

Ганжа О. О., аспірант, Іванчук Н. В., к.т.н., доцент (Національний університет водного господарства та природокористування, м. Рівне)

ГЕНЕРАЦІЯ ТЕКСТОВИХ ОПИСІВ ЗАПЧАСТИН ІЗ ВИКОРИСТАННЯМ МОДЕЛЕЙ ГЛИБОКОГО НАВЧАННЯ ТА ГРАФОВОЇ ТЕХНОЛОГІЇ RETRIEVAL-AUGMENTED GENERATION (GRAPHRAG)

У статті представлено систему автоматизованої генерації технічних описів товарів із використанням трансформерної моделі *mt5-small* та технології *GraphRAG*. Розроблене рішення спрямоване на вирішення проблеми створення точних і структурованих описів запчастин для великих інтернет-магазинів, де ручне формування контенту є трудомістким і ресурсозатратним. Основу системи становить граф знань на базі *Neo4j*, який зберігає інформацію про товари у вигляді взаємопов'язаних вузлів, що забезпечує ефективний семантичний пошук і отримання точних релевантних даних. Модуль, реалізований на *Python*, отримує інформацію за каталожним номером, формує запит до моделі, а згенерований результат подається через *API* з простим вебінтерфейсом для користувача. Оцінювання якості роботи системи за допомогою автоматизованих метрик та експертної перевірки підтверджує високу точність, інформативність і стилістичну відповідність створених описів до експертно підготовлених текстів. Представлена робота демонструє ефективність поєднання трансформерних моделей та технології *GraphRAG* для генерації релевантного контенту в галузі електронної комерції.

Ключові слова: нейронні мережі; глибоке навчання; обробка природної мови; *GraphRAG*; граф знань.

Вступ. Застосування моделей глибокого навчання для генерації текстового контенту є надзвичайно перспективним та ефективним напрямом. Штучний інтелект (ШІ) дозволяє створювати великі обсяги унікального та релевантного тексту з мінімальними витратами часу та ресурсів. Наприклад, в інтернет-магазинах, які містять десятки чи навіть сотні тисяч товарів, кожен з яких потребує унікального, точного та *SEO*-оптимізованого опису, традиційний підхід до написання контенту потребує значних людських ресурсів. Зазвичай залучаються автори, редактори, спеціалісти з предметної галузі та



SEO-фахівці, що значно збільшує витрати та ускладнює процес. Однак сучасні моделі глибокого навчання дозволяють скоротити кількість залучених спеціалістів, підвищити продуктивність та знизити фінансові витрати.

Завдання автоматизованої генерації описів товарів належать до сфери обробки природної мови (Natural Language Processing, NLP). Ця галузь охоплює аналіз, обробку та створення різноманітного текстового контенту. Сучасні генеративні моделі трансформерної архітектури здатні створювати тексти, які мало відрізняються від написаних людиною за структурою, читабельністю та змістовністю. Проте навіть найбільш потужні моделі мають суттєві обмеження. Одним із них є точність згенерованої інформації, особливо коли йдеться про вузькоспеціалізовані технічні описи.

Попередні дослідження використання авторегресивних моделей глибокого навчання для генерації текстових описів запчастин до сільськогосподарської техніки [1] показали, що такі моделі можуть створювати добре структурований і природний текст. Однак були виявлені численні помилки та неточності в описах, зокрема у параметрах і характеристиках запчастин. Це пояснюється тим, що генеративна модель, не маючи доступу до достовірних структурованих даних, може вигадувати або спотворювати інформацію, що є критичною проблемою для технічних товарів. Подібні неточності є неприйнятними, оскільки вони можуть ввести в оману покупців, спричинити повернення товару або навіть викликати серйозні технічні проблеми у разі неправильного використання запчастин. Це демонструє, що генеративні моделі не можуть ефективно працювати у такому контексті без додаткових механізмів контролю точності.

