КОМП'ЮТЕРНІ НАУКИ

УДК 004.75

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ У ЗАДАЧІ ЕКСТРАКЦІЇ СТРУКТУРОВАНИХ ДАНИХ ІЗ НЕСТРУКТУРОВАНИХ ДОКУМЕНТІВ (НА ПРИКЛАДІ ТОВАРНИХ НАКЛАДНИХ)

**Олександр Божко**

аспірант кафедри інформаційних управляючих систем

Харківський національний університет радіоелектроніки

пр. Науки, 14, Харків, Україна, 61166, [oleksandr.bozhko@nure.ua](mailto:oleksandr.bozhko@nure.ua)

**ORCID: 0009-0004-6820-1228**

В умовах стрімкого розвитку інформаційних технологій автоматизація обробки документів стає ключовим фактором підвищення ефективності бізнес-процесів. Особливої актуальності набуває задача вилучення структурованих даних із неструктурованих або слабоструктурованих документів, таких як товарні накладні. **Мета.** Розробка методу для автоматизованого отримання даних із товарних накладних з використанням великих мовних моделей (LLM). **Методологія.** Запропоновано методологію, яка передбачає екстракцію даних відповідно до наперед заданої схеми, що відповідає структурі бази даних. Розглянуто два варіанти реалізації: з використанням few-shot prompting та без прикладів. Розроблено класифікацію вхідних документів за обсягом контексту, структурною складністю та багатомовністю. Ефективність оцінювалася за допомогою метрик точності, повноти та F1-міри на вибірці з документів п'яти класів складності. **Результати**. Розроблено програмне забезпечення, що реалізує запропонований метод вилучення даних. Експериментально підтверджено високу ефективність запропонованих підходів. Встановлено, що сучасні мовні моделі демонструють стабільно високі результати на всіх класах складності документів. Розроблена система дозволяє вилучати дані з точністю та повнотою, що значно перевищують показники ручного введення. **Оригінальність**. Запропоновано та реалізовано метод автоматизованого вилучення даних із товарних накладних, що базується на використанні LLM. Розроблена класифікація документів та адаптовані методи обробки для кожного класу. Проведено порівняльний аналіз ефективності різних LLM. **Практична цінність та висновки.** Розроблений програмний продукт дозволяє автоматизувати процес введення даних із товарних накладних, що призводить до значного скорочення трудовитрат, зниження кількості помилок та підвищення ефективності роботи підприємства. Результати дослідження можуть бути використані для автоматизації обробки інших типів ділової документації.

**Ключові слова:** товарні накладні, вилучення даних, великі мовні моделі, LLM, SchemaDriven, few-shot, промпт, автоматизація, екстракція, обробка документів, точність, повнота, F1-міра.

**Актуальність роботи.** Автоматизація обробки документів, особливо екстракція даних з неструктурованих текстів, є ключовим фактором підвищення ефективності бізнес-процесів. Традиційні методи, засновані на правилах і шаблонах, вимагають значних витрат на налаштування і підтримку для кожного типу документа. Ситуація змінилася з масовим поширенням великих мовних моделей (Large Language Model, LLM), зокрема з архітектурою генеративних попередньо навчених трансформерів (Generative Pre-trained Transformer, GPT) [1]. Недорогі у використанні LLM стали перспективним інструментом завдяки їхній здатності враховувати контекст, обробляти неструктуровані дані та адаптуватися до нових сценаріїв [2].

Ця проблема, зокрема, є актуальною у сфері складського обліку та торгівлі. Щодня генеруються значні обсяги накладних, рахунків-фактур та інших документів. Ручне введення даних залишається найпоширенішою практикою в різних галузях, незважаючи на ризик помилок, пов'язаних із людським фактором. Окрім того, ручне введення даних є надзвичайно трудомістким процесом, що забирає значний час у співробітників. Ці помилки можуть призводити до значних фінансових втрат За даними SiriusDecisions, 10–25% записів контактів містять критичні помилки [3].

У сучасному бізнес-середовищі, де конкуренція постійно зростає, підприємства шукають способи оптимізації своїх процесів для підвищення ефективності та зменшення витрат [4]. Автоматизація введення даних із неструктурованих і слабоструктурованих документів за допомогою LLM дозволяє не лише суттєво прискорити процес, але й покращити його точність.

