https://doi.org/10.31649/1997-9266-2024-177-6-113-120

УДК 004.8

А. А. Яровий[1]
Д. С. Кудрявцев[1]

ПІДХІД ДО ГЕНЕРАЦІЇ ТЕКСТУ НА ОСНОВІ МОВНОЇ МОДЕЛІ BERT

[1]Вінницький національний технічний університет

Запропоновано застосування мовної моделі BERT для задач пошуку і генерації термів у термінологічних базах знань (ТБЗ) з використанням оптимізації для інтелектуальних чат-ботів. Описується архітектура моделі BERT, механізм уваги, алгоритми обробки тексту та основні етапи навчання моделі. Розглянуто використання BERT для семантичного пошуку термів, а також методи адаптації моделі для генерації тексту з урахуванням семантичної цінності кожного терму. Виконано порівняльний аналіз мовної моделі BERT з моделями серії GPT, який продемонстрував сильні та слабкі сторони BERT у контексті пошукових і генеративних задач. У статті також детально розглянуто метрики оцінки якості пошуку термів, такі як Precision, Recall, F1-score, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG) та інші, що дозволяють комплексно оцінювати ефективність пошуку та генерації термів. Розглянуто практичні аспекти інтеграції BERT у системи управління знаннями та надано рекомендації щодо донавчання моделі для вузькоспеціалізованих ТБЗ. До того ж, зосереджено увагу на етичні аспекти використання мовних моделей, зокрема ризики виникнення упередженості (bias) у пошуку та генерації термів, а також важливість забезпечення точності й об'єктивності згенерованих результатів. Обговорюється відповідальне використання BERT для уникнення помилкових або шкідливих висновків під час автоматичної обробки знань. Здійснено розробку програмного забезпечення для тестування мовної моделі BERT. Виконано тестування навчання мовної моделі на різних наборах даних. Результатом тестування доведено високу ефективність використання мовної моделі BERT з урахуванням оптимізації для задачі генерації тексту. Зазначено можливі покращення BERT для роботи з ТБЗ, зокрема методи донавчання моделі на специфічних доменних даних, використання мультимовної версії BERT для обробки багатомовних баз знань, а також техніки оптимізації моделі для підвищення продуктивності в умовах обмежених обчислювальних ресурсів. Запропоновано підходи до тестування та оцінки ефективності пошуку, зокрема використання автоматичних метрик. У заключній частині статті окреслено подальші напрямки досліджень, зокрема інтеграцію BERT з нейронними пошуковими системами, автоматичну генерацію нових термів та розширення функціоналу систем управління знаннями на основі глибокого навчання.

Ключові слова: BERT, термінологічні бази знань, семантичний пошук, мовні моделі, генерація термів.

Вступ

Обробка природної мови (NLP) є однією з ключових галузей штучного інтелекту, що швидко розвивається і охоплює завдання аналізу, розуміння та генерації тексту. Зі зростанням обсягів інформації та необхідністю її швидкої обробки в різних доменах знань постала потреба в створенні інструментів для автоматизованого пошуку, класифікації та генерації тексту. У цьому контексті мовні моделі, такі як BERT (Bidirectional Encoder Representations from Transformers), стали революційним рішенням, забезпечуючи значний прогрес у точності розуміння тексту та ефективності генеративних задач. Модель BERT, розроблена Google у 2018 році, впровадила нові підходи до обробки текстових даних, використовуючи механізм уваги, що дозволяє одночасно враховувати контекст як зліва, так і справа від кожного слова. Це зробило BERT досить ефективною для широкого спектру NLP-задач [1].

Однією з важливих сфер застосування BERT є термінологічні бази знань (ТБЗ), що широко використовуються в різних галузях науки та техніки. У таких базах знань містяться терми, що відображають ключові концепти певного домену, а завданням автоматизованих систем є забезпечення пошуку, класифікації та генерації цих термів. Традиційні методи обробки тексту, такі як вектори-

зація термів за допомогою TF-IDF або використання Word2Vec, часто не можуть ефективно враховувати складні семантичні зв'язки між термами, що обмежує якість пошуку та обробки інформації в термінологічних базах [2]. У цьому контексті модель BERT відкриває нові можливості завдяки своїй здатності до тонкого аналізу контексту та семантики.

