

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ІВРИТУ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ: ОБМЕЖЕНІСТЬ ДАНИХ ТА ВИКЛИКИ МУЛЬТИМОВНОСТІ

Вінницький національний технічний університет

Анотація

У статті розглядаються проблеми, пов'язані з використанням іврит у великих мовних моделях (LLM), зокрема обмежена кількість навчальних даних та нерівномірне розподілення моделювальної уваги у мультимовних архітектурах. Проаналізовано сучасні підходи до покращення підтримки малоресурсних мов.

Ключові слова: великі мовні моделі, іврит, малоресурсні мови, мультимовність, нейронні мережі, обробка природної мови.

Abstract

The paper discusses the challenges of using Hebrew in large language models (LLMs), particularly the limited amount of training data and uneven distribution of modeling capacity in multilingual architectures. Modern approaches to improving low-resource language support are analyzed.

Keywords: large language models, Hebrew, low-resource languages, multilinguality, neural networks, natural language processing.

Проблема малоресурсності івриту

Сучасні великі мовні моделі (LLM) демонструють високі результати у завданнях генерації, перекладу та класифікації текстів. Проте ефективність їхньої роботи значною мірою залежить від обсягу наявних навчальних даних. У випадку івриту, як і багатьох інших малоресурсних мов, цієї умови не дотримано.

Іврит є офіційною мовою Держави Ізраїль, і загальна кількість носіїв оцінюється приблизно у 9 мільйонів осіб, з яких близько 5 мільйонів активно використовують мову у повсякденному житті, освіті та професійній діяльності. Незважаючи на це, за поширеністю і представленістю в цифровому просторі іврит значно поступається іншим мовам з аналогічною кількістю носіїв, зокрема шведській чи чеській, через відносно малий обсяг доступного текстового контенту [1].

У корпусі ROOTS, який використовувався для навчання моделі BLOOM (близько 341 мов та їх варіацій), частка івриту становить лише 0,09% (див. рис. 1). Це приблизно 6 мільярдів токенів із загального обсягу у 6,3 трильйона [2]. Для порівняння, англійська мова займає понад 40% корпусу, а французька і німецька — близько 3–4% кожна. Така непропорційність суттєво впливає на здатність моделі формувати коректні та інформативні висловлювання мовами з меншою представленістю.

