UDC 004.415.532 DOI: https://doi.org/10.18664/ikszt.v30i4.351489

ГАВРАШЕНКО А. О.
Аспірант кафедри електронних обчислювальних машин Харківського національного університету радіоелектроніки
БАРКОВСЬКА О. Ю.
К.т.н, доц.кафедри електронних обчислювальних машин Харківського національного університету радіоелектроніки

Інтелектуальна інформаційна система перекладу штучних мов

Анотація. *Стаття присвячена розробленню інтелектуальної інформаційної системи перекладу штучних мов із використанням методів штучного інтелекту і статистичних алгоритмів, пропонуючи інноваційний підхід для підвищення ефективності машинного перекладу в умовах динамічного розвитку сучасних технологій. Метою роботи є розроблення інтелектуальної інформаційної системи перекладу штучних мов. Дослідницькою складовою в запропонованій системі є оцінювання ефективності використання методів штучного інтелекту для перекладу штучних і звичайних мов, що може призвести до скорочення часових і ресурсних затрат, а також забезпечити високу точність і надійність результатів. У роботі досліджено застосування довгої короткочасної пам'яті (LSTM) для перекладу. Результати демонструють, що LSTM досягає точності до 97.5 % для контрольованої вибірки вхідних даних і до 93 % на випадковому датасеті. Отримані результати підтверджують ефективність розробленої системи, що дає змогу з високою ефективністю перекладати звичайні та штучні мови.*

Ключові слова: *машинне навчання, штучний інтелект, LSTM, датасети, енкодер, декодер, послідовність у послідовність.*

Вступ

У сучасному світі обсяги міжкультурної комунікації стрімко зростають. Глобалізація, міжнародна співпраця, розвиток інформаційних технологій і поширення інтернету зумовлюють дедалі більшу потребу у швидкому, точному та масштабованому перекладі текстів різними мовами. Традиційні методи перекладу, як-от робота людини-перекладача або системи, що базовані на правилах (rule-based systems), хоча й забезпечують високу якість у вузькоспеціалізованих контекстах, не здатні ефективно впоратися з обробкою великих обсягів тексту в реальному часі.

Через це особливої актуальності набувають автоматизовані системи машинного перекладу. Протягом останніх десятиліть ця галузь зазнала суттєвих змін: від статистичних методів (Statistical Machine Translation, SMT) до сучасних підходів, які базовані на штучних нейронних мережах (Neural Machine Translation, NMT). Розвиток нейромережевих технологій кардинально змінив підходи щодо автоматичного перекладу.

На відміну від статистичних методів, які засновані на поєднанні ймовірнісних моделей і частотних залежностей, нейронні мережі допомагають моделювати переклад як складну трансформацію послідовностей з урахуванням контексту, граматики та семантики.

Нейронний машинний переклад побудований на глибокому навчанні (deep learning) — підгалузі машинного навчання, яка використовує багаторівневі штучні нейронні мережі для обробки даних. Основними архітектурами, що використані в системах NMT, є рекурентні нейронні мережі (RNN), довгі короткочасні пам’яті (LSTM), а згодом трансформери (Transformers), які сьогодні вважають найефективнішими для завдань обробки природної мови (Natural Language Processing, NLP). Саме завдяки трансформерам були створені такі успішні моделі, як Google Translate [1], DeepL [2], ChatGPT [3] та інші.

Сучасні системи перекладу, що використовують трансформери, уміють не лише перекладати окремі слова або фрази, але й розуміти

зміст тексту, виявляти приховані значення, граматичні конструкції, стилістичні особливості мови. Вони здатні підтримувати узгодженість термінології в межах одного документа, розпізнавати неоднозначності, а іноді навіть ураховувати контекст попередніх речень або діалогів. Такий рівень розуміння робить нейромережі надзвичайно потужним інструментом для вирішення завдань автоматичного перекладу [4].

Однак, попри досягнення в цій сфері, ще існує чимало викликів. Нейромережеві моделі потребують великих обсягів даних для навчання, значних обчислювальних ресурсів і можуть бути схильними до помилок у разі перекладу рідковживаних мов, технічних текстів або культурно-специфічних висловів. Крім того, виникають питання етичного характеру, як-от авторське право на перекладені тексти або упередження, що виникають через асиметричні дані у процесі навчання.

Ця робота присвячена дослідженню процесу перекладу за допомогою нейронних мереж. У ній розглянуто базові принципи функціонування моделей NMT, проаналізовано архітектури, що найчастіше використовують для перекладу тексту, і розглянуто сучасні тенденції в цій галузі. Особливу увагу приділено практичним аспектам побудови і тренування нейромережевих перекладачів, їх оцінювання.