Одним із найбільш ефективних способів підвищення точності генерації є розширення обсягу вхідної інформації, яка передається моделі. Водночас, для забезпечення зручності користувачів запит має залишатися мінімальним і не вимагати введення великої кількості параметрів вручну. Оптимальним рішенням є використання вбудованої бази знань, яка автоматично доповнює запит користувача необхідними даними перед передачею його моделі для генерації. Одним із найбільш перспективних підходів до реалізації такого рішення є технологія GraphRAG.

GraphRAG (Graph Retrieval-Augmented Generation) – це метод, що поєднує генеративні моделі (наприклад, трансформери) з графовими базами знань. Він є розвитком концепції Retrieval-

Augmented Generation (RAG), яка передбачає використання зовнішніх джерел даних для підвищення точності та релевантності згенерованого тексту. Головна особливість GraphRAG полягає у використанні графів знань, які дозволяють моделі працювати зі структурованими взаємопов'язаними даними.

Графова база знань у цьому контексті містить вузли (наприклад, категорії товарів, окремі товари, їх характеристики) та зв'язки між ними. При отриманні запиту GraphRAG дозволяє знайти релевантну інформацію шляхом семантичного пошуку по графу, витягнути необхідні дані та подати їх у структурованому вигляді генеративній моделі. Це забезпечує не лише підвищену точність описів, а й можливість врахування складних взаємозв'язків між параметрами товару.

У контексті автоматичної генерації описів за частин використання GraphRAG забезпечує низку переваг, які суттєво підвищують якість і ефективність цього процесу. По-перше, технологія дозволяє автоматично доповнювати запит користувача релевантною інформацією про товар, усуваючи потребу в ручному введенні всіх деталей. По-друге, вона гарантує актуальність і точність характеристик завдяки використанню постійно оновлюваної бази знань. Крім того, попередня обробка даних перед генерацією сприяє підвищенню релевантності створюваного тексту. Ще однією важливою перевагою є зменшення ймовірності виникнення помилок або вигаданих фактів (hallucinations), що характерні для авторегресивних моделей, які працюють без зовнішніх джерел знань. Нарешті, підхід на основі GraphRAG є масштабованим рішенням, яке легко оновлюється та доповнюється новими даними, що робить його зручним для динамічного середовища електронної торгівлі.

Застосування GraphRAG у процесі генерації текстових описів товарів відкриває нові можливості для автоматизації контент-маркетингу. Впровадження такого підходу в електронну комерцію не лише підвищує якість і точність створених текстів, а й значно зменшує витрати на їх створення. Це особливо важливо для великих інтернет-магазинів, де ручне написання текстів є надто затратним і неефективним.

Таким чином, технологія GraphRAG є перспективним напрямом розвитку генеративних моделей, який дозволяє поєднувати переваги глибокого навчання з точністю та структурованістю графових баз знань. Це робить її ідеальним рішенням для автоматичної генерації текстового контенту в сферах, де важлива точність та достовірність



інформації, таких як електронна комерція, технічна документація та інформаційні системи.

Аналіз літературних джерел. Зважаючи на актуальність тематики, значна кількість дослідників присвячує свої роботи питанням генерації тексту природньої мови на основі певних вихідних даних. Одним із таких напрямків є створення розгорнутих текстових описів на основі структурованої інформації. Так, Fajri Koto, Jey Han Lau та Timothy Baldwin [2] досліджують можливості створення текстових описів товарів на основі масивів структурованих даних, представлених у вигляді таблиць. Вони показують, що за умови наявності всіх вихідних даних та характеристик товару, можна створити достовірні текстові описи, що відображають ключові особливості продукту.

У роботі Koncel-Kedziorski та ін. [3] запропоновано метод генерації зв'язних багатосентенційних текстів із графів знань, що враховує їхню реляційну структуру без потреби лінеаризації. Використання графового трансформера в межах моделі «енкодер-декодер» дозволяє створювати більш структуровані й інформативні тексти, що перевершують традиційні методи як за автоматичними, так і за експертними оцінками.