Структура моделі BERT, побудована на архітектурі трансформера, що дозволяє їй ефективно обробляти текст завдяки механізму уваги (self-attention). Цей механізм оцінює, які слова в реченні найважливіші для розуміння значення кожного окремого токена, дозволяючи моделі краще визначати взаємозв'язки між термами в тексті. До того ж, існують методи адаптації мовної моделі BERT з використанням маскованого передбачення токенів (Masked Language Model), коли модель поступово заповнює пропущені слова, що дозволяє генерувати нові терми в ТБЗ на основі контексту [3].

Серед відомих натепер мовних моделей, варто порівняти можливості моделі BERT з іншими сучасними моделями, зокрема GPT (Generative Pre-trained Transformer), які більше орієнтовані на генерацію тексту [4]. Порівняльна характеристика з метою оцінки сильних і слабких сторін кожної моделі у контексті пошуку термів та їхньої генерації необхідна для вибору оптимальної та її подальшої оптимізації [4]. Хоча GPT добре підходить для задач покрокової генерації тексту, BERT демонструє вищу точність у задачах розуміння тексту та його класифікації, що є важливим для задач обробки термів у ТБЗ. Оскільки визначення контексту базується на його спорідненості з вже відомими даними з ТБЗ, ключова характеристика, що матиме першочергове значення — це спорідненість термів між собою, що в подальшому може бути використано для формування результату на основі тієї ж спорідненості. Для визначення споріднених термів, необхідно використовувати метрики оцінки якості пошуку термів для визначення ефективності пошукових алгоритмів. Серед них — Precision, Recall, F1-score, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG) [5]. Ці метрики дозволяють оцінити, наскільки ефективно модель знаходить релевантні терми та як добре вона ранжує результати пошуку. Для генерації термів важливим є також врахування семантичної релевантності, що значною мірою забезпечується за допомогою механізму уваги в BERT. Оскільки більшість з відомих мовних моделей все ж таки належать до сімейства GPT, використання мовної моделі BERT для задачі генерації тексту є недостатньо розкритим та потребує детальнішого аналізу завдяки потенційним можливостям поліпшення аналізу контексту.

Метою статті є розробка підходу до генерації тексту на основі мовної моделі BERT з використанням термінологічних баз знань як джерела даних. Одним з основних завдань є адаптація BERT для семантичного пошуку термів, що забезпечує точніше і релевантне знаходження термів для генерації тексту, порівняно з традиційними підходами. Особливу увагу приділено використанню мовної моделі BERT для генерації тексту, зокрема, генерації термів з урахуванням їхньої семантичної цінності в контексті.

Для реалізації поставленої мети необхідно виконати такі *задачі*:

– визначити архітектурні компоненти мовної моделі BERT та її аналогів;

– сформувати критерії ефективності для визначення спорідненості термів;

– скомпонувати процес визначення споріднених термів в ТБЗ та генерації тексту на основі визначених термів;

– визначити можливості для оптимізації на основі порівняльної характеристики з GPT моделями;

– виконати комп'ютерне моделювання для реалізації задачі генерації тексту за використання мовної моделі BERT.

В ході аналізу мовних моделей досить актуальною зазначається проблема з упередженням (bias), що може впливати на результати пошуку або генерації термів, особливо в спеціалізованих базах знань [4]. Одною з можливостей вирішення цієї порблеми є примусова фільтрація. Це рішення випливає з факту, що актуальність даних поступово зменшується, а їхня кількість лише зростає. Таким чином, фільтрація матиме на меті зменшити кількість даних для початкового аналізу контексту, без урахування спорідненості термів. Завдяки цьому пошук споріднених термів на першому етапі комп'ютерного моделювання подано у вигляді задачі оптимізації, де оптимальний набір термів для другого етапу, а саме безпосередньої генерації тексту з використанням багатоцільового механізму уваги, визначатиметься за рахунок відношення середнього арифметичного коефіцієнта спорідненості термів з контексту та релевантних термів з ТБЗ до кількості термів, що будуть використані для генерації тексту.

