

ЧАБАН Олександр

Хмельницький національний університет

<https://orcid.org/0009-0001-4710-3336>

e-mail: chabanolek@khmnu.edu.ua

МЕТОД ПОЄДНАННЯ КОНТЕКСТНИХ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ СЛІВ ІЗ ВЕКТОРНИМ ПОДАННЯМ МЕДИЧНОГО ДОМЕНУ

Основним викликом у галузі медичних висновків природною мовою (англ. NLI) є досягнення високої точності та надійності автоматизованого аналізу клінічних текстів. Такий аналіз ускладнюється специфікою медичної термінології та потребою врахування контекстуальних знань. У статті подано розв'язок задачі для підвищення якості медичного висновку природною мовою (NLI) через інтеграцію структурованих знань предметної області та інформації про тональність тексту. У роботі запропоновано метод, що поєднує сучасний алгоритм вбудовування знань MultE (застосований до Уніфікованої медичної мовної системи) з моделлю BioELMo, а також інтегрує специфічну для предметної області інформацію про тональність за допомогою MetaMap. Запропонований метод використовує вбудовування BioELMo, інтегровані знання та вектори тональності, оброблені через двоспрямовану LSTM та архітектуру на основі уваги. За набором даних MedNLI метод досяг значення F1-міри в 79,85 %, що перевершило базові моделі. Результати обчислювальних експериментів демонструють, що інтеграція специфічних знань підвищує впровадження NLI у медичній галузі, створюючи в такий спосіб надійну основу для покращення підтримки клінічних рішень.

Ключові слова: медичні висновки природною мовою, вбудовування знань предметної області, інтелектуальні системи охорони здоров'я, штучний інтелект, глибоке навчання, підтримка клінічних рішень.

CHABAN Oleksandr

Khmelnitskyi National University

METHOD OF COMBINING CONTEXTUAL EMBEDDINGS WITH A VECTOR REPRESENTATION OF THE MEDICAL DOMAIN

Navigating the intricate logical connections within clinical narratives—a medical natural language inference task—is paramount for advancing applications like AI-assisted clinical decision-making and the automated interpretation of patient records. However, mastering this domain is particularly arduous due to the specialized lexicon, complex conceptual relationships, and subtle semantic variations inherent in medical texts. This research introduces an innovative methodology to elevate medical natural language inference performance by effectively combining structured, field-specific knowledge with insights gleaned from textual sentiment. Our approach capitalizes on MultE, a cutting-edge algorithm for embedding knowledge graphs, to distill profound semantic relationships from the Unified Medical Language System (UMLS). These distilled knowledge representations are then amalgamated with contextual word embeddings generated by BioELMo. To further enrich contextual understanding, sentiment data pertinent to the medical field, extracted via MetaMap, is also integrated. The system architecture processes this composite feature set—BioELMo embeddings augmented by domain knowledge and sentiment vectors—through a bidirectional Long Short-Term Memory (BiLSTM) network, which is subsequently enhanced by an attention mechanism that dynamically assigns importance to different input segments. Validation on the MedNLI benchmark dataset, featuring 14,049 expert-labeled premise-hypothesis pairs, revealed exceptional efficacy. The proposed system achieved 81.14% accuracy, 79.62% recall, an F1-score of 79.85%, and an AUC-ROC of 85.06%, surpassing established baseline techniques. These accomplishments underscore that the deliberate incorporation of specialized knowledge and sentiment cues can dramatically boost natural language inference capabilities in the medical arena, thereby providing a sturdy platform for engineering more dependable and intelligent healthcare solutions.

Keywords: clinical text understanding, knowledge-enhanced NLI, sentiment analysis in healthcare, deep learning for medicine, intelligent decision support.