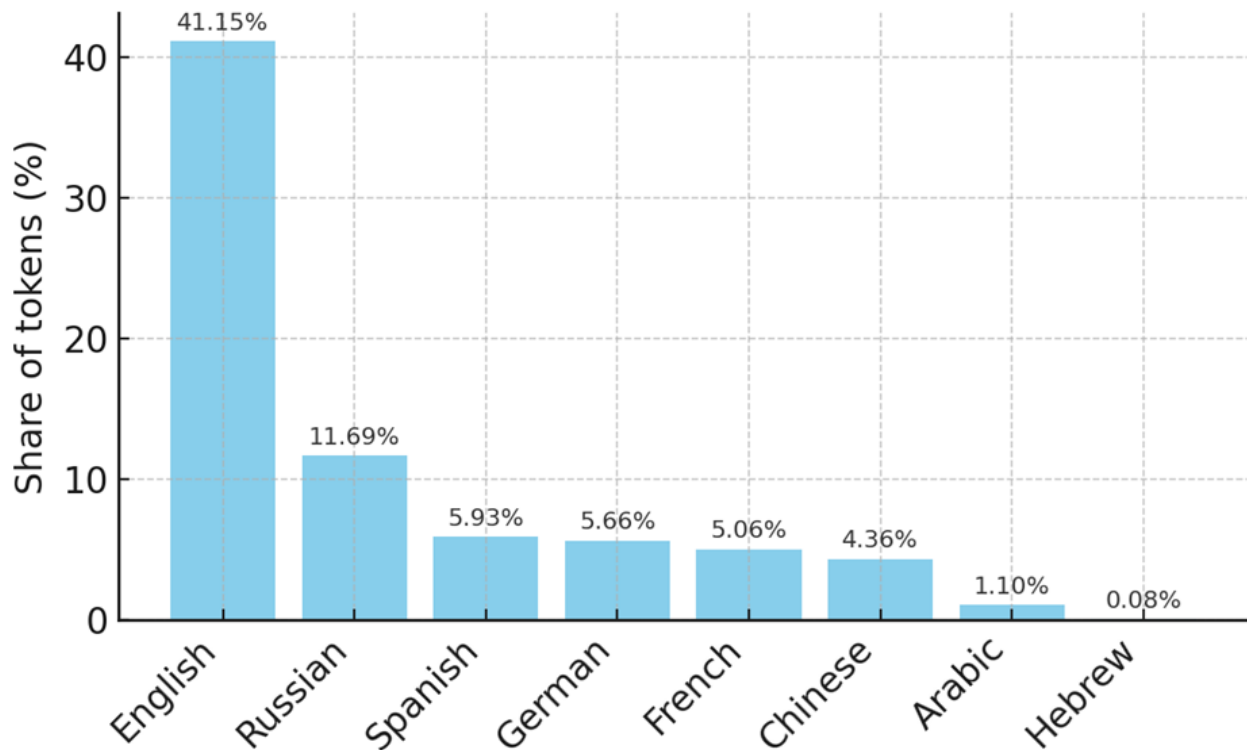


Рис. 1 – Частка івриту у корпусі для навчання ROOTS

У більшості LLM, зокрема GPT-3 або LLaMA, переважну частину навчальних даних становлять англomовні джерела, тоді як тексти івритом здебільшого обмежуються релігійною, енциклопедичною та новинною тематиками [3]. У GPT-3 англійська мова представлена у понад 92% корпусу, тоді як всі інші мови разом — менше 8% [4]. У LLaMA 2 (Meta, 2023) ситуація є подібною: іврит був представлений у вкрай обмеженому обсязі, що не дозволило моделі належним чином засвоїти особливості цієї мови на рівні синтаксису та семантики [5].

Ця недостатність ресурсів зумовлює зниження точності відповідей івритом, особливо в завданнях генерації, пошуку відповідей та машинного перекладу [6]. Наприклад, у міжмовному порівнянні продуктивності моделей GPT-3.5 та GPT-4 за результатами виконання однакових запитань різними мовами, іврит демонстрував на 18–25% нижчий рівень коректності відповідей порівняно з англійською, французькою або іспанською [3].

Морфологічна складність івриту та проблеми токенизації

Значним викликом у контексті обробки івриту є токенизація — процес поділу тексту на окремі елементи (токени), які можуть бути опрацьовані моделлю. Більшість сучасних токенизаторів спроектовано для мов з аналітичною структурою, передусім англійської, що ускладнює їх застосування до флективних та аглютинативних мов, таких як іврит.

Іврит має багату морфологічну структуру: численні словоформи формуються шляхом поєднання коренів із шаблонами, використанням префіксів, суфіксів і частих випадків злиття службових слів (прийменників, артиклів, займенників) з основою. Це призводить до появи складних лексичних одиниць,

які стандартні токенизатори часто розбивають на надмірну кількість фрагментів. Така сегментація збільшує довжину вхідної послідовності та ускладнює навчання, знижуючи ефективність представлення тексту.

Додаткову складність становить відсутність нікудації (вокалізації) — діакритичних позначок голосних звуків. У більшості сучасних текстів іврити ці позначки опускаються, однак у низці випадків саме вони визначають семантичне значення слова. Моделі, що не враховують нікудацію, можуть припускатися серйозних помилок у розпізнаванні та аналізі лексичних одиниць [7].

Виклики мультимовних архітектур

Ще одним системним обмеженням є архітектура мультимовних моделей. У спробі забезпечити підтримку десятків мов за допомогою спільної моделі, обчислювальні ресурси розподіляються нерівномірно. Це призводить до того, що мови з більшим обсягом навчальних даних (наприклад, англійська, іспанська) отримують більшу «вагу» під час навчання, тоді як малоресурсні мови, зокрема іврит, залишаються недостатньо репрезентованими [8]. Це явище описується терміном «виклик мультимовності» [9].

Дослідницькі роботи демонструють, що під час спільного навчання моделі на багатьох мовах може виникати негативна міжмовна інтерференція: покращення якості на одній мові супроводжується погіршенням на іншій. Іврит, як мова, що належить до семітської групи та є типологічно відмінною від більшості індоевропейських мов, особливо вразливий до таких ефектів. У практиці генерації текстів спостерігались випадки, коли модель змішувала синтаксичні структури або навіть вставляла фрагменти іншої мови, наприклад англійської.

Поточні підходи до вирішення проблем

Серед сучасних стратегій подолання зазначених проблем — розширення словників моделей із врахуванням морфологічних особливостей, спеціалізоване донавчання на корпусах цільовою мовою, а також використання мовних адаптерів (adapter layers) і модульних архітектур [10].