Ще одним важливим аспектом, що пропонується розглянути в цьому дослідженні, є можливі покращення BERT для роботи з ТБЗ, а саме вплив структури даних ТБЗ, взаємозв'язки між термами та групами за семантичною ознакою. Оскільки задача генерації тексту є досить складною з

погляду навчання мовної моделі та потребує великих корпусів текстових даних, ефективне рішення для зменшення часу навчання моделі в цьому випадку є технологія відкладеного навчання. Ідея цієї технології полягає у розділенні процесу навчання мовної моделі на етапи, які визначаються за глибиною та обсягом даних, що потребує модель. Таким чином, мовна модель може бути навчена поступово на домен-специфічних даних, що дозволить їй краще розуміти терми певної галузі. При цьому для подальшого використання в інших предметних сферах необхідне буде лише повернення до попереднього етапу та повторне навчання на даних з нових предметних сфер. До того ж, застосування багатомовної версії BERT (mBERT) відкриває можливості для оброблення багатомовних термінологічних баз знань, що є особливо важливим для глобальних наукових та технічних баз [4].

Основна частина

Основна мета BERT полягає в глибшому розумінні контексту тексту через попереднє навчання на великих наборах неанотованих даних. Це дозволяє мовній моделі вивчати контекстну інформацію і будувати зв'язки між словами, термами і фразами, поліпшуючи результати в різних завданнях NLP, включно з генерацією тексту. BERT базується на архітектурі трансформера (Transformer) [6]—[8]. Приклад структури цієї моделі показано на рис. 1. Трансформер складається з двох основних блоків — енкодера та декодера, але у випадку BERT використовується лише енкодерна частина. Енкодер має багатошарову структуру, яка дозволяє моделі ефективно обробляти вхідні текстові дані. Основними компонентами архітектури BERT є:

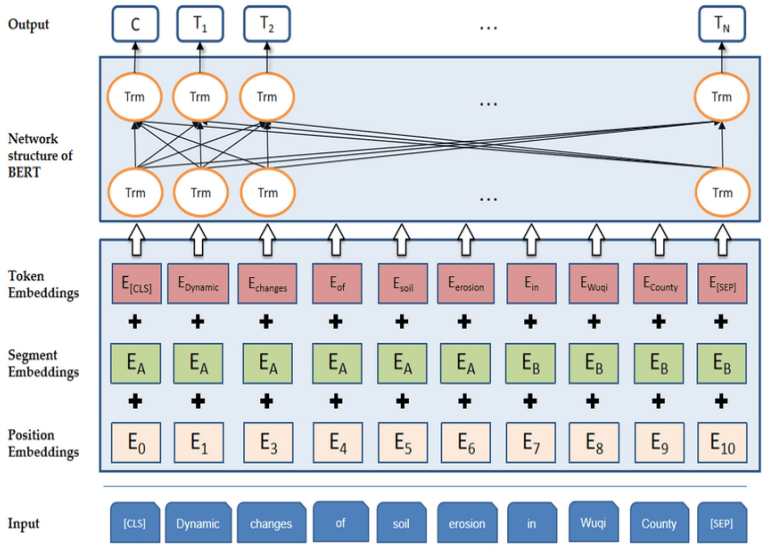

Рис. 1. Структура мовної моделі BERT [6]

1. Механізм уваги (Self-Attention): BERT використовує цей механізм для розуміння залежностей між усіма словами у вхідній послідовності. Це означає, що кожне слово у тексті взаємодіє з усіма іншими словами, що дозволяє моделі захоплювати як локальні, так і глобальні зв'язки.

2. Багатоцільова увага (Multi-head Attention): Кожна «ціль» в даному випадку вивчає різні аспекти взаємозв'язку між словами, що допомагає захоплювати різні аспекти контексту.

3. Позиційне кодування (Positional Encoding): Оскільки BERT не має рекурентної структури, він використовує позиційне кодування для збереження інформації про порядок слів у послідовності.