Стаття надійшла до редакції / Received 24.04.2025

Прийнята до друку / Accepted 11.05.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Висновки природною мовою (NLI) є важливим завданням у галузі штучного інтелекту (ШІ) та розуміння природної мови (NLU). Воно фокусується на визначенні логічних зв'язків між заданим засновком та гіпотезою (наслідок, суперечність, нейтральність). Хоча значного прогресу досягнуто в загальних доменах, медична галузь залишається відносно недостатньо дослідженою [1, 2]. Складність та спеціалізований характер медичної мови (жаргон, аббревіатури, контекстно-залежні значення) створюють унікальні проблеми для створення точних та надійних моделей NLI.

Поява набору даних MedNLI [3], спеціально розробленого для NLI у медичній галузі, стала важливим кроком. Проте, складність медичних текстів посилює проблеми моделювання висновків та заключень. З огляду на такі наукові розробки, зростає важливість пояснюваного ШІ (XAI) [4]. Прозорі та інтерпретовані моделі ШІ [5] є ключовими для отримання довіри медичних працівників.

Проектування передових методів вбудовування слів, таких як ELMo та BERT, значно змінила ландшафт оброблення природної мови (NLP). У спеціалізованій галузі біомедичного NLP, моделі на кшталт BioBERT та BioELMo були донавшені на великих біомедичних корпусах текстових даних, покращуючи розуміння медичних текстів та встановлюючи нові стандарти в медичному NLI [6]. Попри ці досягнення, інтеграція зовнішніх знань предметної області в моделі NLI стала критичною сферою досліджень [7].

Інтеграція специфічної для домену інформації про настрої також є важливим напрямком [8]. Наприклад, у роботі [1] автори інтегрували вбудовування з графа знань на основі Уніфікованої медичної мовної системи (UMLS) з інформацією про настрої в рамках BioELMo. Однак, неповне дослідження ознак клінічних знань залишається проблемою. Ця робота спрямована на подолання цієї прогалини через розроблення нового методу, що інтегрує вбудовування знань предметної області та аналіз тональності, які вилучені з UMLS. Науковий внесок полягає у новому методі вбудовування знань предметної області з UMLS в передові медичні моделі NLU, такі як BioELMo, та вдосконаленому підході до інтегрування інформації щодо тональності, яка пов'язана з медичними концепціями з UMLS.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Запропонований метод поєднання контекстних векторних представлень слів із векторним поданням медичного домену призначений для підвищення якості медичного NLI через інтеграцію структурованих знань предметної області (з UMLS через MultE), контекстуальних векторних представлень слів (BioELMo) та інформації про тональність тексту (з MetaMap). У роботі розглянуто задачу класифікації як стандартну проблему NLI, а саме: необхідно визначити, чи можна гіпотезу вивести із висновку, та класифікувати її як відповідність (entailment), суперечність (contradiction) або нейтральність (neutral). Метод ґрунтується на підході, що вперше був описаний в [2], використовуючи модель вбудовування BioELMo та розширену послідовну модель висновків (ESIM). Схема методу проілюстровано на рис. 1.

Вхідними даними методу є пари текстів: "передумова-висновок" (так звані Premise-Hypothesis), що отримані з клінічних записів:

- 1) корпус текстів-передумов (premise);
- 2) корпус текстів-висновків (hypothesis).

Нижче подамо основні блоки запропонованого методу з розбиттям на елементарні кроки.

Блок 1. Підготовка даних:

Крок 1.1. Токенізація: Кожне речення (висновок та гіпотеза) токенизується за допомогою Classical Language Toolkit (CLTK). Результатом токенизації є такі послідовності tokenів:



Рис. 1. Загальна схема запропонованого методу класифікації пар «висновок-гіпотеза»

1) послідовність tokenів, що отримані з текстів-передумов: $P = (p_1, p_2, \dots, p_m)$, де p_i – i -й token передумови, m – кількість tokenів для текстів-передумов;

2) послідовність tokenів, що отримані з текстів-висновків: $H = (h_1, h_2, \dots, h_n)$, де h_j – j -й token висновку, n – кількість tokenів для текстів-висновків.