Мета дослідження полягає у створенні інтелектуальної інформаційної системи перекладу штучних мов. У межах роботи буде також зроблено спробу побудувати і протестувати власну нейромережеву модель перекладу на базі відкритих корпусів даних і машинного навчання.

Аналіз останніх досліджень і публікацій

У роботах [5-7] було проведено значну кількість досліджень, спрямованих на покращення якості нейронних машинних перекладачів. Класифікація типів перекладачів наведена на рис. 1.

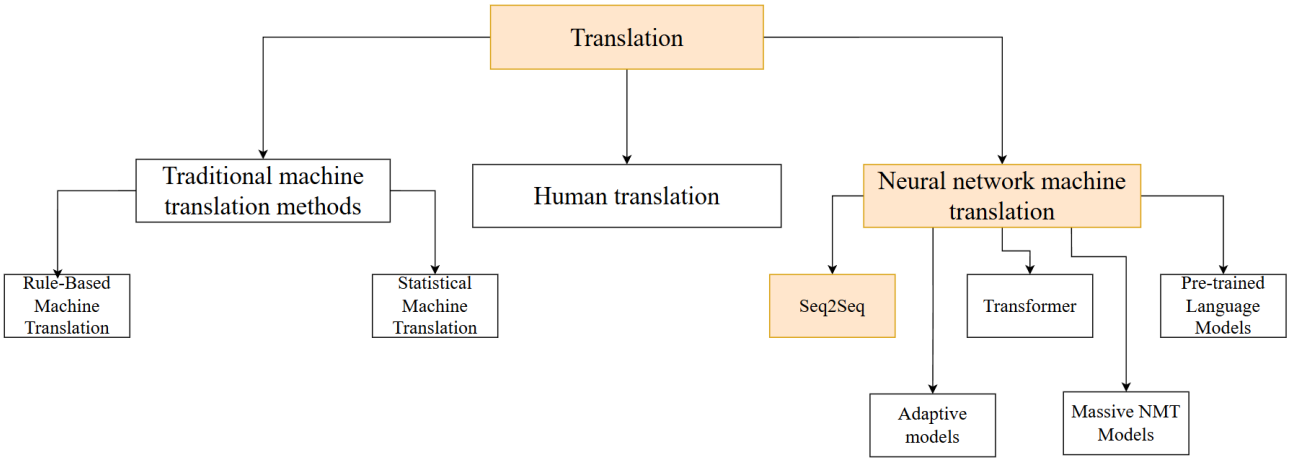

Рис. 1. Класифікація перекладів

Нейронні методи машинного перекладу стали головною технологією в автоматичному перекладі завдяки своїй здатності моделювати складні мовні залежності і створювати переклади, що наближені до людських. Їхня еволюція розпочалась із простих архітектур, які поступово ускладнювались і вдосконалювалися, що дало змогу досягати високої точності перекладу навіть між мовами з дуже різною структурою.

Перші моделі нейронного перекладу були базовані на принципі «послідовність у послідовність». У цьому підході використано дві рекурентні нейронні мережі — одна для кодування вхідного речення у вектор фіксованої довжини, а інша для декодування цього вектора в речення цільовою мовою. Під час навчання така модель поступово навчається асоціювати цілі послідовності слів між собою, не розбиваючи речення на окремі слова чи фрази, як це відбувалось у статистичних методах. Основною перевагою цього підходу була його здатність вивчати шаблони в мовленні без явного програмування правил. Проте моделі такого типу мали суттєве обмеження — вони погано працювали з довгими реченнями, оскільки весь сенс вхідного речення мусив вміститися в один фіксований вектор [8].

Щоб розв’язати проблему втрати інформації з кодуванням довгих речень, було запропоновано механізм уваги. Увага допомагає моделі фокусуватися не на всьому вхідному реченні одночасно, а вибірково звертати увагу на окремі його частини під час генерації кожного слова на виході. Інакше кажучи, декодер, створюючи переклад, отримує доступ до повного набору проміжних станів енкодера і динамічно зважує, які з них є найбільш релевантними для поточного моменту перекладу. Це дало змогу

значно покращити точність і узгодженість перекладу, зокрема в тих випадках, коли структура речення в одній мові значно відрізняється від іншої.

Найбільший прорив у нейронному перекладі стався з появою архітектури трансформерів. Цей підхід кардинально відмовився від рекурентних мереж і замість них використав механізм багатоголової самоуваги. У трансформерах модель може одночасно аналізувати всі слова вхідного речення і встановлювати між ними зв'язки незалежно від їхнього порядку. Завдяки цьому можна досягти набагато ефективнішого моделювання довготривалих залежностей у мові. Крім того, трансформери дозволяють навчання на значно більших обсягах даних і з використанням паралельної обробки, що зробило їх дуже ефективними на практиці. Саме трансформери стали основою більшості сучасних систем перекладу, таких як Google Translate, DeepL та інші [1, 2].