4. Шари прямого розповсюдження (Feedforward layers): Ці шари допомагають моделі вивчати складні перетворення та взаємодії між словами.

5. Нормалізація шарів та залишкові зв'язки (Layer Normalization and Residual Connections): Ці техніки стабілізують процес навчання і дозволяють ефективно навчати дуже глибокі моделі.

Однією з ключових особливостей BERT є двонаправлене навчання. На відміну від інших мовних моделей, які обробляють текст або зліва-направо, або справа-наліво, BERT одночасно враховує контекст обох напрямків [9]. Це дозволяє моделі краще розуміти зв'язки між словами, які зустрічаються як до, так і після поточного слова. Навчання BERT відбувається у два етапи. Попереднє навчання (Pre-training), що включає тренування моделі на великих масивах неанотованих текстів за допомогою двох основних завдань: маскованого мовного моделювання (Masked Language Model, MLM), що полягає в заміні частини слів у тексті випадково замінюється на масковане значення і модель повинна вгадати, які саме слова були замасковані. Це дозволяє BERT захоплювати контекст з обох сторін маскованого слова. А друге полягає в прогнозуванні наступного речення (Next Sentence Prediction, NSP), де модель отримує пару речень і повинна визначити, чи є друге речення логічним продовженням першого. Це допомагає BERT вивчати зв'язки між реченнями.

Другим етапом є специфічне налаштування (Fine-tuning) під вибрану задачу [9]. На цьому етапі,

у дослідженні такою задачею є генерація тексту. Після попереднього навчання модель BERT може бути адаптована до конкретних завдань шляхом специфічного налаштування. Це означає, що модель навчається на менших і специфічніших наборах даних для виконання конкретного завдання, такого як класифікація тексту, відповіді на питання або генерація тексту. Хоча основною метою BERT є розуміння тексту, його також можна адаптувати для завдань генерації тексту. Проте для цього потрібно модифікувати модель або застосовувати специфічні стратегії. Одними з найдієвіших є використання адаптації BERT-2-BERT [10]. Оскільки BERT сам по собі є моделлю для оброблення тексту, а не для його генерації, його можна поєднувати з іншими моделями або використовувати у власній адаптації. Один з підходів — використання двох моделей BERT: одна для розуміння вхідного тексту, а інша — для генерування відповіді. Це дає змогу використовувати можливості BERT для аналізу та розуміння контексту під час генерації тексту. В ході цього дослідження така модифікація є найкращою з погляду вибірки даних та архітектурного підходу до практичної реалізації.

Після визначення архітектурних компонентів мовної моделі BERT, необхідно охарактеризувати спорідненість термів в ТБЗ, а саме визначення залежності між кількістю зв'язків термів між собою та загальною частотою зв'язків для терму в ТБЗ. Для комп'ютерного моделювання цього процесу використано середовище розробки PyCharm та графову базу даних neo4j, що містить значну кількість даних для підтвердження ефективності введення критерію спорідненості термів.

```python
from neo4j import GraphDatabase
import networkx as nx

class TermAffinityAnalyzer:
    def __init__(self, uri, user, password):
        self.driver = GraphDatabase.driver(uri, auth=(user, password))

    def close(self):
        self.driver.close()

    def fetch_graph_data(self):
        query = """
        MATCH (n)-[r]->(m)
        RETURN n.name AS source, m.name AS target
        """
        with self.driver.session() as session:
            result = session.run(query)
            return [(record["source"], record["target"]) for record in result]

    def calculate_affinity(self, graph_data):
        G = nx.Graph()
        G.add_edges_from(graph_data)

        total_edges = G.number_of_edges()

        from networkx.algorithms.community import greedy_modularity_communities
        clusters = list(greedy_modularity_communities(G))

        num_clusters = len(clusters)

        if num_clusters > 0:
            average_affinity = total_edges / num_clusters
        else:
            average_affinity = 0

        return {
            "total_edges": total_edges,
            "num_clusters": num_clusters,
            "average_affinity": average_affinity
        }
```

Рис. 2. Фрагмент коду для визначення спорідненості термів в ТБЗ

Для програмної реалізації використано мову програмування Python та бібліотеки neo4j. Фрагмент програмної реалізації визначення спорідненості термів в ТБЗ показано на рис. 2, а результати комп'ютерного моделювання спорідненості подано в табл. 1.

