КОМП'ЮТЕРНІ НАУКИ

УДК 004.421 https://doi.org/10.31713/vt3202437

Ляшко Д. А., аспірант (Національний університет водного господарства та природокористування, м. Рівне, d.a.lyashko@nuwm.edu.ua)

ПРОБЛЕМИ ОБРОБКИ ПРИРОДНОЇ МОВИ У МАЛОРЕСУРСНОМУ СЕРЕДОВИЩІ. АНАЛІЗ ПЕРСПЕКТИВНИХ МЕТОДІВ РІШЕННЯ

Проблема технічних та архітектурних рішень для малоресурсних середовищ, хоча і є важливою для досліджень, проте зазвичай ігнорується більшістю науковців, які фокусуються на рішеннях для середовищ з великим ресурсом. У роботі проводиться аналіз сучасних методів обробки малоресурсних мов, зосереджуючись на проблемах і обмеженнях, пов'язаних із відсутністю якісних мовних ресурсів та інструментів. Визначено ключові виклики, як-от нестача даних, неможливість застосування стандартних методів для малоресурсних мов, проблеми оцінювання моделей та безпекові проблеми. Розглянуто сучасні підходи, які включають використання багатомовних, мономовних та великих мовних моделей, а також методів покращення навчання для цих мов, таких як міжмовні репрезентації, Task specific fine-tuning, розширення словника та інші. Розглядаються нові, перспективні архітектури нейромереж. Наприклад Mamba, у майбутньому має потенціал замінити стандартну модель трансформера. Мережі Колгоморова – Арнольда є принципово новим архітектурним рішенням класичної багатошарової мережі і може показувати непогану ефективність в порівнянні зі звичайними методами. За результатами роботи робиться висновок про неоднозначність кожної технології у ефективності виконання задач у малоресурсному середовищі. Спільною проблемою усіх представлених рішень є потреба у великій кількості даних. Мультимовні та великі мовні моделі дають кращі результати за відсутності адекватних даних, ніж мономовні, через можливість навчання на корпусах схожих мов. У свою чергу мономовні моделі є більш прозорими, передбачуваними та ефективними для вузької задачі.

Висновком даної статті є рекомендація вибору технології виходячи з умов поставленої задачі, кількості даних та опираючись

на емпіричний метод, оскільки жоден з методів не має абсолютної переваги над іншими і може давати неочікувані результати.

Ключові слова: **малоресурсна мова; модель; архітектура; мовна модель; метод.**

1. Вступ

Попри популярність досліджень у області обробки природної мови у малоресурсних середовищах сам термін «малоресурсне середовище» історично мав змінне та нечітке визначення. Наприклад, малоресурсна мова для виникаючих інцидентів (англ. Low Resource Language for Emergent Incidents, LORELEI) – американський проєкт розвитку малоресурсних мов – визначає малоресурсні середовища як ті, для яких не існує автоматизованих технологій обробки природної мови. Дане визначення фокусується лише на аспекті малоресурсності, який спричинений відсутністю якісної мовної анотації, ігноруючи можливу специфічність завдання, через яку технології можуть бути відсутні. Інші автори, наприклад (В. Бермент та ін.) [1], пропонують набір евристик для визначення низькоресурсності середовища, які включають наявність мінімальних наборів даних для різних підзадач обробки природної мови, інструментів та людських ресурсів.

Метою дослідження є подання та аналіз сучасних методів обробки малоресурсних мов, Було поставлено наступні завдання:

- аналіз сучасних підходів та рішень;
- визначення сильних та слабких сторін кожного розглянутого методу;
- виявити перспективні та нові напрями досліджень, та проаналізувати їх науковий потенціал.

2. Малоресурсні середовища. Виклики та проблеми обробки малоресурсних мов

Більшість прикладних областей теж є малоресурсними середовищами. Біомедичний текст, юридичні документи, літературні твори – це все приклади специфічних доменів, для яких моделі, які працюють добре на текстах загального спрямування, показують гірші результати [2]. Більше того, коли знаходиться нове застосування для методів обробки природної мови, дані для тренування та перевірки роботи створених рішень відсутні або дуже дорогі та повинні бути створені силами зацікавлених сторін. Цей факт – одна з найважливіших мотивацій для розробки нових методів обробки природної мови у малоресурсних середовищах.

