УДК 81’322.4
DOI: 10.18523/2617-3808.2025.8.97-101

Махаммедов Ж. Ж., Кирієнко О. В., Ткаченко В. О.

ОЦІНКА ТРАНСФОРМЕРНИХ МОДЕЛЕЙ mT5 ДЛЯ УКРАЇНСЬКО-АНГЛІЙСЬКОГО ПЕРЕКЛАДУ

Цю статтю присвячено кількісному вивченню впливу розміру архітектури Transformer на точність українсько-англійського машинного перекладу з використанням моделі mT5. Досліджено ефективність роботи моделей mT5 різних розмірів (small, base, large) щодо часу навчання, часу генерації перекладів і якості перекладу, оціненої метриками BLEU та chrF++. Результати показують, що більші моделі mT5 демонструють вищу якість перекладу, але потребують значно більше обчислювальних ресурсів. Результати дослідження підтверджують доцільність застосування моделей mT5 для українсько-англійського перекладу, навіть на типових обчислювальних системах.

Ключові слова: трансформер, оброблення природної мови, машинний переклад, нейронний машинний переклад, mT5, HPLT, BLEU, chrF++, NLLB-200.

Вступ

Машинний переклад є однією з ключових галузей оброблення природної мови. Історично ця сфера зазнала значної еволюції, здійснивши перехід від ранніх систем, основаних на правилах, до статистичного машинного перекладу [5], який домінував на початку XXI століття. Проте парадигматичний зсув відбувся з появою нейронного машинного перекладу, який продемонстрував здатність генерувати значно більш якісні, контекстуально обґрунтовані та точні переклади. На відміну від статистичного машинного перекладу, що оперував на рівні фраз, перші моделі нейронного машинного перекладу на базі рекурентних нейронних мереж дали можливість обробляти речення як єдине ціле, що стало значним кроком уперед [1].

Ключовим моментом, що визначив сучасний стан нейронного машинного перекладу, стало впровадження архітектури Transformer у 2017 р. [8]. Ця архітектура відмовилася від рекурентних і згорткових шарів на користь механізмів уваги, зокрема самоуваги (self-attention). Такий підхід дозволив моделі зважувати важливість різних слів у вхідній послідовності та одночасно обробляти всі її елементи. Можливість паралельного оброблення даних не лише кардинально прискорила процес навчання та перекладу, а й значно підвищила ефективність у захопленні довготривалих залежностей у тексті. Завдяки своїй ефективності та масштабованості, архітектура Transformer стала фундаментальним будівельним блоком для переважної більшості сучасних великих мовних моделей і де-факто стандартом для найсучасніших систем машинного перекладу.

Останнім часом популярність багатомовних моделей значно зросла, цю сферу також активно досліджують і в Україні [4], однією з таких моделей є mT5 (Massively Multilingual Text-to-Text Transfer Transformer) [9]. Дослідження їхньої ефективності для менш поширених або складних мов, зокрема української, залишаються актуальними. Хоча автори mT5 у своїй статті охоплюють загальні відомості про роботу моделі та демонструють загальні спроможності архітектури, вони не надали детального аналізу або тестування, сфокусованих на задачі перекладу. Це створює прогалину в розумінні того, як різні розміри моделей mT5 проявляють себе у таких задачах, тому ця робота прагне дослідити цю проблему й оцінити можливості моделі саме для українсько-англійського перекладу.

Передумови дослідження

Модель mT5 є багатомовною версією моделі T5 (Text-to-Text Transfer Transformer), головною особливістю якої є уніфікація завдань з оброблення природної мови у формат «text-to-text», що дозволяє використовувати одну архітектуру кодера-декодера для вирішення широкого спектра завдань: від машинного перекладу та узагальнення до відповідей на запитання та класифікації тексту. Модель

mT5 була попередньо навчена (pre-trained) на багатомовному корпусі даних mC4, що дає їй можливість ефективно працювати зі 101 мовою. Це робить її надзвичайно цінним інструментом для завдань, що потребують роботи з текстами або генерації тексту різними мовами.

Мовний корпус mC4 (Massively Multilingual Common Crawl Cleaned Corpus) охоплює 101 мову та має обсяг у 27 терабайтів. Цей датасет містить 2733 мільярди токенів англійською і 41 мільярд токенів українською, що становить 5,67 % і 1,51 % відповідно. Він є багатомовною версією корпусу C4 (Colossal Clean Crawled Corpus), який використовувався для навчання оригінальної моделі T5. Обидва корпуси базуються на даних, зібраних проєктом Common Crawl, що систематично сканує веб.