У дослідженні Junyi Li та ін. [4] запропоновано підхід до генерації природномовних текстів на основі графів знань в умовах обмеженої кількості навчальних прикладів (few-shot). Автори ефективно використовують можливості попередньо навчених мовних моделей (PLMs), доповнюючи їх трьома ключовими компонентами: узгодженням представлень для зменшення семантичного розриву між графами та текстами, спеціалізованою лінеаризацією графів з урахуванням типів зв'язків, а також багатозадачним навчанням для кращого узгодження структури графа з відповідним текстом.

Таким чином, розвиток методів генерації тексту з використанням трансформерів та графових технологій відкриває нові можливості для автоматизації створення описів товарів, що забезпечує підвищену точність, структурованість і відповідність вимогам онлайн-торгівлі.

Набір даних. Для проведення експерименту було сформовано два взаємопов'язані набори даних. Перший – це структуровані дані, призначені для побудови графа знань, що мав максимально охоплювати інформацію про запчастини до сільськогосподарської техніки. Другий – навчальний корпус для мовної моделі, який містить пари: структуроване подання товару у форматі, зручному для

обробки моделлю, та відповідний текстовий опис, складений людиною.

Обидва набори були створені на основі відкритих джерел, насамперед – з описів товарів, опублікованих на вебсайтах інтернет-магазинів, що спеціалізуються на продажі запчастин до агротехніки. Для первинної обробки даних використовувалася модель машинного навчання [5], яка автоматично виокремлювала основні складові опису: назву товару, його коротку характеристику та технічні параметри. Технічні характеристики при цьому приводилися до уніфікованого формату виду: «Назва характеристики: Значення».

Усього було оброблено понад 19 000 неструктурованих описів. Після початкової структуризації дані додатково очищувалися від дублікатів, семантичного шуму та можливих помилок, що виникли під час автоматичного аналізу. Саме ці структуровані й очищені дані були використані для побудови графа знань.

Для створення навчального набору мовної моделі структуровані дані поєднувалися з оригінальними описами, зібраними з вебсайтів. Крім того, для донавчання моделі було сформовано додатковий корпус із 200 описів, написаних вручну спеціально для потреб дослідження. Цей набір дозволив краще налаштувати модель на стилістику та формат подання, що відповідає вимогам електронної торгівлі.

Експеримент. Розроблений програмний комплекс базується на поєднанні графової бази знань, генеративної мовної моделі та вебінтерфейсу для взаємодії з користувачем. Його архітектура побудована таким чином, щоб забезпечити гнучке й масштабоване середовище для обробки запитів на генерацію опису товарів на основі структурованих характеристик.

Фундаментом розробленої системи є графова база даних, реалізована з використанням Neo4j. У цій базі кожен товар представлений окремим вузлом із власними властивостями, зокрема: назвою, каталожним номером та коротким описом. Характеристики товару зберігаються у вигляді окремих вузлів, з'єднаних із вузлом товару орієнтованими ребрами типу HAS_ATTRIBUTE.

Ключовою особливістю запропонованої структури є збереження значень характеристик (наприклад, «Вага: 0.388 кг») не як властивостей вузлів, а саме як властивостей ребер, які з'єднують вузли товарів і характеристик. Такий підхід забезпечує гнучке масштабування бази – нові характеристики можна додавати без

модель mt5-small від Google [6]. Вибір саме цієї моделі обумовлений її підтримкою української мови та легкістю інтеграції через бібліотеку HuggingFace Transformers. Модель вирішує задачу Text2Text Generation, перетворюючи структуровані дані у повноцінні текстові описи, зручні для сприйняття людиною.