Наступним кроком у розвитку нейронного перекладу стало використання великих попередньо навчених мовних моделей. Ці моделі навчаються не лише на паралельних корпусах, а і величезних обсягах неструктурованого тексту різними мовами, що допомагає їм набувати глибшого розуміння мовної структури, граматики і стилю. Однією з таких моделей є mBART, яка поєднує можливості трансформера і додаткове переднавчання на завданні відновлення зашумленого тексту. У процесі перекладу така модель демонструє високу точність навіть у випадках малоресурсних мов, оскільки здатна узагальнювати знання, набуті з інших мов. Інший приклад — модель T5, яка трактує всі завдання, включаючи переклад, як задачу генерації тексту у відповідь на текстовий запит. Наприклад, щоб перекласти речення, їй достатньо дати інструкцію «переклади з англійської на українську», після чого вона сформує відповідний результат [9].

Окрему категорію складають масивні багатомовні моделі, які здатні перекладати між сотнями мов без необхідності мати паралельні дані для кожної мовної пари. Такі системи використовують узагальнене мовне представлення та принципи трансферного навчання. Наприклад, модель M2M-100 або NLLB (No Language Left Behind) розробляють з акцентом на підтримку рідковживаних мов і забезпечення якісного перекладу в умовах обмежених ресурсів. Вони навчаються на багатомовних корпусах, використовуючи принципи навчання без учителя, вирівнювання мовних просторів і багаторівневої самоуваги, що робить їх надзвичайно потужними та універсальними [10].

Значний інтерес також викликають моделі, які поєднують нейронний переклад із методами підкріплення або адаптації до зворотного зв'язку від користувачів. У таких системах модель може бути додатково навчена на основі людських оцінок перекладу або редагувань. Це дає змогу не лише підвищити якість, а й забезпечити відповідність стилю, термінології та специфічним вимогам користувача. Такі системи активно досліджують для впровадження в бізнес-інструменти, де переклад має відповідати корпоративним стандартам.

Отже, нейронні методи машинного перекладу пройшли довгий шлях — від простих рекурентних моделей до масштабованих трансформерних архітектур та універсальних багатомовних систем. Кожен новий етап розвитку наближає машинний переклад до рівня розуміння, властивого людині, забезпечуючи не лише переклад слів, а й передавання змісту, стилю та інтенції висловлювання.

Мета та задачі роботи

Метою роботи є розроблення інтелектуальної інформаційної системи перекладу штучних мов. Дослідницькою складовою в запропонованій системі є оцінювання ефективності нейромережевого перекладу і вимірювання часу роботи на стрес-тестах, що може призвести до скорочення часових і ресурсних затрат, а також забезпечити високу точність і надійність результатів.

Для досягнення поставленої мети мають бути вирішені такі завдання:

- порівняльний аналіз типів перекладачів;
- створення робочого датасету для подальшого дослідження;
- розроблення методики проведення експерименту;
- створення моделі інтелектуальної інформаційної системи перекладу штучних мов;
- оцінювання впливу нейромережевих моделей на результат перекладу штучних мов у межах запропонованої системи на основі методів машинного навчання.

Проведені експерименти є основою для подальшого покращення методів перекладу на основі нейромережевих моделей і частиною загального комплексу перекладу.

Обмеженнями системи є необхідність навчання нейромережі на паралельному корпусі текстів різними мовами. Також до обмежень можна віднести необхідність навчання системи на словах, що будуть використані для перекладу.

Отримані наукові результати. Обговорення результатів

У цьому досліджені вперше запропоновано інтелектуальну інформаційну систему перекладу штучних мов, яка генерує нові словники на основі існуючого набору паралельних текстів. Інформаційну систему наведено у вигляді узагальненої моделі (рис. 2) перекладача, яка складається зі взаємозалежних модулів: нейронного перекладу і статистичного перекладу

Рис. 2. Інтелектуальна інформаційна система перекладу штучних мов

Методика проведення експерименту, необхідного для досягнення поставленої мети, є такою: спершу було підготовлено робочий датасет через отримання набору паралельних речень англійською, українською, французькою мовами і штучною мовою есперанто. Кожен із датасетів містив однакові речення різними мовами.