Моделі mT5 існують у 5 розмірах: mT5-small, mT5-base, mT5-large, mT5-xl, mT5-xxl. Розмір моделі Transformer визначається кількістю її параметрів: кількістю блоків у кодері та декодері, розмірністю векторів ембедінгу, розмірністю матриці проєкції ключів / значень, кількістю голів уваги, розмірністю проміжного вектора всередині блоку трансформера. Ці параметри є критично важливим фактором, що впливає на продуктивність моделі та її обчислювальні вимоги. Моделі Transformer можуть варіюватися від сотень мільйонів до сотень мільярдів параметрів, що безпосередньо корелює з їхньою здатністю до захоплення складних мовних закономірностей і контекстів. Зазвичай, що більша модель і що більше даних вона обробляє під час навчання, то кращої якості перекладу вона може досягти. Однак збільшення розміру тягне за собою значне зростання обчислювальних ресурсів та пам'яті, необхідних для навчання та розгортання. Це створює важливий компроміс між бажаною точністю перекладу та практичною можливістю використання моделі на доступному обладнанні.

Методологія дослідження

Для оцінки було використано три версії mT5 — small (300 млн параметрів, 1,2 ГБ), base (580 млн параметрів, 2,4 ГБ), large (1,3 млрд параметрів, 4,9 ГБ). Ці версії було обрано, оскільки більшість споживчих графічних прискорювачів мають об'єм відеопам'яті від 8 до 12 ГБ, а отже більші моделі, як-от mT5-xl (3,7 млрд, 15 ГБ), не вмістилися б у такий об'єм.

Для кількісної оцінки якості машинного перекладу використовуються автоматичні метрики, які порівнюють згенерований моделями переклад з еталонним перекладом з корпусу. У цьому дослідженні для оцінки були обрані дві широко використовувані метрики: BLEU та chrF++, вони дають змогу швидко та об'єктивно порівнювати різні моделі.

BLEU (Bilingual Evaluation Understudy) [6] — це одна з найдавніших і найпоширеніших метрик для оцінки машинного перекладу, представлена у далекому 2002 році. Вона вимірює схожість між машинним і еталонним перекладами шляхом розрахунку модифікованої точності для n-грам. Кінцеве значення BLEU є геометричним середнім модифікованих точностей для n-грам різної довжини, помноженим на штраф за стислість. Значення BLEU лежить у діапазоні від 0 до 100, де вищі показники вказують на кращу якість.

chrF++ (character n-gram F-score) [7] — метрика, розроблена 2017 року як розширення оригінальної chrF. Вона працює на рівні символьних n-грам і додатково враховує словесні уніграми та біграми, що дозволяє враховувати як дрібні лінгвістичні деталі, так і ширший контекст. Як F-міра, chrF++ балансує між точністю та повнотою.

Для тонкого налаштування (fine-tuning) моделей було використано англійсько-українську частину корпусу HPLT v2 (High-Performance Language Technologies version 2) [2]. Цей корпус є розширеним та очищеним багатомовним набором даних, спеціально розробленим для задач оброблення природної мови. Його вибір зумовлений високою якістю і релевантністю для завдань машинного перекладу. Корпус було відфільтровано за показниками score-aligner не менше ніж 0,4 і score-bicleaner не менше ніж 0,9 для збільшення якості тренувальних даних. Після цього з отриманих пар було взято 10 % випадковим чином. Усього для навчання моделей було використано 1 068 199 пар речень.

Для оцінки якості перекладів був використаний тестовий набір даних із 10 000 українсько-англійських речень, відібраних з корпусу HPLT v2, які не використовувалися під час тонкого налаштування. Вхідні українські речення отримували префікс «translate Ukrainian to English: ». Оцінку BLEU проводили з параметром «tokenize='flores200'». Для chrF++ були вказані параметри «word_order=2» та «char_order=6», що відповідає стандартним налаштуванням.

Для розуміння результатів роботи mT5 були включені дві дистильовані моделі з сімейства NLLB-200 (No Language Left Behind) [3]: NLLB-200-distilled-600M та NLLB-200-distilled-1.3B. Обидві моделі NLLB-200 базуються на архітектурі Transformer і були навчені за допомогою техніки дистиляції,

коли менша модель вчиться на виходах більшої моделі, вони навчалися за допомогою моделі NLLB-MOE-54B (54 млрд параметрів, 250 ГБ), що дозволяє їм перекладати безпосередньо між будь-якими з 200 підтримуваних мов без використання проміжної мови.