Первинне навчання моделі здійснювалося на масиві з 2500 пар: структуровані дані – опис товару. Надалі модель було донавчено на меншому, але якісніше підготовленому наборі даних, що дозволило значно підвищити точність генерації та узгодженість змісту.

Для перевірки якості генерації моделі застосовувалася метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [7], що є одним з найпоширеніших методів оцінки якості автоматичних систем текстового резюмування та машинного перекладу. Результати оцінки моделі в процесі навчання наведені в табл. 1.

Таблиця 1

Результати оцінки моделі в процесі навчання

Epoch	Training Loss	Validation Loss	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
1	0.3689	0.3225	0.2565	0.2040	0.2554	0.2559
2	0.4074	0.2778	0.2638	0.2118	0.2631	0.2635
3	0.3624	0.2705	0.2641	0.2124	0.2636	0.2636

Генерація опису реалізована як REST API, створене за допомогою фреймворку FastAPI. На бекенді функціонує єдиний ендпоінт, який приймає запит із каталожним номером, зчитує відповідні дані з бази, форматує їх та передає до мовної моделі. У відповідь повертається згенерований текстовий опис товару.

Для зручності користувачів реалізовано простий вебінтерфейс, створений на основі HTML і JavaScript (Vanilla JS), без використання сторонніх фреймворків. Через нього користувач може ввести каталожний номер і миттєво отримати опис. Приклад вебінтерфейсу в роботі наведено на рис. 2.



Генерація опису товару

661226

Згенерувати

Захист ходового варіатора молотарки комбайна 661226 Claas - D-120 [Original] Захист ходового варіатора зернозбиральних комбайнів Claas (Dominator, Mega, Tucano). Діаметр - 120 мм, кількість отворів кріплення - 3. Каталогів номери 661226.0, 0006612260 Параметри: Вага , W kg 0.28 Зовнішній діаметр , D mm 120 Товщина матеріалу , S mm 1.5 Параметри можуть відрізнятися від фактичних Оддиниці продажу Оддиниця продажу 1 штука Кількість в упаковці 1 штука

Рис. 2. Вигляд вебінтерфейсу генерації опису товару

Усі компоненти системи – графова база, API-сервер і вебінтерфейс – упаковані в окремі Docker-контейнери. Це забезпечує простоту розгортання системи в будь-якому середовищі без складних налаштувань. Для запуску використовується Docker Compose, який координує одночасну роботу всіх сервісів. Окремий том зберігання даних гарантує збереження інформації навіть після перезавантаження системи.

Таким чином, представлена система є комплексним рішенням для автоматизованого формування описів товарів на основі структурованих графових даних, що ефективно поєднує сучасні підходи до зберігання знань і генерації природної мови.

Результати дослідження. Розроблений програмний комплекс забезпечує автоматичну генерацію описів товарів – зокрема, запчастин для сільськогосподарської техніки – на основі лише їх каталожного номера. Приклад згенерованого опису та відповідного зразка, підготовленого експертом, наведено в таблиці 2.

Як видно з прикладу, обидва описи подібні за стилістикою та структурою, проте не є повністю ідентичними. Застосування генеративної моделі дозволяє досягти необхідного рівня унікальності тексту, зберігаючи при цьому точність і коректність передачі характеристик, що є критично важливим у сфері електронної торгівлі технічними товарами.