Крок 1.2. Оброблення MetaMap: Речення обробляються MetaMap для вилучення концептів UMLS та пов'язаної інформації про тональність. Кожен концепт узгоджується зі словами, що його складають.

2. Генерація векторних представлень:

Крок 2.1. Вбудовування BioELMo: Контекстуальні вбудовування слів генеруються для кожного токена за допомогою моделі BioELMo, попередньо навченої на десяти мільйонах рефератів PubMed.

Крок 2.2. Вбудовування знань предметної області: Модель MultE використовується для генерації векторних представлень для кожного концепту UMLS, витягнутого MetaMap. Для побудови знань предметної

області весь набір даних MedNLI оброблено MetaMap, витягнуто 7 496 унікальних медичних концептів. Списки ребер знань предметної області включають 117 467 трійок з Metathesaurus та 23 824 105 трійок з Semantic Network. Покращена формалізація векторних представлень MultE:

$$\text{score}_{\text{MultE}} = \text{ReLU}(W_r r + b_r) \text{ReLU}(W_h h + b_h) \cdot \text{ReLU}(W_t t + b_t), \quad (1)$$

де W_r, W_h, W_t – вагові матриці, а b_r, b_h, b_t – вектори зміщень.

Крок 2.3. Інтеграція тональності: Для кожного токена створюється вектор тональності (1-D), що вказує на позитивну (0) або негативну (1) тональність.

Блок 3. Конкатенація векторних представлень: Вбудовування BioELMo, MultE та вектори тональності для кожного токена об'єднуються. Покращений підхід до об'єднання векторних представлень для слова w :

$$e^w = \alpha e_{\text{BioELMo}}^w + \beta e_{\text{MultE}}^w, \quad (2)$$

де α та β – коефіцієнти, що навчаються.

Далі інтегрується інформація про тональність s^w з механізмом уваги:

$$e_{\text{final}}^w = \sum_{i=1}^n a_i (e_{\text{BioELMo},i}^w + e_{\text{MultE},i}^w + s_i^w), \quad (3)$$

де a_i – ваги уваги.

Блок 4. Архітектура моделі (ESIM):

Крок 4.1. Кодування речень: Шари двоспрямованої LSTM (BiLSTM) використовуються для кодування векторних представлень висновків та гіпотез окремо.

Крок 4.2. Механізм уваги: Обчислюється попарна матриця уваги між закодованими висновком та гіпотезою.

Крок 4.3. Другий шар BiLSTM: Застосовується до представлень висновку та гіпотези з урахуванням уваги. Виконується максимальне та середнє агрегування на виходах BiLSTM.

Крок 4.4. Класифікація Softmax: Агреговані виходи подаються на шар Softmax для класифікації.

Вихідними даними запропонованого методу є результат класифікації кожної пари "передумова-висновок" за одним із трьох класів: а) відповідність (істина), б) суперечність (хибність) або нейтральність (невизначено).

Для проведення експериментальних досліджень над запропонованим методом використано набір даних MedNLI з MIMIC-III, що містить 14 049 пар "передумова-висновок" (11 232 для навчання, 1 395 для валідації, 1 422 для тестування). Середня довжина висновків – 20 слів, гіпотез – 5,8 слів.

Для моделі MultE розмірність векторних представлень слів становила 100. Використовувався оптимізатор Stochastic Gradient Descent з початковою швидкістю навчання 10^{-3} та розміром пакету 64. Для ESIM використовувалися мережі BiLSTM з розміром прихованого стану 500. Застосовувався Dropout з коефіцієнтом 0,4. Початкова швидкість навчання для ESIM також 10^{-3} , розмір пакету 32. Навчання моделі обмежувалося максимум 64 епохами з механізмом ранньої зупинки (якщо втрати на валідаційному наборі не зменшувалися протягом 5 послідовних епох).