Апаратне та програмне забезпечення

Процес тонкого налаштування моделей mT5, який потребує значних обчислювальних ресурсів і об'ємів відеопам'яті, виконувався на графічному прискорювачі Nvidia RTX 5090. Відеокарту було обрано через високу продуктивність (104.8 TFLOPS) і великий обсяг відеопам'яті (32 ГБ), вона є найпотужнішим споживчим графічним прискорювачем на момент проведення дослідження. Етап генерації перекладів для оцінки моделей виконувався на графічному прискорювачі Nvidia RTX 3070, який добре відображає умови типових сучасних обчислювальних систем. Цей графічний прискорювач має меншу продуктивність (20.31 TFLOPS) і відеопам'ять (8 ГБ) порівняно з RTX 5090, але його можливостей було достатньо для швидкого отримання перекладів на тестовому наборі даних.

Програмне середовище було побудовано на основі мови Python і провідних бібліотек для глибинного навчання:

- PyTorch — основний фреймворк для глибинного навчання, забезпечує гнучкість у роботі з нейронними мережами та оптимізовану роботу з GPU;
- Hugging Face Transformers — для взаємодії з попередньо навченими моделями Transformer, такими як mT5, і спрощення процесу тонкого налаштування;
- Hugging Face Datasets — для ефективної роботи з наборами даних;
- sacrebleu — інструмент для автоматичної оцінки машинного перекладу.

Результати

Дані, отримані після попереднього оброблення даних і тонкого налаштування моделей mT5, наведено в табл. 1 і 2.

Таблиця 1. **Час тренування моделей на графічному прискорювачі Nvidia RTX 5090**

Модель	Кількість параметрів	Час тренування (години)
mT5-small	300 млн	6:06
mT5-base	580 млн	15:20
mT5-large	1.2 млрд	48:39

Таблиця 2. **Оцінка перекладу моделей на графічному прискорювачі Nvidia RTX 3070**

Модель	Кількість параметрів	BLEU	chrF++	Час генерації прикладів (години)
mT5-small	300 млн	41,53	62,33	0:14
mT5-base	580 млн	49,97	68,17	0:21
mT5-large	1.2 млрд	54,50	71,13	1:06
NLLB-200-distilled-600M	600 млн	42,89	63,44	0:27
NLLB-200-distilled-1.3B	1.3 млрд	46,61	66,90	1:41

Аналіз табл. 2 засвідчує, що більші моделі mT5 демонструють вищу якість перекладу за обома метриками. Найкращі результати серед моделей mT5 показала mT5-large, що підтверджує гіпотезу про те, що збільшення розміру моделі позитивно корелює з якістю перекладу.

Порівняно з моделями сімейства NLLB-200 тонко налаштовані моделі mT5 демонструють кращі результати. NLLB-200-distilled-600M показала, що вона працює трохи краще, ніж mT5-small. Однак mT5-base перевершує NLLB-200-distilled-600M за обома метриками, незважаючи на схожий розмір. Найцікавіше, що NLLB-200-distilled-1.3B досягла другого за якістю результату, хоча і була найбільшою моделлю; mT5-large значно перевершує показники NLLB-200-distilled-1.3B, що свідчить про ефективність тонкого налаштування mT5 на спеціалізованому українсько-англійському корпусі даних і демонструє перевагу моделей вузького спрямування.

Отримані результати чітко ілюструють компроміс між якістю перекладу та обчислювальними ресурсами. Хоча mT5-large забезпечує найвищу якість перекладу, вона потребує суттєво більше часу для навчання та генерації порівняно з меншими моделями. Водночас mT5-base демонструє значне

покращення якості порівняно з mT5-small за помірного збільшення часу навчання та генерації, що робить її привабливим вибором для сценаріїв, де важливий баланс між якістю і доступністю ресурсів. Навіть на типових обчислювальних системах, що представлені, mT5-small, mT5-base і навіть mT5-large показують прийнятний час генерації перекладів.

Для глибшого розуміння якісних відмінностей у перекладах, згенерованих моделями різних розмірів, було проведено ручний аналіз 50 речень. У табл. 3 наведено розбір одного типового речення.