У лінгвістичній типології прийнято розрізняти добре та недостатньо описані мови. Добре описані мови зазвичай приваблюють більше дослідників; існує безліч граматик і наукових праць, що описують правила та структуру таких мов. Однією з причин класифікації мови як малоресурсна, це неможливість застосування стандартизованих методів, які були створені опираючись на задачу обробки великоресурсних мов. Наприклад мультилінгвістичні мовні моделі на основі Нейронних мереж мають, проблеми з текстами нелатинської писемності, ізольованими мовами та сімействами мов, менш пов'язаними з мовами високого ресурсу [3, С. 17].

Серед основних проблем обробки малоресурсних мов можна виділити наступні:

- *Впровадження системи може вимагати залучення експертів*
  Підходи, які засновані на правилах та евристиках, вимагають значної роботи над лінгвістичними конструкціями і потребують великої кількості роботи експерта. На приклад, у роботі (S. Holoshchuk та ін.) [4, С. 3, 39], автори розглядають граматичні характеристики англійських іменникових та предикатних груп і їх зв'язок з займенниками. Автори виявляють ключові відмінності у синтаксичному та семантичному аналізі тексту і приходять до висновку, що українська мова потребує нових підходів обробки мови та адаптацій вже існуючих.
- *Рішення для специфічних задач є важко досяжними*
  Прикладні задачі у полі малоресурсних мов можуть потребувати специфічних даних (професійна термінологія, діалоги тощо). Однак за визначенням малоресурсності отримати такі дані є складною або неможливою задачею.
- *Проблема оцінювання моделі*
  У випадку, якщо модель можливо побудувати, виникає проблема оцінки точності. Невідомо з чим порівнювати результати моделі та які критерії порівняння (Benchmarks) використовувати, нерідко дослідникам доводиться бути першовідкривачами адекватного методу оцінювання.
- *Теоретична неможливість досягти тими ж методами конкурентної ефективності і точності у порівнянні з великоресурсними мовами*
  Згідно з дослідженням [3, С. 12–15], автори доводять що кількість наявних даних є ключовим аспектом побудови якісної моделі. Мови з меншою кількістю даних також мають і меншу якість даних [3, С. 17]. Нерівномірність кількості даних

у відсотковому співвідношенні, для тренування сучасних мовних моделей, мають проблему в збереженні культурних особливостей, та адекватному моделюванні мови [5].

- *Безпекові проблеми комп'ютерних систем*
  У дослідженні [6], авторам вдалось обійти безпекову систему моделі GPT4 використовуючи команди на мовах, які не були в достатній кількості представлені у тренувальному наборі даних. Дослідники дійшли висновку, що причиною ігнорування моделлю системи безпеки стала проблема невідповідного узагальнення (mismatched generalization), коли вхідні дані не потрапили до корпусу для навчання моделі на безпеку, але перебувають у межах ширшого та різноманітнішого набору даних попереднього навчання моделі на різні завдання.

3. Методи обробки малоресурсних мов

Аналізуючи останні тенденції сфери обробки природної мови, можна зробити висновок, що увага наукової спільноти сфокусована на дослідження великих мовних моделей (LLM), та інших моделей, які використовують технологію мереж трансформерів (Transformer network). Моделі, які використовують простий механізм токенізації і Рекурентних мереж значно програють у точності та ефективності першим. Враховуючи цю інформацію виділимо основні сучасні архітектури та методи обробки малоресурсних мов.

- *Прості багатомовні моделі (Multilingual LM)*
  Прості моделі характеризуються малою кількістю параметрів і невеликим обсягом навчальних даних. Багатомовні моделі, такі як mBERT і XLM-R, замінили традиційні репрезентації (word2vec, fastText, GloVe) завдяки кращому врахуванню контексту. Проте їх точність знижується для малоресурсних мов. У статті [7] автори пропонують неконтрольований підхід для покращення міжмовних репрезентацій цих мов (cross-lingual representations), автоматично отримуючи пари перекладу слів з монолінгвальних корпусів. Модель протестована на кількох мовах, включаючи бенгальську, баскську та непальську, і показала суттєві покращення.
- *Мономовні моделі (Monolingual LM)*
  Мономовні моделі використовують підхід навчання на монолінгвістичному корпусі даних, що стало можливим завдяки архітектурі BERT. Такий підхід дозволяє відійти від