Для оцінювання результатів класифікації використовуються стандартні метрики: Accurasy (точність класифікації), Precision (влучність класифікатора), Recall (повнота вгадувань), F1-міра та площа під ROC-кривою (AUC-ROC).

На рисунку 2 показано ROC-криві за отриманими результатами.

З рис. 2 бачимо, що BioELMo має AUC 82,17%, ESIM-know – 83,08%. BioELMo + Sentiment досягає AUC 84,03%. Запропонований метод перевершує всі моделі з AUC 85,06%.

У таблиці 1 наведено порівняльний аналіз різних моделей. Аналіз експериментальних результатів, що подані у статті, демонструє значне покращення класифікації пар "передумова-висновок" за запропонованим методом поєднання контекстних векторних представлень слів (BioELMo) із векторним поданням медичного домену (MultE з UMLS) та інформацією про тональність (MetaMap). Запропонований підхід досяг найвищих показників поміж усіх порівнюваних моделей на наборі даних MedNLI: accurasy – 81,14%, recall – 79,62%, F1-міра – 79,85% та AUC-ROC – 85,06%. Це суттєво перевершує як базові моделі, наприклад, fastText (AUC-ROC 75,16%), так і більш просунуті, такі як BioELMo окремо (AUC-ROC 82,17%) чи ESIM-know (що інтегрує знання, AUC-ROC 83,08%).

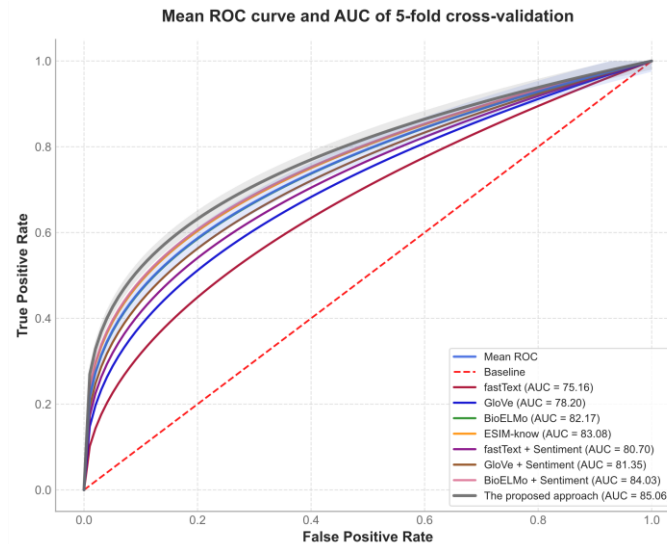


Рис. 2. ROC-криві та значення AUC для досліджуваних моделей глибокого навчання за набором даних MedNLI (5-кратна перехресна валідація)

Таблиця 1

Порівняння продуктивності різних моделей на наборі даних MedNLI (значення у %).

Підхід	Accuracy	Recall	Precision	F1-міра	AUC-ROC
fastText	73,50	71,05	72,00	71,57	75,16
GloVe	76,42	74,60	75,12	74,82	78,20
BioELMo	79,73	78,15	78,55	78,23	82,17
ESIM-know	80,22	78,67	79,05	78,76	83,08
fastText + Sentiment	78,59	77,80	77,58	77,17	80,70
GloVe + Sentiment	79,10	77,43	78,08	77,74	81,35
BioELMo + Sentiment	80,60	79,30	79,48	79,19	84,03
Запропонований метод	81,14	79,62	80,08	79,85	85,06

Якісний аналіз показує, що поданий метод правильно класифікує пари, де семантична еквівалентність або негативна тональність є ключовими. Наприклад, "Diagnosis of COPD" та "Patient has chronic obstructive pulmonary disease" класифікуються як наслідок. Однак, існують помилкові класифікації, особливо з тонкими нюансами тональності або тимчасовим контекстом, наприклад, "He was walking steadily at that moment" та "The patient never had a steady walk" помилково класифіковано як наслідок.