Далі за допомогою цього датасету було створено нейронну модель для машинного перекладу, побудовану за принципом «послідовність у послідовність» (sequence-to-sequence, скорочено seq2seq). Основна її мета — навчитися перетворювати послідовність слів з однієї мови на логічно відповідну послідовність іншою мовою як імітацію перекладу за допомогою глибокої нейронної мережі.

На високому рівні модель складається з двох основних компонентів: енкодера (encoder) і декодера (decoder). Обидва реалізовані на основі багатошарового LSTM (Long Short-Term Memory) — різновиду рекурентної нейронної мережі, яка добре обробляє послідовності та зберігає інформацію в часі. LSTM створена для зберігання довготривалої залежності між словами, що особливо важливо для завдань перекладу, де слово на початку речення може впливати на переклад наприкінці. На відміну від аналогічних систем, ця система працює з набором токенів, записаних у словниках, що дає змогу використовувати систему на звичайних і штучних мовах.

На першому етапі енкодер приймає вхідне речення у вигляді індексів слів (які відповідають токенам із попередньо побудованого словника), перетворює їх у векторні представлення за допомогою вбудованого шару (embedding layer), а потім послідовно обробляє ці вектори через LSTM. На кожному кроці LSTM генерує прихований стан, але в результаті виводить лише фінальні приховані і клітинні стани, які зберігають стиснену інформацію про все вхідне речення. Саме ці фінальні стани передавані в декодер як початковий контекст для генерації перекладу.

Декодер на кожному кроці приймає попереднє слово вихідного речення (під час тренування це або правильне слово з навчального прикладу — teacher forcing, або попереднє передбачене), перетворює його на вектор, проводить через свою LSTM з оновленими станами і формує передбачення наступного слова. Це передбачення є вектором розміру зі словника вихідної мови, де кожен елемент вказує на ймовірність відповідного слова. Вибирає слово з найвищою ймовірністю, яке або знову подає на вхід, або порівнює з еталонним у випадку тренування.

Навчання складається з таких етапів:

- для зміни внутрішніх станів нейромережі використано метод зворотного поширення помилки (backpropagation through time), за якого оптимізатор Adam змінює ваги мережі, щоб зменшити функцію втрат;
- як функцію втрат використано кросентропійну втрату (CrossEntropyLoss), яка прийнятна для завдань класифікації, зокрема вибору одного правильного слова з великого словника;
- нейромережа навчається 100 епох, на кожній із яких аналізують ефективність навчання.

Для оцінювання нейромережі була розроблена також оцінка точності (accuracy), яку розраховують як відсоток правильних передбачень серед усіх токенів у реченнях.

Щоб модель працювала, перед навчанням створюють словники для обох мов: кожному слову відповідає індекс, а кожне речення перетворюється на фіксованої довжини послідовність цих індексів. Дані беруть із двох вхідних текстових файлів, наприклад англійський текст у першому файлі та його український переклад в іншому [11, 12]. Оскільки для перекладу окремого слова потрібно на ньому навчити нейронну модель, то поділ датасету на навчальний і тестовий неможливий. Кожне тестоване слово має пройти навчання. Алгоритм описаної моделі зображено на рис. 3.