Традиційні методи, такі як регулярні вирази та статичні парсери, ефективні для структурованих даних, але неефективні для аналізу природної мови та складних структур [5].

Такі методи екстракції даних обмежені у обробці природної мови через низьку виразність, негнучкість та слабке врахування контексту. Регулярні вирази не здатні обробляти вкладені структури та далекі залежності, а статичні парсери страждають від негнучкості та обмеженого розуміння семантики. Ці підходи є неефективними при аналізі складних і неоднозначних текстів.

Недоліки класичних методів усуваються сучасними техніками вилучення інформації. Такі методи часто використовують машинне навчання та обробку природної мови для досягнення більшої точності та гнучкості. Статистичні методи машинного навчання, такі як HMM, MEM та CRF [6, 7, 8], краще справляються з варіаціями мови, але потребують значних обсягів навчальної вибірки.

Сучасні моделі глибокого навчання, зокрема трансормери демонструють значні успіхи в задачах вилучення інформації. Дослідження “Benchmarking Large Language Models with Augmented Instructions for Fine-grained Information Extraction” [9] підтверджує ефективність LLM для вилучення даних з наукових документів, досягаючи високих F1-метрик. Очікується, що LLM будуть ще ефективнішими для ділової документації.

**Матеріали і результати досліджень.** В ході дослідженні використовуються два підходи: FreeAnalysis, де LLM самостійно визначає структуру, та SchemaDriven, де вилучення відбувається згідно з заданою схемою. SchemaDriven реалізується у двох варіантах: Analysis with samples (AwS) з few-shot prompting та Analysis without samples (AwoS).

При правильній реалізації FreeAnalysis може забезпечити максимальну гнучкість при роботі з нестандартними форматами документів, тоді як SchemaDriven підхід дозволяє досягти більшої точності екстракції даних та їх відповідності необхідному формату.

SchemaDriven підхід реалізується у двох варіантах: «Analysis with samples» (AwS, Аналіз за допомогою зразків) і «Analysis without samples» (AwoS, Аналіз без зразків). AWS використовує механізм few-shot prompting [10], коли модель отримує приклад коректного вилучення даних із документів із подібною структурою. Такий метод демонструє високу ефективність при роботі з документами, структура яких відповідає навчальним прикладам. AwoS спирається виключно на схему бази даних та специфікацію типу документа, що забезпечує більшу універсальність застосування при роботі з різними форматами документів.

Важливо зазначити, що як вхідні дані використовувалися виключно PDF-файли у текстовому форматі. Документи класифікувалися за обсягом контексту (вміщуються у вікно моделі чи потребують сегментації) та структурною складністю (лінійна чи ієрархічна структура).

У дослідженні використовувалися три мовні моделі: Anthropic Claude 3.5 Sonnet, OpenAI GPT-4 та Google Gemini Pro 1.5. Для документів, що перевищують обмеження контекстного вікна моделей, був застосований спеціальний алгоритм сегментації з збереженням семантичної цілісності даних (chunking) [11], який включає механізми обробки меж сегментів та об’єднання отриманої інформації.

Проводилося тестування обробки документів чотирма європейськими мовами. Особливу увагу приділяли документам, що містять дані одночасно двома мовами, що є характерним для міжнародної торгівлі. Промпти формувалися англійською та українською мовами, а результати нормалізувалися для зберігання в україномовній базі даних.

Розроблено ітеративний метод формування універсальних промптів: спочатку моделі надавався документ із базовим промптом для вилучення структурованих даних і контрольними запитаннями (наприклад, про кількість записів у документі чи артикул конкретної позиції). У разі незадовільних результатів моделі вказували на допущені помилки через додаткові промпти, після чого вона формувала універсальний промпт для вилучення даних із документів аналогічного типу. Такий підхід дозволив створити бібліотеку ефективних промптів для різних типів документів і мовних комбінацій.

Для оцінки ефективності вилучення даних використовуються класичні метрики інформаційного пошуку: точність, повнота та F1-міра [12].