З погляду використання як джерела даних для навчання ТБЗ, задача генерації поділяється на дві складових, а саме визначення набору термів ТБЗ, що належать контексту для генерації та, власне, генерації на основі контексту. Структура ТБЗ переважно складається з графової бази даних, де вузлами виступають терми, а ребра — зв'язками. Зв'язки в цьому випадку будуть використані для визначення семантичної спорідненості та належності до ТБЗ, а також з урахуванням зв'язку між даними для генерації можливих протирічь, що може виникати внаслідок некоректних даних, чи часткову втрату семантичних зв'язків [11]. Основними рішеннями в цьому випадку є фільтрація семантичних зв'язків на предмет цілісності та обмеження на розмір семантичних груп, що поділяються за належністю то ТБЗ чи за певною семантичною ознакою. Для формування контексту або векторного представлення набору термів необхідно визначити розмір такого вектора, коефіцієнти семантичної спорідненості між термами, що входять до вектора та представлені у вигляді матриці спорідненості.

Для поліпшення та оптимізації роботи мовної моделі з ТБЗ, необхідно розглянути можливість використання механізму багатоцільової уваги.

Таблиця 1

Комп’ютерне моделювання спорідненості термів в ТБЗ

Кількість термів	Кількість предметних областей	Кількість зв’язків між термами	Спорідненість
1000	23	10602	460,9565
2000	30	23837	794,5667
3000	38	36998	973,6316
4000	46	50116	1089,478
5000	53	64120	1209,811
6000	61	77360	1268,197
7000	69	90499	1311,58
8000	76	104364	1373,211
9000	84	117751	1401,798
10000	92	130444	1417,87

Основний зміст механізму для цього дослідження полягає у визначенні спорідненості та зв’язку між термами в ТБЗ на основі різних комбінацій або семантичних груп за різними критеріями, такими як семантична складова, мовна група чи подвійний контекст.

В ході комп’ютерного моделювання зазначається зростання спорідненості термів зі збільшенням кількості термів в ТБЗ, що показано на рис. 3. Насамперед це пов’язано з предметними областями, що виступають в ролі класів, а при визначенні спорідненості, маркери класів невідомі в тренувальній вибірці, завдяки чому, отримані групи термів формують кластери, що фактично відповідають предметним сферам. На противагу BERT, розглянемо мовну модель GPT. Ця модель вже містить чимало підвидів, що формують групу мовних моделей, які базуються на архітектурі трансформерів та містять значну кількість оптимізацій під специфічні задачі. Однією з найсучасніших мовних моделей цієї групи є GPT 4 та GPT-4Turbo [12]. Специфіка та відмінність від попередніх мовних моделей цієї групи полягає у підвищенні продуктивності, здатності до контекстуалізації та адаптивності за рахунок збільшення кількості параметрів, врахування багатоетапності задач та врахування стилю мови, тональності, що допомагає краще адаптуватись до користувача та покращити якість взаємодії. В цьому дослідженні повноцінне навчання мовних моделей не є доцільним за рахунок обмеженості тренувальних вибірок, тому основну увагу зосереджено на тестуванні оптимізації у вигляді примусової фільтрації термів на першому етапі знаходження релевантних термів з ТБЗ, а також тестування механізму багатоцільової уваги з використанням декількох вхідних точок. Для забезпечення прийнятного набору даних для навчання використано наявний датасет з понад 10000 термів [13].