cross-lingual representations, та уникнути неточностей перекладу, жертвуючи при цьому деякими перевагами цього методу. Відносна дешевизна навчання таких моделей дозволила створити більше тисячі мономовних мереж [3, С. 20]. Серед них можна виділити модель української мови LiBERTa [8]. Автори підкреслюють, що LiBERTa демонструє порівнянну продуктивність з попередньою найсучаснішою моделлю NER-UK, з незначним покращенням результатів (+0.03 п.п.). Цікаво, що для цієї задачі друга велика модель XLM-R показує результати гірші за всі базові моделі і має найвищу дисперсію. Цей результат підкреслює необхідність навчання мовно-специфічних моделей. У статті також підкреслюється необхідність створення критеріїв тестування для української мови для покращення майбутніх досліджень.

- *Великі мовні моделі(LLM)*
  Щодо великих мовних моделей, одним з методів є Continual Training, який є варіацією трансферного навчання. Результати досліджень показали, що цей спосіб може покращити виявлення семантичних та синтаксичних зв'язків специфічної мови. Continal Training може використовуватись на невеликих моделях типу BERT, для донавчання моделей типу GPT, LLaMa, BLOOM, застосовується процес низькорангової адаптації (LoRA, QLoRA). Іншим методом є Instruction Fine-tuning, який покращує можливості LLM до виконання інструкцій і допомагає моделям з домінантною англійською мовою значно ефективніше працювати з цільовою. Task-specific fine tuning [9] є способом Instruction tuning для вузькоспеціалізованої задачі. Цей метод також показав хорошу ефективність в багатьох задачах обробки природної мови. Vocabulary Extension також показав кращі результати, в порівнянні з базовою моделлю. У своїй роботі [10], автори комбінують ці методи і приходять до результату, що одночасне застосування Task-specific tuning та Instruction tuning, значно виграють в точності та ефективності всі інші комбінації.

4. Обговорення

Перейдемо до визначення переваг та недоліків представлених рішень. Основними перевагами ***багатомовних моделей*** в контексті малоресурсності є широка доступність попередньо навчених моделей, ефективні міжмовні проєкції, особливо, якщо модель навчена лише на мовах однієї сім'ї, що при критично малій кількості

даних цільової мови робить цей підхід практично єдиним рішенням певної задачі. Більш інтуїтивне донавчання у порівнянні з LLM дає змогу пришвидшувати експерименти та розробку моделей. Варто відмітити, що властивість багатомовності є доволі дискусійним питанням, оскільки у роботі [11], автори доводять, що навчання на великій кількості мов призводить до погіршення результатів моделі на кожній з цих мов. Водночас інші дослідження підтверджують протилежне [12]. Ще однією перевагою є апаратна невибагливість та швидкість навчання з невеликими обчислюваними ресурсами. Серед недоліків можна виділити обмеження на кількість вхідних токенів, збільшення часу виводу результату для великих вхідних послідовностей. Великоресурсні мови мають великий вплив на вивід моделі, гірша точність на вузьконаправлених завданнях специфічної мови.

Щодо ***мономовних моделей***, головною перевагою є ефективна, оптимізація під одну задачу, якісне збереження культурного контексту, через відсутність впливу репрезентацій інших мов. Варто зауважити, що відсутність міжмовних проєкцій в мономовних моделях є і недоліком і перевагою, так як жертвуючи більшою повнотою моделювання мови ми отримуємо відсутність хибних трансляцій та негативного ефекту дисбалансу даних. Оскільки мономовні моделі наслідують ту ж архітектуру, що й прості багатомовні, швидкість навчання та дешевизна експлуатації також зберігаються. Головною причиною більшості недоліків є дані. Можна виділити неможливість досягти адекватної якості при дуже малій кількості даних [13, С. 9–10], важкодоступність даних для формування навчального набору, відсутність адекватного способу тестування моделей [8].

Великі мовні моделі, за наявності великої кількості даних, можуть перевершити інші алгоритми майже в будь-яких задачах. Саме наявність величезного корпусу даних є суттєвим недоліком створення LLM виключно для малоресурсних мов. Останні досягнення дослідників дозволяють налаштувати LLM під вузькі задачі та мови, що є дуже обіцяючим методом. Головними перевагами LLM є мультимодальність, кратно більша кількість параметрів, покращує здатність Великих мовних моделей до навчання. Важливою особливістю є застосування механізмів Zero-shot, One-shot, Few-shot, які дозволяють LLM виконувати задачі на мовах, які були практично відсутні в тренувальних даних [14]. Недоліками LLM є значне споживання обчислювальних ресурсів, для специфічної мови та задачі значно підвищується ризик галюцинацій

та неточних відповідей. Закритість розробок передових мереж Великих мовних моделей є великим недоліком, адже це унеможливлює доступ для досліджень переважній більшості науковців.