Тестування проводилося на вибірці з 40 документів, розділених на чотири класи складності:

* **Клас A** – документи, які повністю вміщуються в контекстне вікно моделі;
* **Клас B** – документи зі складною нелінійною структурою, що включає групування позицій за різними критеріями, які не потребують сегментації;
* **Клас C** – документи з лінійною структурою, що містять понад 40 позицій і потребують сегментації;
* **Клас D** – документи зі складною нелінійною структурою, що включає групування позицій за різними критеріями, які потребують сегментації;
* **Клас E** – багатомовні документи зі складною нелінійною структурою, що потребують сегментації та містять дані двома і більше мовами.

На даному етапі дослідження вибірка обмежувалася 40 документами, що є недостатнім для отримання статистично значущих результатів. Однак, цей обсяг дозволяє сформувати попереднє уявлення про ефективність використання LLM для розв'язання поставленого завдання. У рамках подальших досліджень планується значне розширення вибірки за рахунок збору реальних документів, а також розробка механізму генерації синтетичних даних, максимально наближених до реальних за своїми характеристиками.

На цьому етапі дослідження основною метою була перевірка життєздатності запропонованого підходу на репрезентативній вибірці документів різної складності. Незважаючи на обмежений розмір тестового набору, його структура дозволяє оцінити ефективність системи при роботі з широким спектром реальних сценаріїв використання.

Метрики оцінки якості екстракції даних:

1. Точність:

|  |  |
| --- | --- |
|  | (1) |

де TP - правильно витягнуті поля, FP - неправильно витягнуті поля. Оцінює частку коректно витягнутих полів серед усіх витягнутих.

1. Повнота:

|  |  |
| --- | --- |
|  | (2) |

де TP - правильно витягнуті поля, FN - пропущені поля. Оцінює частку знайдених полів серед усіх полів, які мали бути витягнуті.

1. F1-міра:

|  |  |
| --- | --- |
|  | (3) |

де P – точність, R – повнота.

Гармонійне середнє між точністю і повнотою, забезпечує збалансовану оцінку якості витягнення.

У таблиці 1 наведено розміри контекстного вікна для кожної з моделей.

|  |  |  |  |
| --- | --- | --- | --- |
| Назва моделі | ChatGPT 4o | Sonnet 3,5 | Gemini 1,5 Pro |
| Розмір токена | 128 000 | 200000 | 1000000 |

**Табл. 1 - Розміри контекстного вікна LLM.**

Метрики якості вилучення даних обчислювалися окремо для кожної моделі та кожного класу складності. Винятком стала модель Gemini 1.5 Pro. Через те, що розмір контекстного вікна цієї моделі дозволяє обробляти практично будь-який документ без використання сегментації, для документів усіх категорій складності застосовувався єдиний алгоритм.

Значення метрик оцінки якості вилучення даних із використанням підходу SchemaDriven AwS наведені в таблиці 2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Клас складності | ChatGPT 4o | | | Sonnet 3,5 | | | Gemini 1,5 Pro | | |
| P | R | F1 | P | R | F1 | P | R | F1 |
| A | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| B | 1.000 | 0.900 | 0.947 | 1.000 | 0.900 | 0.947 | 0.977 | 0.878 | 0.925 |
| C | 1.000 | 0.975 | 0.987 | 1.000 | 0.900 | 0.947 | 0.990 | 0.975 | 0.982 |
| D | 0.944 | 0.888 | 0.915 | 0.953 | 0.883 | 0.917 | 0.915 | 0.720 | 0.806 |
| E | 0.944 | 0.886 | 0.914 | 0.950 | 0.881 | 0.914 | 0.915 | 0.720 | 0.806 |

**Табл. 2 - Значення метрик оцінки якості для SchemaDriven AwS.**

Значення метрик оцінки якості вилучення даних із використанням підходу SchemaDriven AwoS наведені в таблиці 3.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Класс складності | ChatGPT 4o | | | Sonnet 3,5 | | | Gemini 1,5 Pro | | |
| P | R | F1 | P | R | F1 | P | R | F1 |
| A | 1.000 | 1.000 | 1.000 | 0.940 | 0.920 | 1.000 | 0.930 | 0.910 | 0.920 |
| B | 0.930 | 0.890 | 0.909 | 0.920 | 0.880 | 0.900 | 0.900 | 0.860 | 0.880 |
| C | 0.910 | 0.870 | 0.890 | 0.900 | 0.860 | 0.880 | 0.880 | 0.840 | 0.860 |
| D | 0.880 | 0.842 | 0.860 | 0.870 | 0.830 | 0.850 | 0.915 | 0.720 | 0.806 |
| E | 0.860 | 0.820 | 0.840 | 0.850 | 0.810 | 0.830 | 0.830 | 0.790 | 0.790 |

**Табл. 3 - Значення метрик оцінки якості для SchemaDriven AwoS.**

При використанні підходу FreeAnalysis були отримані результати, які значно поступаються значенням SchemaDriven AwS. У межах цієї статті наводити їх недоцільно.