Порівняння згенерованого і зразкового опису

Згенерований опис	Зразковий опис
<p>Захист ходового варіатора молотарки комбайна 661226 Claas - D-120 [Original]</p> <p>Захист ходового варіатора зернозбиральних комбайнів Claas (Dominator, Mega, Tucano). Діаметр - 120 мм, кількість отворів кріплення - 3. Каталогів номери 661226.0, 0006612260 Параметри: Вага , W kg 0.28</p> <p>Зовнішній діаметр , D mm 120</p> <p>Товщина матеріалу , S mm 1.5</p> <p>Параметри можуть відрізнятися від фактичних</p> <p>Одиниці продажу</p> <p>Одиниця продажу 1 штука</p> <p>Кількість в упаковці 1 штука</p>	<p>Захисний кожух варіатора молотарки комбайна 661226 Claas - D-120 [Original]</p> <p>Захист механізмів варіатора зернозбиральних комбайнів Claas (Dominator, Mega, Tucano). Діаметр - 120 мм, кількість отворів кріплення - 3.</p> <p>Каталогові номери 661226.0, 0006612260</p> <p>Параметри</p> <p>Вага , Wkg0.28</p> <p>Зовнішній діаметр , Dmm 120</p> <p>Кількість отворів кріплення, N 3</p> <p>Діаметр отворів кріплення , d4mm 6</p> <p>Товщина матеріалу , Smm1.5</p> <p>Параметри можуть відрізнятися від фактичних</p> <p>Одиниці продажу</p> <p>Одиниця продажу1 штука</p> <p>Кількість в упаковці1 штука</p>

Водночас, не всі характеристики, присутні у зразковому описі, відображено у згенерованому. Це зумовлено частковими втратами інформації на етапі підготовки бази знань для графової структури, зокрема під час обробки та очищення вихідних даних. Уникнути подібних втрат можливо шляхом більш ретельної підготовки вихідного датасету або використання повніших і точніших джерел інформації.

Аналіз згенерованих описів у порівнянні з еталонними дозволив виявити низку недоліків і специфічних випадків. Зокрема, наявність у базі даних товарів з однаковими каталожними номерами, але різними характеристиками (наприклад, деталі-аналоги від різних виробників), може призводити до неточностей у генерації. Подібні помилки також виникають у випадках, коли одна й та сама назва



використовується для кількох характеристик одного товару – наприклад, якщо в описі фігурує кілька параметрів із назвою «довжина», що стосуються різних частин виробу. Це може спричинити плутанину або порушення структури згенерованого тексту.

Варто підкреслити, що зазначені недоліки не стосуються безпосередньо роботи самої генеративної моделі, а пов'язані переважно з якістю та повнотою вихідної бази знань. За умови належної попередньої обробки даних, результати генерації цілком відповідають вимогам до точності, структури й інформативності текстових описів.

Висновки. Результати дослідження підтверджують доцільність використання трансформерних моделей у поєднанні з методами генерації з доповненням на основі графів знань (GraphRAG) для створення текстових описів запчастин. Запропонований підхід забезпечує високу точність, структурованість та релевантність згенерованих текстів, що є критично важливим для застосування в електронній комерції. Водночас виявлені потенційні недоліки, зумовлені неповнотою або неоднорідністю вихідних даних, вказують на необхідність ретельної підготовки бази знань. За умови якісного опрацювання вхідної інформації, розроблений алгоритм є ефективним інструментом для автоматизації створення товарних описів у сфері онлайн-торгівлі.

1. Ганжа О. О., Іванчук Н. В. Порівняння ефективності моделей глибокого навчання в задачах генерування текстових описів запчастин до сільськогосподарської техніки. *Вісник Національного університету водного господарства та природокористування. Сер. Технічні науки*. 2024. Вип. 1(105). С. 273–281. 2. Fajri Koto, Jey Han Lau, and Timothy Baldwin. Can Pretrained Language Models Generate Persuasive, Faithful, and Informative Ad Text for Product Descriptions?. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. Association for Computational Linguistics, Dublin, Ireland, 2022. P. 234–243. 3. Koncel-kedziorski, rik, et al. Text Generation from Knowledge Graphs with Graph Transformers. In: *Proceedings of NAACL-HLT*. 2019. P. 2284–2293. 4. Li Junyi, et al. Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021. P. 1558–1568. 5. Ганжа О. О., Іванчук Н. В. Застосування моделей глибокого навчання в задачах структуризації текстових описів запчастин до сільськогосподарської техніки. *Вісник Національного університету водного господарства та природокористування. Сер. Технічні науки*. 2024. Вип. 4(108). С. 302–309. 6. Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya

Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *Association for Computational Linguistics*. 2021. P. 483–498. **7.** Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain. Association for Computational Linguistics, 2004. P. 74–81.