Яскравим прикладом успішної адаптації є модель DictaLM 2.0, побудована на основі LLaMA та донавчена на корпусі обсягом понад 35 мільярдів слів івритом [3]. Вона продемонструвала суттєве зростання якості генерації, класифікації та перекладу івритом у порівнянні з універсальними мультимовними LLM. Іншим підходом є застосування адаптивних токенизаторів, які враховують морфологічні особливості конкретної мови, а також методів distillation — перенесення знань із великої універсальної моделі до компактної, але цільової мовної архітектури.

Перспективи розвитку

На сьогодні активно розвиваються ініціативи зі створення відкритих репрезентативних корпусів і бенчмарків для іврити. Наприклад, Hebrew LLM Leaderboard дозволяє порівнювати якість роботи моделей саме на івритських завданнях [3]. Також формуються й удосконалюються корпуси, як-от Hebrew Oscar та добірки із проєктів Wikimedia.

Очікується, що наступне покоління мультимовних LLM впроваджуватиме механізми динамічного розподілення параметрів залежно від мови, що сприятиме підвищенню інклюзивності та збалансованості якості для малоресурсних мов. Такий підхід може забезпечити кращу підтримку для іврити та інших типологічно відмінних мов.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Ethnologue: Languages of the World*. 26-е вид. Dallas, TX: SIL International, 2023. – С. 539.
2. Mou C., Nguyen A., Shao Y. та ін. *CulturaX: Investigating Large-Scale Multilingual Data for LLMs* [Електронний ресурс]. – arXiv:2305.14556, 2023. – Режим доступу: <https://arxiv.org/abs/2305.14556>
3. Shmidman S., Shmidman A., Koppel M. та ін. *Adapting LLMs to Hebrew: Unveiling DictaLM 2.0* [Електронний ресурс]. – arXiv:2401.06230, 2024. – Режим доступу: <https://arxiv.org/abs/2401.06230>
4. Brown T., Mann B., Ryder N. та ін. *Language Models are Few-Shot Learners* [Електронний ресурс]. – arXiv:2005.14165, 2020. – Режим доступу: <https://arxiv.org/abs/2005.14165>
5. Touvron H., Lavril T., Izacard G. та ін. *LLaMA 2: Open Foundation and Fine-Tuned Chat Models*. [Електронний ресурс]. – arXiv:2307.09288, 2023. – Режим доступу: <https://arxiv.org/abs/2307.09288>
6. Eyal M., Noga H., Aharoni R., Szpektor I., Tsarfaty R. *Multilingual Sequence-to-Sequence Models for Hebrew NLP* // Findings of the Association for Computational Linguistics: ACL 2023. – 2023. – С. 7700–7708.
7. Tsarfaty R., Seker A., Sadde S., Klein S. *What's Wrong with Hebrew NLP? And How to Make it Right* [Електронний ресурс]. – arXiv:1903.10682, 2019. – Режим доступу: <https://arxiv.org/abs/1903.10682>
8. Wang Z., Lipton Z., Tsvetkov Y. *On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment* [Електронний ресурс]. – arXiv:2010.03017, 2020. – Режим доступу: <https://arxiv.org/abs/2010.03017>
9. Conneau A., Lample G., Ranzato M. та ін. *Unsupervised Cross-lingual Representation Learning at Scale* // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020. – С. 8440–8451.
10. Pfeiffer J., Rücklé A., Gurevych I., Cho K. *AdapterFusion: Non-Destructive Task Composition for Transfer Learning* [Електронний ресурс]. – arXiv:2005.00247, 2020. – Режим доступу: <https://arxiv.org/abs/2005.00247>

Недашківський Євген Анатолійович – аспірант кафедри комп'ютерних наук, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця., e-mail: yevhen@allstarsit.com.

Науковий керівник: **Колесницький Олег Константинович** – к.т.н., доцент, професор кафедри факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця., e-mail: kolesnytskiy@vntu.edu.ua.

Yevhen Anatoliiovych Nedashkivskiy – PhD student at the Department of Computer Science, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine. E-mail: yevhen@allstarsit.com.

Scientific advisor: **Oleh Kostiantynovych Kolesnytskiy** – PhD in Engineering, Associate Professor, Professor at the Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine. E-mail: kolesnytskiy@vntu.edu.ua.