Перспективні технології

Архітектура Mamba успішно успадкувала ключові характеристики від трансформерів, як-от увага до контексту та мультимодальність, відкриваючи при цьому нові перспективи для майбутнього розвитку. Здатність Моделі ефективно працювати в різних доменах, особливо в модальностях, де потрібне врахування великого обсягу контексту, як-от геноміка, аудіо та відео, виділяє її серед передових розробок.

Мережа Колгоморова – Арнольда (KAN) останнім часом отримала багато уваги, адже є новою архітектурою багатошарового персептрона, де замість стандартних функцій активацій, мережа вивчає нові, опираючись на дані. Автори роботи [15] реалізували архітектуру Transformer використовуючи KAN як MLP і дійшли до висновку, що при умові адаптації архітектури під задачу, KAN-Transformer має майже ідентичну ефективність у порівнянні з стандартною моделлю. Хоча, варто відмітити вагомий недолік KAN, який, в той же час є його основною особливістю – це вивчення функцій активацій. В контексті великих даних навчання великих мереж, які ґрунтуються на KAN може бути катастрофічно дорогим у часі та обчислювальних ресурсах. Незважаючи на це, ця архітектура залишається перспективним об'єктом дослідження.

5. Висновки

У цій роботі визначено основні проблеми роботи з малоресурсними мовами у контексті їх обробки методами глибокого навчання, було визначено основні сучасні архітектурні рішення цих проблем. В результаті аналізу кожного запропонованого рішення, можна зробити висновок про неоднозначність кожної технології. Спільною проблемою усіх представлених рішень є потреба у великій кількості даних. Мультимовні та великі мовні моделі дають кращі результати за відсутності адекватних даних, ніж мономовні, через можливість навчання на корпусах схожих мов. Водночас мономовні моделі є більш прозорими, передбачуваними та ефективними для вузької задачі. Пропонується підходити до вибору моделей виключно через умови самої задачі та емпіричний метод, адже кожен із

запропонованих підходів не може гарантувати абсолютну перевагу над іншими.

1. V. Berment. Méthodes pour informatiser les langues et les groupes de langues peudotées. (Methods to computerize 'little equipped' languages and groups of languages). Univ. Joseph-Fourier-Grenoble I Diss., 2004. **2.** D. R. Mortensen. Low-Resource NLP. URL: http://demo.clab.cs.cmu.edu/algo4nlp20/slides/low-resource-nlp.pdf. (дата звернення: 12.06.2024). **3.** G. Nicholas, A. Bhatia. CDT Research: Lost in translation: Large Language Models in non-English content analysis. 2023. **4.** V. Vysotska, S. Holoshchuk. A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach. Lviv Polytechnic National University, 2021. **5.** M. Khan, A. Hanna. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. *Forthcoming, 19 Ohio St. Tech. L. J.* 2023. С. 17–20. С. 68–78. **6.** Zheng-Xin Yong, C. Menghini, S. H. Bach. Low-Resource Languages Jailbreak GPT-4. *Department of Computer Science*. Brown University, 2023. С. 3–6. **7.** V. Hangya, H. Shaikh Saadi, A. Fraser. Improving Low-Resource Languages in Pre-Trained Multilingual Language Models. Munich Center for Machine Learning, Germany, 2022. **8.** M. Haltiuk, A. Smywiński-Pohl. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. *AGH*. University of Krakow, Enelpol, 2024. С. 3–9. **9.** P. Budzianowski, I. Vulic. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. Engineering Department, Cambridge University, UK 2019. **10.** C. Toraman. Llama Turk: Adapting Open-Source Generative Large Language Models for Low-Resource Language. *Department of Computer Engineering Middle East Technical University*. Ankara : Turkey, 2024. С. 5–7. **11.** Alexis Conneau et al. Unsupervised cross-lingual representation learning at scale. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. С. 6–9. **12.** Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? *In Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2020. С. 120–130. **13.** T. A. Chang et al. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. Data Science Institute University of California San Diego, 2023. **14.** M. Usman Hadi et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. *MD Anderson Cancer Center*. 2024. С. 20–26. **15.** R. Genet, H. Inzirillo. A Temporal Kolmogorov-Arnold Transformer for Time Series Forecasting. Université Paris Dauphine, PSL. 2024.