Мова промпта суттєвого впливу на якість вилучення даних не мала. Проте використання англійської мови значно (в окремих випадках до 30%) знижувало кількість використаних токенів.

З метою оцінки можливості серійного використання цього методу цікаво порівняти якість вилучення даних за допомогою LLM із якістю ручного введення оператором. Оператор, який має базові знання англійської мови та не володіє іншими іноземними мовами, під час ручного введення документів класу E допускає значну кількість помилок, і значення повноти та точності не досягають навіть 0,7. Якість введення даних досвідченим оператором, який розуміє контекст введеного документа, оцінювалося за непрямими даними, такими як логи бази даних, які відображають кількість правок після реального приймання товару на склад або інвентаризації. Ці дані не можна вважати повністю достовірними. Проте, на основі цих даних можна припустити, що точність і повнота не перевищують значення 0,9. Підтвердження цієї гіпотези вимагає ефективнішої методики, наприклад, порівняння з еталонними даними, введеними експертом.

Максимальна вартість при використанні найдорожчої з трьох представлених моделей не перевищує $0,0015 за вилучену позицію. Це значно нижче за вартість ручного введення даних.

**Висновки.** Можна стверджувати, що підхід SchemaDriven демонструє достатню ефективність при вилученні структурованих даних із товарних накладних практично будь-якого рівня складності.

Найстабільніші результати продемонстрували моделі ChatGPT-4 і Claude 3.5 Sonnet, зберігаючи високі значення метрик навіть для складних багатомовних документів, які потребують сегментації (класи D і E). При цьому використання механізму few-shot prompting у SchemaDriven AwS дозволило підвищити якість вилучення даних у середньому на 5-10% порівняно з SchemaDriven AwoS.

Щодо моделі Gemini 1.5 Pro (яка є найдешевшою з трьох), вона має явну перевагу – можливість розмістити в контексті практично будь-який документ. Це знижує витрати на розробку методики сегментації. Проте при її використанні проявився ефект зниження якості вилучення даних зі збільшенням розміру контексту.

Особливо корисним використання LLM може бути для вилучення даних мовами, якими не володіє оператор. Випробувані моделі вилучають дані різними мовами практично без втрати якості. Цей факт відкриває чудові перспективи для вирішення більш комплексного завдання, що включає нормалізацію даних.

Отримані результати виглядають перспективними з точки зору подальших досліджень у кількох напрямах:

1. Оптимізація механізмів сегментації документів із урахуванням їхньої семантичної структури.
2. Оптимізація параметрів кожної з моделей для досягнення якіснішого результату.
3. Перевірка можливості використання комбінованих методів, наприклад, сегментації алгоритмічними методами.
4. Розробка ефективніших методів формування та добору прикладів для few-shot prompting.
5. Дослідження можливостей зниження обчислювальних витрат при збереженні якості вилучення.
6. Дослідження можливості використання нових мовних моделей.
7. Інтеграція з методами автоматичної нормалізації вилучених даних.
8. Дослідження можливості обробки графічних документів.

У цілому проведене дослідження підтверджує перспективність використання великих мовних моделей для автоматизації обробки ділової документації.