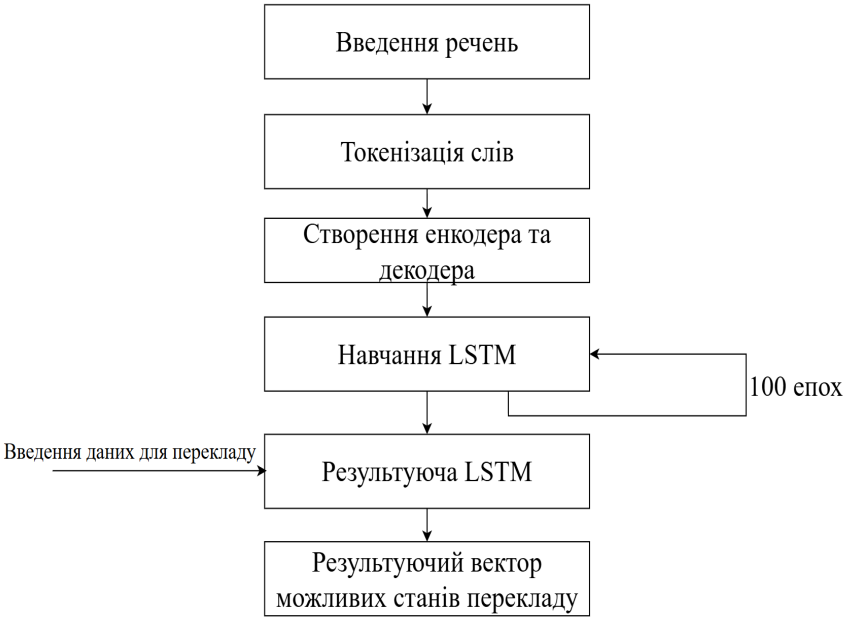

Рис. 3. Алгоритм використання LSTM нейромережі для перекладу

Для тестування моделі було проведено набір експериментів. У кожному з них брали спеціально вибраний датасет, що складається з набору паралельних речень двома вибраними мовами, після чого на цьому наборі даних проходило навчання нейромережі та перевірка якості перекладу. Для перевірки якості кожен набір словосполучень із початкової мови перекладали на цільову мову за допомогою нейромережі. У результаті нейромережа генерувала вектор, який оцінює можливі переклади. Після цього вектор значень аналізували і порівнювали з правильною відповіддю. Загальний результат – це відсоток розміру вхідного словника слів, для яких значення порогу правильно визначених завчасно вибраних відповідей було найбільшим серед усіх варіантів.

Для повного тестування проводили три різних набори експериментів:

1. Тестування на невеликій (40 речень) контрольованій вибірці підібраних речень англійською, українською, французькою мовами і штучною мовою есперанто.

2. Тестування на більшій вибірці (40 контрольованих речень і 40 випадкових) із додаванням до першої вибірки випадкових слів і речень.

3. Тестування на великій вибірці (стрес-тести до 1100 речень у наборі) для тестування максимального часу роботи системи.

У першому наборі експериментів було проведено 12 тестувань і отримано результати, наведені в табл. 1.

Таблиця 1

Точність перекладу першого набору

	Англійська	Українська	Французька	Есперанто
Англійська	-	93.4 % (1)	96.1 % (3)	97.2 % (5)
Українська	94.7 % (2)	-	94.1 % (7)	94.6 % (9)
Французька	95.3 % (4)	93.9 % (8)	-	97.5 % (11)
Есперанто	96.8 % (6)	96.3 % (10)	96.8 % (12)	-

У таблиці подано результати експериментів. Рядок означає вихідну мову, стовпчик – цільову. У дужках описано номер експерименту, який зображено на рис. 4.

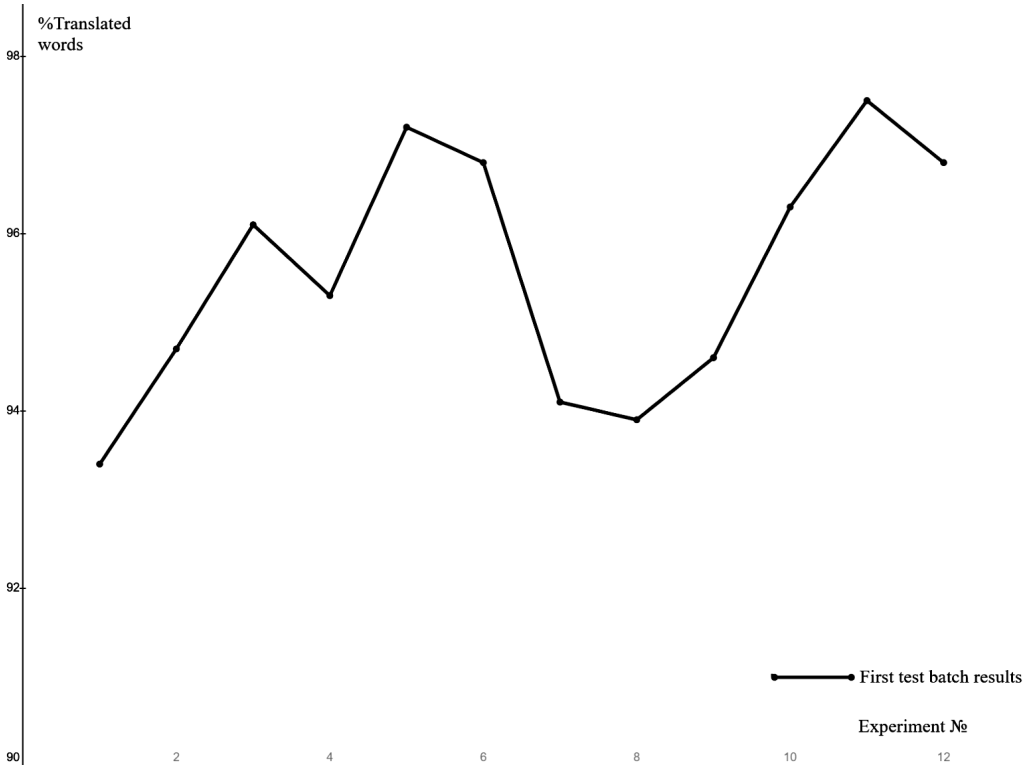

Рис. 4. Результат першого набору експериментів

У другому наборі експериментів було проведено 12 тестувань із розширеним датасетом і отримано результати, наведені в табл. 2.

