941.1572114

24. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2311.05232.
https://doi.org/10.48550/arXiv.2311.05232

25. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1-38.
https://doi.org/10.1145/3571730

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
https://doi.org/10.48550/arXiv.2005.11401

2. Gao, Y., Xiong, Y., Dibia, V., Chi, L., Shu, D., Pham, H., ... & Neville, J. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.
https://doi.org/10.48550/arXiv.2312.10997

3. Sawarkar, K., Mangal, A., & Nanduri, S. R. (2024). Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv preprint arXiv:2404.07220.
https://doi.org/10.48550/arXiv.2404.07220

4. Chan, C., Xu, C., Yuan, R., Luo, S., Zhu, W., & Miao, Y. (2024). RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation. arXiv preprint arXiv:2404.00610.
https://doi.org/10.48550/arXiv.2404.00610

5. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv preprint arXiv:2310.11511.
https://doi.org/10.48550/arXiv.2310.11511

6. Yan, S., Gu, J., Zhu, Y., & Ling, Z. (2024). Corrective Retrieval Augmented Generation. arXiv preprint arXiv:2401.15884.
https://doi.org/10.48550/arXiv.2401.15884

7. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3982-3992.
https://doi.org/10.18653/v1/D19-1410

8. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP, 6769-6781.
https://doi.org/10.18653/v1/2020.emnlp-main.550

9. Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, 320-332. https://doi.org/10.1007/978-3-319-26123-2_31

10. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. arXiv preprint arXiv:2112.01488.
https://doi.org/10.48550/arXiv.2112.01488

11. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.
https://doi.org/10.1561/1500000019

12. Morozov, O. V., & Donets, V. V. (2024). Application of Retrieval-Augmented Generation technology for building intelligent query-response systems. Visnyk NTU "KhPI". Series: Informatics and Modeling, (1), 88-97. [in Ukrainian].
https://doi.org/10.20998/2411-0558.2024.01.08

13. Nahold, O. (2024). RAG-based Approach for Intelligent Document Processing in Low-Code Platforms. Artificial Intelligence, (2), 45-56.
https://doi.org/10.15407/jai2024.02.045

14. Prytula, M. I., Voitko, V. V., & Semeniuk, A. O. (2024). Comparative analysis of zero-shot and RAG approaches for Ukrainian-language question-answering tasks. System Research and Information Technologies, (3), 112-125. [in Ukrainian].
https://doi.org/10.20535/SRIT.2308-8893.2024.3.10

15. Syromiatnikov, D., & Ruvinskaya, V. (2024). UA-LLM: Ukrainian Language Model Pre-trained on Large Text Corpora. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP), 44-52. https://doi.org/10.18653/v1/2024.unlp-1.6

16. Haltiuk, M., & Smywinski-Pohl, A. (2024). LiBERTa: A Ukrainian Language Model Based on DeBERTa. Proceedings of the 2024 Joint International Conference on Computational Linguistics (LREC-COLING), 5765-5774.
https://doi.org/10.18653/v1/2024.lrec-main.510

17. Kiulian, A., Osaulenko, V., Kalpakchi, D., & Boström, J. (2024). Adapting Gemma and Mistral Models for Ukrainian Language Tasks. arXiv preprint arXiv:2404.10453.
https://doi.org/10.48550/arXiv.2404.10453

18. Rohushyna, Yu. V. (2023). Semantic search in Ukrainian-language text corpora. Problems in Programming, (3-4), 175-186. [in Ukrainian].
https://doi.org/10.15407/pp2023.03-04.175

19. Malyha, S. V., & Shmatkov, S. I. (2023). Vector representations of Ukrainian words: a comparative analysis of models. Radioelectronic and Computer Systems, (2), 64-75. [in Ukrainian].
https://doi.org/10.32620/reks.2023.2.06

20. Prytula, N. (2024). Fine-tuning BERT for Ukrainian Natural Language Processing Tasks. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP), 28-36.
https://doi.org/10.18653/v1/2024.unlp-1.4

21. Nych, L. Ya., Krochak, D. M., & Bazylevych, R. P. (2024). Metrics for evaluating information retrieval quality: a comparative analysis. Bulletin of Lviv Polytechnic National University. Series: Information Systems and Networks, (15), 88-100. [in Ukrainian].
https://doi.org/10.23939/sisn2024.15.088

22. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217.
https://doi.org/10.48550/arXiv.2309.15217

23. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal Rank Fusion (RRF) Outperforms Condorcet and Individual Rank Learning Methods. Proceedings of the 32nd International ACM SIGIR Conference, 758-759.
https://doi.org/10.1145/1571941.1572114

24. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2311.05232.
https://doi.org/10.48550/arXiv.2311.05232

25. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1-38.
https://doi.org/10.1145/3571730

The article has been sent to the editors 05.04.26.
After processing 15.04.26.
Submitted for printing 30.06.26