Попри досягнуті високі результати, автори статті вказують на певні обмеження запропонованого методу. Якісний аналіз показав, що хоча модель успішно справляється з випадками, де семантична еквівалентність або явна (особливо негативна) тональність є визначальними, вона може мати труднощі з інтерпретацією більш тонких нюансів. Зокрема, це стосується точного виявлення прихованої негативної тональності, коли негативний відтінок не виражений явно, а також обробки тимчасової інформації в клінічних даних, де часові маркери можуть кардинально змінювати логічний зв'язок між передумовою та гіпотезою. Прикладом є помилкова класифікація пари "He was walking steadily at that moment" та "The patient never had a steady walk" як наслідку, що вказує на недостатнє врахування тимчасового аспекту "at that moment".

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ

І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У підсумку, результати проведеного дослідження продемонстрували, що інтеграція вбудовуваних знань предметної області (MultE з UMLS) та аналізу тональності (MetaMap) з BioELMo значно покращує продуктивність завдань медичного NLI. Запропонований метод досяг точності 81,14%, F1-міри 79,85% та AUC-ROC 85,06%, перевершивши базові моделі. Ці покращення підкреслюють високу результативність поєднання знань предметної області та аналізу тональності для врахування складних нюансів медичних текстів. Однак, варто виділити обмеження такого підходу, зокрема, в точному виявленні прихованої негативної тональності та обробці тимчасової інформації в клінічних даних.

Майбутня наукова робота ґрунтуватиметься на вдосконаленні аналізу тональності в межах запропонованого методу для кращого врахування тонких ознак. Розширення цього методу на інші медичні набори даних та дослідження його застосовності до ширших завдань підтримки клінічних рішень також буде критично важливим для просування цієї галузі.

Література

1. Incorporating domain knowledge into medical NLI using knowledge graphs / S. Sharma et al. *Proc. Conf. Empir. Methods Nat. Lang. Process. Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)* : Proceedings, Hong Kong, China, 3–7 November 2019. Stroudsburg, PA, USA, 2019. P. 6092–6097. URL: <https://doi.org/10.18653/v1/d19-1631>
2. Chaban O., Manziuk E. Enhancing medical NLI with integrated domain knowledge and sentiment analysis. *Inf. Control Syst. Technol. (ICST-2024)* : CEUR-WS Proc., Odesa, Ukraine, 23–25 September 2024 / ed. by A. Pakštas et al. Aachen, 2024. P. 262–272. URL: <https://ceur-ws.org/Vol-3790/paper23.pdf>
3. Romanov A., Shivade C. Lessons from natural language inference in the clinical domain. *Proc. Conf. Empir. Methods Nat. Lang. Process.* : Proceedings, Brussels, Belgium, 2–4 November 2018. Stroudsburg, PA, USA, 2018. P. 1586–1596. URL: <https://doi.org/10.18653/v1/d18-1187>
4. Explainable deep learning: A visual analytics approach with transition matrices / P. Radiuk et al. *Mathematics*. 2024. Vol. 12, no. 7. P. 1024. URL: <https://doi.org/10.3390/math12071024>
5. Representation of trustworthiness components in AI systems: A formalized approach considering interconnections and overlap / E. Manziuk et al. *Proc. Int. Workshop IT-prof. Artif. Intell. (ProfIT AI 2024) 2024* : CEUR-WS Proc., Cambridge, MA, USA, 25–27 September 2024 / ed. by D. Chumachenko, P. Golland, S. Yakovlev. Aachen, 2024. P. 405–417. URL: <https://ceur-ws.org/Vol-3777/paper25.pdf>
6. Towards efficient methods in medical question answering using knowledge graph embeddings / S. Sengupta et al. *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)* : Proceedings, Lisbon, Portugal, 3–6 December 2024. New York, NY, USA, 2024. P. 5089–5096. URL: <https://doi.org/10.1109/bibm62325.2024.10821824>
7. Чабан О. Р., Манзюк Е. А. Метод інтегрування доменних знань у багатостратегічну класифікацію медичних зображень. *Вісник Хмельницького національного університету. Серія: Технічні науки*. 2024. Т. 337, № 3(2). С. 231–236. URL: <https://doi.org/10.31891/2307-5732-2024-337-3-34>
8. Robust R-peak detection using deep learning based on integrating domain knowledge / O. Kovalchuk et al. *Proc. of the 6th Int. Conf. on Inform. & Data-Driven Med.* : CEUR-Workshop Proceedings, Bratislava, Slovakia, 17–19 November 2023 / ed. by N. Shakhovska et al. Aachen, 2024. P. 1–14. URL: <https://ceur-ws.org/Vol-3609/paper1.pdf>