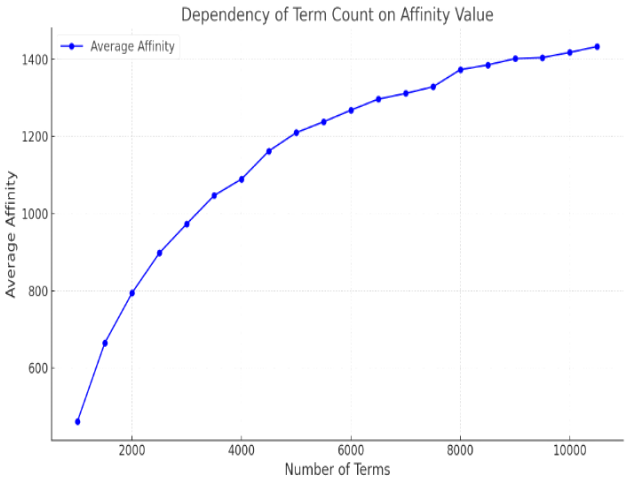

Рис. 3. Графік залежності спорідненості термів від кількості термів в ТБЗ

Для оцінювання мовної моделі BERT за використання оптимізації та GPT, необхідно застосувати метрики оцінювання, що визначатимуть результативність, а саме метрики на основі збігів та на основі семантики. Причому використовуючи BERT-2-BERT, необхідно також використати метрики Precision, Recall, F1-score для визначення точності мовної моделі на предмет знаходження релевантних термів. Враховуючи контекст програмної реалізації методу генерації тексту на основі набору термів з ТБЗ, необхідно описати процес генерації тексту на прикладі реального застосування, а саме інтелектуального чат-бота з можливістю діалогу з користувачем. В цьому випадку процес генерації тексту базується на отриманні ключових термів з повідомлення користувача, знаходження релевантних термів з ТБЗ, що є основним джерелом даних як для пошуку термів, що відповідають контексту вхідного повідомлення користувача, так і джерелом для генерації відповіді користувачу.

Практична реалізація здійснена мовою програмування Python з використанням платформи Open AI [14] в середовищі розробки PyCharm. Результати тестування подані в табл. 2 та 3.

Таблиця 2

Результати визначення контексту за метриками Precision, Recall, F1-Score

Метрика / Мовна модель	Precision			Recall			F1-score		
	ТБЗ 1	ТБЗ 2	ТБЗ 1&2	ТБЗ 1	ТБЗ 2	ТБЗ 1&2	ТБЗ 1	ТБЗ 2	ТБЗ 1&2
BERT	0,656	0,724	0,702	0,753	0,826	0,864	0,701	0,772	0,774
BERT*	0,755	**0,783**	0,765	0,828	0,878	0,875	0,789	0,827	0,816
BERT**	**0,841**	0,745	**0,863**	**0,914**	**0,942**	**0,963**	**0,876**	**0,832**	**0,911**
GPT	0,819	0,709	0,829	0,903	0,938	0,942	0,859	0,808	0,881
Word2Vec	0,538	0,673	0,697	0,521	0,504	0,587	0,529	0,576	0,637

Примітки: 1. BERT* — це реалізація мовної моделі BERT з використанням підходу BERT-2-BERT та механізму уваги з однією точкою входу; 2. BERT** — з двома цілями механізму уваги.

Таблиця 3

Результати генерації тексту за метриками MRR, nDCG

Метрика / Мовна модель	MRR			nDCG		
	ТБЗ 1	ТБЗ 2	ТБЗ 1&2	ТБЗ 1	ТБЗ 2	ТБЗ 1&2
BERT	0,455	0,449	0,394	0,624	0,825	0,826
BERT*	0,487	**0,483**	0,426	0,655	0,842	0,829
BERT**	0,499	0,473	0,451	**0,683**	**0,879**	**0,859**
GPT	**0,502**	0,47	**0,458**	0,651	0,872	0,841

Також розглянемо метрики Mean Reciprocal Rank (MRR), яка визначає середнє значення зворотного рангу релевантного терму в векторному представленні та Normalized Discounted Cumulative Gain (nDCG), що визначає вагу терму у відповідності до позиції у векторному представленні.

Для нормалізації використовується IDCG, коли найближчі сусідні терми за семантичною ознакою знаходяться на перших позиціях в згенерованому тексті.