Таблиця 3. **Приклади перекладу**

Модель	Оригінал речення	Еталонний переклад	Переклад моделі
mT5-small	Швидка бура лисиця перестрибує через лінивого собаку.	A quick brown fox jumps over a lazy dog.	Quick blade of the fox spreads through a linear dog.
mT5-base	Швидка бура лисиця перестрибує через лінивого собаку.	A quick brown fox jumps over a lazy dog.	Quick brown fox jumps through a lazy dog.
mT5-large	Швидка бура лисиця перестрибує через лінивого собаку.	A quick brown fox jumps over a lazy dog.	The fast brown fox jumps over a lazy dog.

Почнемо з mT5-small. Його переклад значною мірою некоректний. Модель не змогла правильно передати значення багатьох слів («бура» — «blade», «перестрибує» — «spreads», «лінивого» — «linear») і загалом спотворила зміст речення. Спостерігається втрата сенсу та граматичні помилки. Модель mT5-base показує значно кращий переклад, ніж mT5-small. Більшість ключових фраз перекладено правильно: «quick brown fox», «lazy dog». Проте є контекстуальна помилка у прийменнику «через» — «through» замість «over», що спотворює сенс речення. Модель mT5-large створила переклад речення, дуже схожий на еталонний, навіть контекстуально правильно переклала слово «через». Варто зазначити, що mT5-large іноді дає більш якісний переклад, ніж еталон з HPLT, mT5-base зазвичай тримається на рівні еталону, а mT5-small часто показує гірший результат, ніж еталон, хоча і ненабагато.

Ці результати підкреслюють доцільність використання моделей mT5 для українсько-англійського перекладу після тонкого налаштування.

Висновок

Це дослідження було присвячене кількісному вивченню впливу розміру моделі mT5 на точність українсько-англійського перекладу. Ми проаналізували продуктивність трьох версій моделі mT5 — small, base та large — з огляду на час навчання, час генерації перекладів та якість перекладу, використовуючи метрики BLEU та chrF++. Для повнішого контексту результати моделей mT5 було порівняно з показниками дистильованих версій NLLB-200.

Наше тестування підтвердило значний вплив розміру моделі Transformer на точність. Важливо зазначити, що mT5-large показала кращий результат та працювала швидше, ніж навіть більша модель широкої направленості NLLB-200-distilled-1.3B після тонкого налаштування на корпусі у мільйон мовних пар. Це демонструє користь моделей вузького направлення, які спеціалізовані на конкретній мовній парі та завданні й тому можуть перевершувати більш універсальні рішення.

Результати цього дослідження показують доцільність використання моделей mT5 для українсько-англійського перекладу навіть на типових обчислювальних системах.

Список літератури

1. Bahdanau D. Neural Machine Translation by Jointly Learning to Align and Translate [Electronic resource] / D. Bahdanau, K. Cho, Y. Bengio // arXiv. — 2014. — Mode of access: https://arxiv.org/pdf/1409.0473. — Title from screen.
2. Burchell L. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT) [Electronic resource] / L. Burchell, O. Gilbert, et al. // arXiv. — 2025. — Mode of access: https://arxiv.org/pdf/2503.10267. — Title from screen.
3. Costa-jussa M. No Language Left Behind: Scaling Human-Centered Machine Translation [Electronic resource] / M. Costa-jussa, J. Cross, et al. // arXiv. — 2022. — Mode of access: https://arxiv.org/pdf/2207.04672. — Title from screen.
4. Glybovets M. Natural Language Processing Using Large Language Models and Machine Learning Methods [Електронний ресурс] / M. Glybovets, D. Zadokhin, et al. // Наукові записки НаУКМА. Комп'ютерні науки. — 2024. — Т. 7. — С. 102–111. — Режим доступу: https://doi.org/10.18523/2617-3808.2024.7.102-111. — Назва з екрана.
5. Och F. J. Statistical Machine Translation. European Association for Machine Translation [Electronic resource] / F. J. Och, H. Ney. — Ljubljana, Slovenia, 2000. 5th EAMT Workshop: Harvesting Existing Resources. — Mode of access: https://aclanthology.org/2000.eamt-1.5.pdf. — Title from screen.
6. Papineni K. Bleu: a Method for Automatic Evaluation of Machine Translation [Electronic resource] / K. Papineni, S. Roukos, et al. — Association for Computational Linguistics. Philadelphia, Pennsylvania, USA, 2002. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — Pp. 311–318. — Mode of access: https://aclanthology.org/P02-1040.pdf. — Title from screen.