References

1. Incorporating domain knowledge into medical NLI using knowledge graphs / S. Sharma et al. *Proc. Conf. Empir. Methods Nat. Lang. Process. Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)* : Proceedings, Hong Kong, China, 3–7 November 2019. Stroudsburg, PA, USA, 2019. P. 6092–6097. URL: <https://doi.org/10.18653/v1/d19-1631>
2. Chaban O., Manziuk E. Enhancing medical NLI with integrated domain knowledge and sentiment analysis. *Inf. Control Syst. Technol. (ICST-2024)* : CEUR-WS Proc., Odesa, Ukraine, 23–25 September 2024 / ed. by A. Pakštas et al. Aachen, 2024. P. 262–272. URL: <https://ceur-ws.org/Vol-3790/paper23.pdf>
3. Romanov A., Shivade C. Lessons from natural language inference in the clinical domain. *Proc. Conf. Empir. Methods Nat. Lang. Process.* : Proceedings, Brussels, Belgium, 2–4 November 2018. Stroudsburg, PA, USA, 2018. P. 1586–1596. URL: <https://doi.org/10.18653/v1/d18-1187>
4. Explainable deep learning: A visual analytics approach with transition matrices / P. Radiuk et al. *Mathematics*. 2024. Vol. 12, no. 7. P. 1024. URL: <https://doi.org/10.3390/math12071024>
5. Representation of trustworthiness components in AI systems: A formalized approach considering interconnections and overlap / E. Manziuk et al. *Proc. Int. Workshop IT-prof. Artif. Intell. (ProfIT AI 2024) 2024* : CEUR-WS Proc., Cambridge, MA, USA, 25–27 September 2024 / ed. by D. Chumachenko, P. Golland, S. Yakovlev. Aachen, 2024. P. 405–417. URL: <https://ceur-ws.org/Vol-3777/paper25.pdf>
6. Towards efficient methods in medical question answering using knowledge graph embeddings / S. Sengupta et al. *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)* : Proceedings, Lisbon, Portugal, 3–6 December 2024. New York, NY, USA, 2024. P. 5089–5096. URL: <https://doi.org/10.1109/bibm62325.2024.10821824>
7. CHABAN, O., MANZIUK, E., & DUKA, O. (2024). METHOD OF DOMAIN KNOWLEDGE INTEGRATION IN MULTI-STRATEGIC CLASSIFICATION OF MEDICAL IMAGES. *Herald of Khmelnytskyi National University. Technical Sciences*, 337(3(2)), 231–236. <https://doi.org/10.31891/2307-5732-2024-337-3-34>
8. Robust R-peak detection using deep learning based on integrating domain knowledge / O. Kovalchuk et al. *Proc. of the 6th Int. Conf. on Inform. & Data-Driven Med.* : CEUR-Workshop Proceedings, Bratislava, Slovakia, 17–19 November 2023 / ed. by N. Shakhovska et al. Aachen, 2024. P. 1–14. URL: <https://ceur-ws.org/Vol-3609/paper1.pdf>