Точність перекладу першого набору експериментів та другого набору

	Англійська	Українська	Французька	Есперанто
Англійська	-	89.7%(1)	92.2%(3)	92.8%(5)
Українська	90.3%(2)	-	89.6%(7)	90.5%(9)
Французька	91.1%(4)	89.2%(8)	-	92.9%(11)
Есперанто	92.7%(6)	92.1%(10)	93.0%(12)	-

Результати таблиці описано в тому самому форматі, що і табл. 1. Порівняння результатів двох наборів експериментів подано на рис. 5.

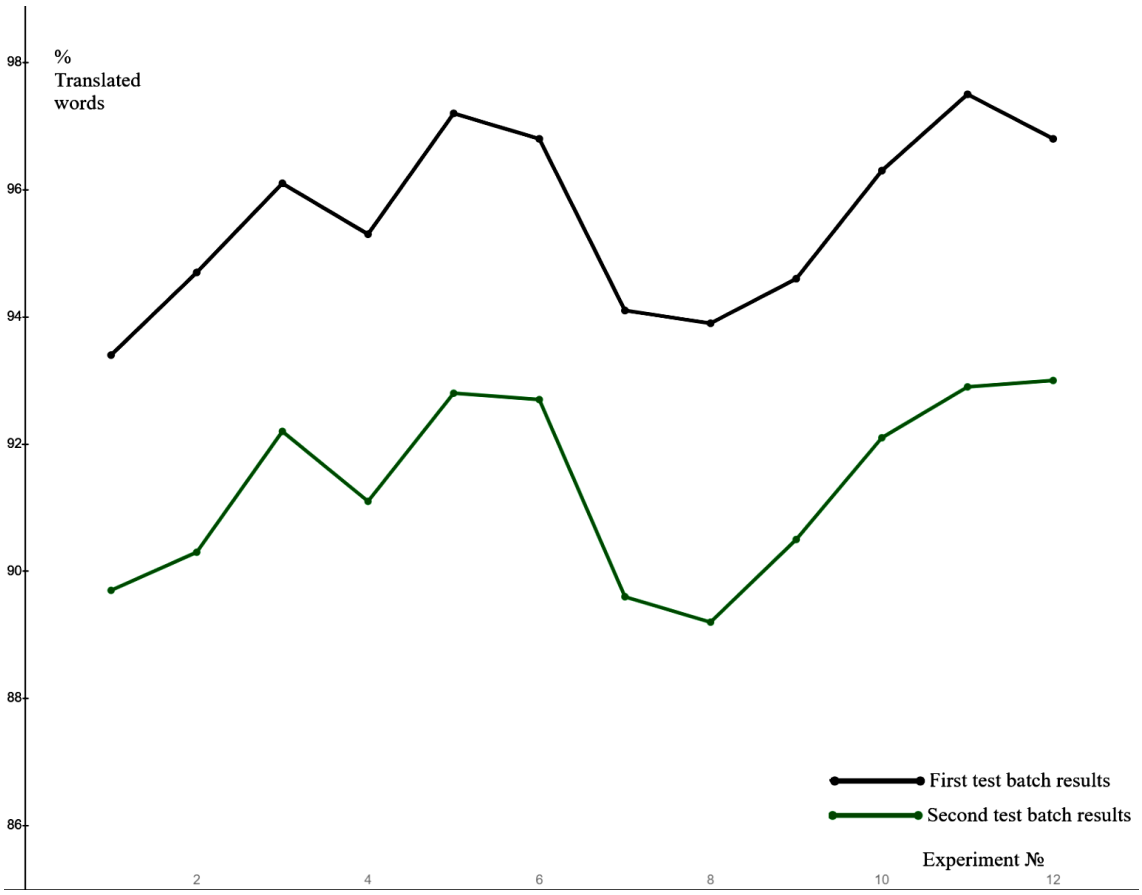

Рис. 5. Порівняльний результат першого та другого набору експериментів

У результаті ми можемо бачити, що якість перекладу досить висока, але нижча, ніж за ідеального підбору речень. Оскільки слова перетворені в числа в словнику, на результат менше впливають окремі слова, а більше впливає структура мови. І з правильним підбором речень результат роботи може бути кращим.