З цих результатів варто оцінити різницю між мовними моделями GPT та BERT з використанням підходу BERT-2-BERT та механізмом багатоцільової уваги, що дали найвищі показники, але різниця незначна, що пояснюється схожістю їхньої архітектури та принципу роботи. Але незважаючи на це, вплив механізму багатоцільової уваги, що є ключовою відмінністю моделей типу BERT від моделей GPT, надав кращі результати у разі збільшення кількості цілей «уваги».

Відповідно до отриманих результатів, варто зазначити доцільність використання підходу BERT-2-BERT, а також використання ТБЗ та графової структури даних як оптимальної для знаходження релевантних термів за рахунок семантичних зв’язків між термами. За метрикою F1-score результати майже ідентичні та кращі від GPT в межах 1,983…3,217 % за рахунок обмеженості вибірки даних. Але, беручи до уваги основну задачу, а саме генерацію тексту, результативність використання підходу BERT-2-BERT є нижчою за GPT на 0,598…1,529 % за метрикою MRR, але вищою за GPT на 0,803…4,916 % за метрикою nDCG. Таким чином, застосування механізму багатоцільової уваги та підходу BERT-2-BERT є ліпшим для задачі генерації тексту з використанням ТБЗ, але з невеликою перевагою у порівнянні з GPT.

Висновки

За результатами дослідження можна дійти висновку про високу ефективність застосування мовної моделі BERT з використанням оптимізації для задачі генерації тексту. Для порівняння застосовано мовну модель GPT, яка є однієї з найкращих в сфері задач генерації тексту, а результат порівняння лише підтвердив це твердження за рахунок оптимізації в архітектурі мовної моделі. На противагу використанню термінологічних баз знань як джерела даних, показники точності визначення контексту для подальшої генерації, все ж таки менш ефективні за використання підходу BERT-2-BERT. Таким чином, планується проведення подальших досліджень та розробки повноцінної мовної моделі на базі інформаційної технології створення інтелектуальних чат-ботів з використанням мовної моделі BERT та підходу BERT-2-BERT для задачі генерації тексту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” *J. Big Data*, no. 9, 15, 2022. https://doi.org/10.1186/s40537-022-00564-9 .

[2] W. Liu, et al., “K-BERT: Enabling Language Representation with Knowledge Graph,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 34 (03), 2020, pp. 2901-2908. https://doi.org/10.1609/aaai.v34i03.5681 .

[3] S. Shen, et al., “Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 2020, pp. 8815-8821. https://doi.org/10.1609/aaai.v34i05.6409 .

[4] M. Pankiewicz, “Large Language Models (GPT) for automating feedback on programming assignments,” in *31st International Conference on Computers in Education (ICCE)*, vol. 13, 2023. https://doi.org/10.48550/arXiv.2307.00150 .

[5] A. Moffat, “Computing Maximized Effectiveness Distance for Recall-Based Metrics,” in *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 198-203, 1 Jan. 2018. https://doi.org/10.1109/TKDE.2017.2754371 .

[6] J. Sun, et al., “Deep learning-based methods for natural hazard named entity recognition,” *Sci Rep*., no. 12, pp. 4598, 2022. https://doi.org/10.1038/s41598-022-08667-2.

[7] A. Bello, S.-C. Ng, and M.-F. Leung, “A BERT Framework to Sentiment Analysis of Tweets,” *Sensors*, no. 23, pp. 506, 2023. https://doi.org/10.3390/s23010506 .

[8] Y. Chen, X. Kou, J. Bai, and Y. Tong, “Improving BERT With Self-Supervised Attention, ” in *IEEE Access*, vol. 9, pp. 144129-144139, 2021. https://doi.org/10.1109/ACCESS.2021.3122273 .

[9] M. H. Syed, S-T. Chung, “MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain,” *Applied Sciences*, no. 11(13), pp. 6007, 2021. https://doi.org/10.3390/app11136007 .

[10] A. Wang, and K. Cho, *BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model*, 2019. https://doi.org/10.48550/arXiv.1902.04094 .