REFERENCES:

1. Hanzha O. O., Ivanchuk N. V. Porivniannia efektyvnosti modelei hlybokoho navchannia v zadachakh heneruvannia tekstovykh opysiv zapchastyn do silskohospodarskoi tekhniki. *Visnyk Natsionalnoho universytetu vodnoho hospodarstva ta pryrodokorystuvannia. Ser. Tekhnichni nauky*. 2024. Vyp. 1(105). S. 273–281. **2.** Fajri Koto, Jey Han Lau, and Timothy Baldwin. Can Pretrained Language Models Generate Persuasive, Faithful, and Informative Ad Text for Product Descriptions?. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. Association for Computational Linguistics, Dublin, Ireland, 2022. P. 234–243. **3.** Koncel-kedziorski, rik, et al. Text Generation from Knowledge Graphs with Graph Transformers. In: *Proceedings of NAACL-HLT*. 2019. P. 2284–2293. **4.** LI Junyi, et al. Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021. P. 1558–1568. **5.** Hanzha O. O., Ivanchuk N. V. Zastosuvannia modelei hlybokoho navchannia v zadachakh strukturyzatsii tekstovykh opysiv zapchastyn do silskohospodarskoi tekhniki. *Visnyk Natsionalnoho universytetu vodnoho hospodarstva ta pryrodokorystuvannia. Ser. Tekhnichni nauky*. 2024. Vyp. 4(108). S. 302–309. **6.** Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *Association for Computational Linguistics*. 2021. P. 483–498. **7.** Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain. Association for Computational Linguistics, 2004. P. 74–81.

Hanzha O. O., Post-graduate Student, Ivanchuk N. V., Candidate of Engineering (Ph.D.), Associate Professor (National University of Water and Environmental Engineering, Rivne)

GENERATION OF SPARE PARTS TEXT DESCRIPTIONS USING DEEP LEARNING MODELS AND GRAPH-BASED RETRIEVAL-AUGMENTED GENERATION (GRAPH-RAG) TECHNOLOGY

This paper presents the development and evaluation of a system for the automated generation of technical product descriptions using



deep learning models and GraphRAG technology. The proposed approach addresses the challenges of producing accurate, structured, and SEO-optimized descriptions for spare parts in large-scale e-commerce environments, where manual content creation is time-consuming and resource-intensive. Transformer-based generative models can produce fluent and human-like text, but often suffer from factual inaccuracies when lacking access to domain-specific data. To overcome this, the system integrates a Neo4j-based knowledge graph, which stores products and their attributes as interconnected nodes and relationship properties, enabling semantic enrichment and accurate data retrieval.

A custom Python module extracts relevant information based on catalog numbers, structures it into a standardized prompt, and sends it to a multilingual transformer model (mt5-small) fine-tuned on domain-specific data. The Graph Retrieval-Augmented Generation (GraphRAG) approach allows the model to incorporate external structured knowledge, improving factual consistency and reducing hallucinations. The system includes a REST API built with FastAPI and a lightweight HTML+JS interface, with all components containerized using Docker for flexible deployment.

Evaluation using ROUGE metrics demonstrates that the generated texts are stylistically and structurally aligned with expert-written descriptions, while the use of a knowledge graph significantly improves accuracy. Identified issues such as inconsistent or duplicated source data underline the importance of high-quality knowledge base design. Overall, the study confirms that combining transformer models with GraphRAG-based knowledge augmentation offers a scalable and effective solution for generating precise and contextually relevant product content in the domain of agricultural machinery parts.

Keywords: neural networks; deep learning; natural language processing; GraphRAG; knowledge graph.