У третьому наборі експериментів було визначено вплив розміру датасету на час навчання однієї епохи нейромережі. Експеримент проводили на обчислювачі 13th Gen Intel(R) Core(TM) i7-13620H 2.40 GHz. Для проведення експериментів було вибрано штучну мову есперанто та українську мову. Отримані результати подані в табл. 3.

Час роботи третього набору експериментів

Експеримент	Кількість паралельних речень	Середній час навчання однієї епохи
1	40	2 с
2	400	15 с
3	1100	75 с

Результати у графічному вигляді подані на рис. 6.

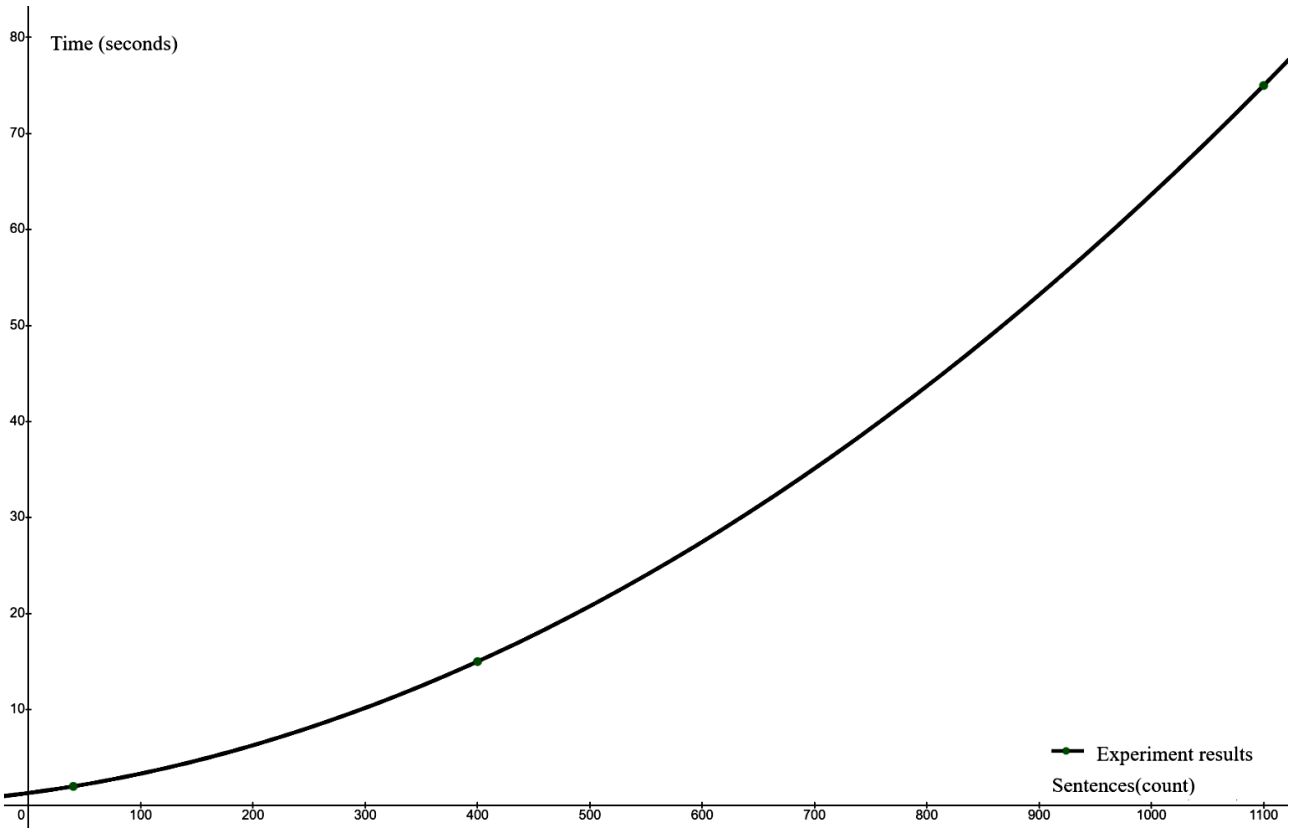

Рис. 6. Залежність часу навчання однієї епохи нейромережі від кількості вхідних речень

Як можемо бачити, зі зростанням розміру датасету час збільшується майже з квадратичною складністю. Це можна пояснити тим, що передбачення є вектором розміру зі словник вихідної мови. І зі збільшенням розміру датасету збільшується складність внутрішньої структури нейромережі. Тому цей метод не можна використовувати як основний метод перекладу, але він може бути окремим модулем для покращення результатів.

Висновки

У дослідженні вперше було розроблено інтелектуальну інформаційну систему перекладу штучних мов. У рамках дослідження вдалося досягти встановлених цілей і вирішити поставлені завдання.