REFERENCES:

1. V. Berment. Méthodes pour informatiser les langues et les groupes de

langues peudotées. (Methods to computerize 'little equipped' languages and groups of languages). Univ. Joseph-Fourier-Grenoble I Diss., 2004. **2.** D. R. Mortensen. Low-Resource NLP. URL: http://demo.clab.cs.cmu.edu/algo4nlp20/slides/low-resource-nlp.pdf. (data zvernennia: 12.06.2024). **3.** G. Nicholas, A. Bhatia. CDT Research: Lost in translation: Large Language Models in non-English content analysis. 2023. **4.** V. Vysotska, S. Holoshchuk. A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach. Lviv Polytechnic National University, 2021. **5.** M. Khan, A. Hanna. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. *Forthcoming, 19 Ohio St. Tech. L. J.* 2023. С. 17–20. С. 68–78. **6.** Zheng-Xin Yong, C. Menghini, S. H. Bach. Low-Resource Languages Jailbreak GPT-4. *Department of Computer Science*. Brown University, 2023. С. 3–6. **7.** V. Hangya, H. Shaikh Saadi, A. Fraser. Improving Low-Resource Languages in Pre-Trained Multilingual Language Models. Munich Center for Machine Learning, Germany, 2022. **8.** M. Haltiuk, A. Smywiński-Pohl. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. *AGH*. University of Krakow, Enelpol, 2024. С. 3–9. **9.** P. Budzianowski, I. Vulic. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. Engineering Department, Cambridge University, UK 2019. **10.** C. Toraman. Llama Turk: Adapting Open-Source Generative Large Language Models for Low-Resource Language. *Department of Computer Engineering Middle East Technical University*. Ankara : Turkey, 2024. С. 5–7. **11.** Alexis Conneau et al. Unsupervised cross-lingual representation learning at scale. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. С. 6–9. **12.** Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? *In Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2020. С. 120–130. **13.** T. A. Chang et al. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. Data Science Institute University of California San Diego, 2023. **14.** M. Usman Hadi et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. *MD Anderson Cancer Center*. 2024. С. 20–26. **15.** R. Genet, H. Inzirillo. A Temporal Kolmogorov-Arnold Transformer for Time Series Forecasting. Université Paris Dauphine, PSL. 2024.

Liashko D. A., Post-graduate Student (National University of Water and Environmental Engineering, Rivne)

PROBLEMS OF NATURAL LANGUAGE PROCESSING IN LOW-RESOURCE ENVIRONMENTS. ANALYSIS OF PROMISING SOLUTION METHODS

The problem of technical and architectural solutions for low-resource environments, although important for research, is usually ignored by most scientists who focus on solutions for high-resource environments. The paper analyzes current methods of processing low-resource languages, focusing on the problems and limitations associated with the lack of high-quality language resources and tools. Key challenges such as lack of data, inability to apply standard methods for low-resource languages, model evaluation problems, and security issues are identified. Modern approaches are considered, including the use of Multilingual, Monolingual, and Big Language models, as well as methods for improving learning for these languages, such as cross-lingual representations, task-specific fine-tuning, vocabulary expansion, and others. New, promising Neural Network architectures are considered. For example, Mamba has the potential to replace the standard Transformer model in the future. Kolgomorov – Arnold networks are a fundamentally new architectural solution to the classical multilayer network and can show good efficiency compared to conventional methods. Based on the results of the work, it is concluded that each technology is ambiguous in terms of the efficiency of performing tasks in a low-resource environment. The common problem of all the presented solutions is the need for a large amount of data. Multilingual and Big Language models provide better results in the absence of adequate data than Monolingual models due to the possibility of training on corpora of similar languages. In turn, Monolingual models are more transparent, predictable and efficient for a narrow task.

The conclusion of this article is a recommendation to choose a technology based on the conditions of the task, the amount of data, and the empirical method, since none of the methods has an absolute advantage over the others and may produce unexpected results.

***Keywords*: low-resource language; model; architecture; language model; method.**