7. Popović M. chrF++: words helping character n-grams [Electronic resource] / M. Popović. — Association for Computational Linguistics. Copenhagen, Denmark, 2017. Proceedings of the Second Conference on Machine Translation. — Pp. 612–618. — Mode of access: https://aclanthology.org/W17-4770.pdf. — Title from screen.
8. Vaswani A. Attention Is All You Need [Electronic resource] / A. Vaswani, N. Shazeer, et al. // arXiv. — 2017. — Mode of access: https://arxiv.org/pdf/1706.03762. — Title from screen.
9. Xue L. mT5: A massively multilingual pre-trained text-to-text transformer [Electronic resource] / L. Xue, N. Constant, et al. // arXiv. — 2020 — Mode of access: https://arxiv.org/pdf/2010.11934. — Title from screen.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv. https://arxiv.org/abs/1409.0473.

Burchell, L., Gilbert, O., Arefyev, N., Aulamo, M., Banon, M., Chen, P., Fedorova, M., Guillou, L., Haddow, B., Haije, J., Helel, J., Hentiksson, E., Klimaszewski, M., Komulainen, V., Kutuzov, A., Kytoniemi, J., Laippala, V., Maehlum, P., Malik, B., ... Zaragoza-Bernabeu, J. (2025). *An expanded massive multilingual dataset for high-performance language technologies (HPLT)*. arXiv. https://arxiv.org/abs/2503.10267.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia, G., Hansanti, P., Hoffman, J., ... Wang, J. (2022). *No language left behind: Scaling human-centered machine translation*. arXiv. https://arxiv.or/g/abs/2207.04672.

Glybovets, M., Zadokhin, D., Dekhtiar, B., & Pyechkurova, O. (2024). Natural language processing using large language models and machine learning methods. *NaUKMA Research Papers. Computer Science*, *7*, 102–111. https://doi.org/10.18523/2617-3808.2024.7.102-111.

Och, F. J., & Ney, H. (2000). Statistical machine translation. In *Proceedings of the 5th EAMT Workshop*: *Harvesting Existing Resources*. European Association for Machine Translation. https://aclanthology.org/2000.eamt-1.5.pdf.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://aclanthology.org/P02-1040.pdf.

Popović, M. (2017). chrF++: Words helping character n-grams. *In Proceedings of the Second Conference on Machine Translation* (pp. 612–618). Association for Computational Linguistics. https://aclanthology.org/W17-4770.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv. https://arxiv.org/abs/1706.03762.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). *mT5: A massively multilingual pre-trained text-to-text transformer.* arXiv. https://arxiv.org/pdf/2010.11934.

Z. Makhammedov, O. Kyriienko, V. Tkachenko

EVALUATING MT5 TRANSFORMER MODELS FOR UKRAINIAN-ENGLISH TRANSLATION

This study quantitatively investigates the impact of Transformer architecture size on the accuracy of Ukrainian-to-English machine translation using the multilingual mT5 model. The research evaluates three distinct mT5 versions (small, base, and large) by fine-tuning them on a subset of the HPLT v2 corpus. The technical implementation relied on a standard Python-based deep learning stack, utilizing PyTorch and Hugging Face (Transformers, Datasets) libraries for model management and training. Fine-tuning was executed on a high-performance GPU to handle the significant computational load, while inference speed was benchmarked on a typical consumer-grade GPU to reflect real-world deployment scenarios. Translation quality was assessed using the standard BLEU and chrF++ metrics.

The results confirm a direct correlation between model size and translation quality, with larger models consistently achieving higher scores on both evaluation metrics. This improved accuracy, however, comes at the cost of significantly increased computational demand for both training and inference. Notably, when benchmarked against other large-scale, general-purpose translation models such as NLLB-200 (distilled-600M and distilled-1.3B), the fine-tuned mT5 variants demonstrated superior performance for Ukrainian-English translation, underscoring the benefits of task-specific adaptation. This study confirms the feasibility of using all three mT5 models for this task on typical computing systems, presenting users with a clear trade-off between desired translation quality and available resources.

Keywords: transformer, natural language processing, machine translation, neural machine translation, mT5, HPLT, BLEU, chrF++, NLLB-200.

Матеріал надійшов 27.06.2025