Провели детальне дослідження та порівняли перекладачі. Це дало змогу визначити, які техніки перекладу найкращі для штучних мов.

Було проведено дослідження одного з модулів системи, а саме нейромережевий переклад.

У результаті ми отримали те, що нейромережевий переклад працює на всіх типах вибраних мов і показує результати до 97.5 % точності перекладу для спеціально вибраного датасету.

Було проаналізовано максимальне навантаження на нейромережу та з’ясовано, що за великих обсягів даних час навчання однієї епохи може сягати декількох хвилин.

Ураховуючи отримані результати, можна зробити висновок, що цей метод не можна використовувати як основний метод перекладу системи для штучних мов, але він може бути окремим модулем для покращення результатів.

Наукове значення роботи полягає в поглибленні розуміння механізмів, що лежать в основі нейромерж. Викладені висновки можуть бути основою для подальших досліджень у галузі машинного навчання та розроблення інтелектуальних систем, що сприятиме прогресу у сферах штучного інтелекту і машинного перекладу

Практична цінність цього дослідження полягає у створенні інтелектуальної інформаційної системи перекладу, що не залежить від початкових мов. Це дасть змогу використовувати автоматичний переклад для звичайних, штучних і сленгових мов без додаткових затрат.

Список використаних джерел

1. Wu Yonghui et al. Google's neural machine translation system: Bridging the gap

between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
2. Kamaluddin Mohamad Ihsan et al. Accuracy analysis of DeepL: Breakthroughs in machine translation technology. *Journal of English Education Forum (JEEF*). 2024. Vol. 4, No. 2.
3. Jiang Zhaokun et al. Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation. *arXiv preprint arXiv:2401.05176* (2024).
4. Lakew Surafel M., Mauro Cettolo and Marcello Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *arXiv preprint arXiv:1806.06957* (2018).
5. Sennrich Rico and Barry Haddow. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892* (2016).
6. He Wei et al. Improved neural machine translation with SMT features. *Proceedings of the AAAI conference on artificial intelligence*. 2016. Vol. 30, No. 1.
7. Peris Álvaro, Miguel Domingo and Francisco Casacuberta. Interactive neural machine translation. *Computer Speech & Language* 45 (2017): 201-220.
8. Gong Gangjun et al. Research on short-term load prediction based on Seq2seq model. *Energies* 12.16 (2019): 3199.
9. Chipman Hugh A. et al. mBART: multidimensional monotone BART. *Bayesian Analysis* 17.2 (2022): 515-544.
10. Costa-Jussà Marta R. et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).
11. Havrashenko Anton and Olesia Barkovska. Analysis of text augmentation algorithms in artificial language machine translation systems. *Advanced Information Systems* 7.1 (2023): 47-53.
12. Барковська О., Гаврашенко А., Холєв В., Севостьянова О. (2021). Automatic text translation system for artificial llanguages. *Computer systems and information technologies*. (3). 21-30.

Intelligent information system for translation of artificial languages

Abstract. *The article is devoted to the development of an intelligent information system for translation of artificial languages using artificial intelligence methods and statistical algorithms, proposing an innovative approach to increasing the efficiency of machine translation in the conditions of dynamic development of modern technologies. The aim of the work is to develop an intelligent information system for translation of artificial languages. The research component in the proposed system is an assessment of the effectiveness of using artificial intelligence methods for the translation of artificial and ordinary languages, which can lead to a reduction in time and resource costs, as well as ensure high accuracy and reliability of the results. The work investigates the use of long short-term memory (LSTM) for translation. The results demonstrate that LSTM achieves accuracy up to 97.5 % with controlled sampling of input data, and up to 93 % on a random dataset. The results obtained confirm the effectiveness of the developed system, which allows for highly efficient translation of ordinary and artificial languages.*

***Keywords**: machine learning, artificial intelligence, LSTM, datasets, encoder, decoder, sequence-to-sequence.*

About the authors

Гаврашенко Антон Олегович, аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна.
E-mail: anton.havrashenko@nure.ua. ORCID ID https://orcid.org/0000-0002-8802-0529.
Anton Havrashenko, PhD student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
E-mail: anton.havrashenko@nure.ua. ORCID ID https://orcid.org/0000-0002-8802-0529.

Барковська Олеся Юріївна, кандидат технічних наук, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна. E-mail: olesia.barkovska@nure.ua. ORCID ID http://orcid.org/0000-0001-7496-4353.
Olesia Barkovska, Candidate of Technical Sciences, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
E-mail: olesia.barkovska@nure.ua. ORCID ID http://orcid.org/0000-0001-7496-4353.