[11] Dr. A. Shaji George, A. S. Hovan George, Dr. T. Baskarand A. S. Gabrio Martin, “Revolutionizing Business Communication: Exploring the Potential of GPT-4 in Corporate Settings,” *Partners Universal International Research Journa*l (PUIRJ), vol. 2, no. 1, pp. 149-157, Mar. 2023, https://doi.org/10.5281/zenodo.7775900 . ISSN: 2583-5602.

[12] S. Jacobs, and S. Jaschke, “Leveraging Lecture Content for Improved Feedback: Explorations with GPT-4 and Retrieval Augmented Generation, ” in *36th International Conference on Software Engineering Education and Training (CSEE&T)*, Würzburg, Germany, 2024, pp. 1-5, https://doi.org/10.1109/CSEET62301.2024.10663001 .

[13] Gabriel A., *Kensho Derived Wikimedia Dataset*, 2020. [Electronic resource]. Available: https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data. Accessed on September 1, 2024.

[14] B. D. Lund, and T. Wang, “Chatting about ChatGPT: how may AI and GPT impact academia and libraries?” *Library Hi Tech News*, vol. 40, no. 3, pp. 26-29, 2023. https://doi.org/10.1108/LHTN-01-2023-0009 .

Рекомендована кафедрою комп’ютерних наук, ВНТУ

Стаття надійшла до редакції 17.11.2024

Яровий Андрій Анатолійович — д-р техн. наук, професор, завідувач кафедри комп’ютерних наук, e-mail: a.yarovyy@vntu.edu.ua ;

Кудрявцев Дмитро Станіславович — аспірант, асистент кафедри комп’ютерних наук, e-mail: dmytro_k@vntu.edu.ua .

Вінницький національний технічний університет, Вінниця

A. A. Yarovyi[1]
D. S. Kudriavtsev[1]

Method of Text Generation Based on the BERT LLM

[1]Vinnytsia National Technical University

The application of the BERT language model for tasks of term search and generation in terminological knowledge bases (TKB) with optimization for intelligent chatbots is proposed. The architecture of the BERT model, its bidirectional attention mechanism, text processing algorithms, and the main stages of model training are described. The use of BERT for semantic search of terms and methods for adapting the model for text generation, considering the semantic value of each term, are considered. A comparative analysis of the BERT language model with models from the GPT series is carried out, highlighting the strengths and weaknesses of BERT in the context of search and generative tasks. The paper also thoroughly examines metrics for evaluating the quality of term search, such as Precision, Recall, F1-score, Mean Reciprocal Rank (MRR),

Normalized Discounted Cumulative Gain (nDCG), and others, which allow for a comprehensive assessment of the effectiveness of term search and generation. Practical aspects of integrating BERT into knowledge management systems are discussed, and recommendations are provided for fine-tuning the model for specialized TKBs. Additionally, the ethical aspects of using language models are emphasized, particularly the risks of bias in term search and generation, as well as the importance of ensuring accuracy and objectivity in the generated results. The responsible use of BERT is discussed to avoid incorrect or harmful conclusions during the automatic processing of knowledge. Software was developed for testing the BERT language model, and training of the language model was tested on various datasets. The testing results demonstrated the high efficiency of using the BERT language model, considering optimizations for text generation tasks. Potential improvements to BERT for working with TKBs are discussed, including methods for fine-tuning the model on domain-specific data, using the multilingual version of BERT for processing multilingual knowledge bases, as well as optimization techniques for improving performance in resource-constrained environments. Approaches for testing and evaluating search effectiveness are proposed, including the use of expert evaluations and automatic metrics. The final part of the article outlines future research directions, including the integration of BERT with neural search systems, automatic generation of new terms, and the expansion of knowledge management systems' functionality based on deep learning.

Keywords: BERT, terminological knowledge bases, semantic search, language models, term generation.

Yarovyi Andrii A. — Dr. Sc. (Eng.), Professor, Head of the Chair of Computer Science, e-mail: a.yarovyy@vntu.edu.ua ;

Kudriavtsev Dmytro S. — Post-Graduate Student, Assistant of the Chair of Computer Science, e-mail: dmytro_k@vntu.edu.ua