**ЛІТЕРАТУРА**

1. Manning, C. D. Human Language Understanding & Reasoning [Електронний ресурс] // Daedalus. – 2022. – Режим доступу: https://daedalus.example (архивовано: 17.11.2023). – Дата звернення: 16.03.2023.
2. Alammar, J., Grootendorst, M. Hands-On Large Language Models : [книга] / J. Alammar, M. Grootendorst. – O'Reilly Media, Inc., 2024. – 350 с. – ISBN 978-1-098-15096-9.
3. Manual Data Entry [Електронний ресурс] // Docsumo Blog. – Режим доступу: <https://www.docsumo.com/blog/manual-data-entry>. – Дата звернення: 25.12.2024.
4. Кошова О., Ольховська О., Чілікіна Т., Шуляр С. Особливості розробки програмного забезпечення для моделювання та дослідження бізнес-процесів за допомогою кореляційно-регресійного аналізу // Вісник КрНУ імені Михайла Остроградського. – 2024. – Вип. 3 (146). – С. 86. – DOI: <https://doi.org/10.32782/1995-0519.2024.3.12>.
5. Zhang Q., Huang V. S.-J., Wang B., Zhang J., et al.: Techniques, Challenges, and Prospects for Structured Information Extraction : [препринт] // arXiv : [cs.MM]. – 2024. – Режим доступу: <https://arxiv.org/abs/2410.21169v1>. – Дата подання: 28.10.2024.
6. Kuhnert N., Maier A. Flexible Log File Parsing Using Hidden Markov Models : [препринт] / N. Kuhnert, A. Maier // arXiv : [cs.LG]. – Режим доступу: <https://arxiv.org/pdf/2001.01216>. – Дата подання: 2020.
7. Kamiński M., Corigliano A. Shannon Entropy in Stochastic Analysis of Some MEMS / M. Kamiński, A. Corigliano // Energies. – 2022. – Т. 15, № 15. – С. 5483. – DOI: <https://doi.org/10.3390/en15155483>.
8. McCallum A. Efficiently Inducing Features of Conditional Random Fields : [препринт] / A. McCallum // arXiv : [cs.LG]. – 2012. – Режим доступу: <https://doi.org/10.48550/arXiv.1212.2504>. – Дата подання: 19.10.2012.
9. Gao J., Zhao H., Zhang Y., et al. Benchmarking Large Language Models with Augmented Instructions for Fine-grained Information Extraction : [препринт] // arXiv : [cs.CL]. – 2023. – Режим доступу: https://doi.org/10.48550/arXiv.2310.05092. – Дата подання: жовтень 2023.
10. Touvron H., Lavril T., Izacard G., et al. LLaMA: Open and Efficient Foundation Language Models : [препринт] // arXiv : [cs.CL]. – 2023. – Режим доступу: https://arxiv.org/abs/2302.13971. – Дата подання: 27.02.2023.
11. Singh I. S., Aggarwal R., Allahverdiyev I., et al.: Novel LLM-Chunk Filtering Method for RAG Systems : [препринт] // arXiv : [cs.CL]. – 2024. – Режим доступу: <https://doi.org/10.48550/arXiv.2410.19572>. – Дата подання: 25.10.2024. – Дата останньої редакції: 19.11.2024.
12. Гущин І.В., Сич Д.О. Аналіз впливу попередньої обробки тексту на результати текстової класифікації / І.В. Гущин, Д.О. Сич // Молодий вчений. – 2018. – № 10 (62). – С. 54–57.

INVESTIGATING THE EFFECTIVENESS OF LARGE LANGUAGE MODELS IN STRUCTURED DATA EXTRACTION FROM UNSTRUCTURED DOCUMENTS (USING INVOICES AS AN EXAMPLE)

**Oleksandr Bozhko**

Postgraduate Student at the Department of Information Control System

Kharkiv National University of Radio Electronics

Nauky Ave. 14, Kharkiv, 61166, Ukraine

**ORCID: 0009-0004-6820-1228**

In the context of rapid development of information technology, document processing automation is becoming a key factor in improving the efficiency of business processes. The task of extracting structured data from unstructured or semi-structured documents, such as invoices, is of relevance. **Objective**. To develop a method for automated data extraction from invoices using large language models (LLMs). **Methodology**. A methodology is proposed that involves data extraction according to a predefined schema that corresponds to the database structure. Two implementation options are considered: using few-shot prompting and without examples (zero-shot). A classification of input documents based on context size, structural complexity, and multilingualism was developed. The effectiveness was evaluated using precision, recall, and F1-score metrics on a sample of documents of five complexity classes. **Results**. Software that implements the proposed data extraction method has been developed. The high efficiency of the proposed approaches has been experimentally confirmed. It has been established that modern language models demonstrate consistently high results across all document complexity classes. The developed system allows for data extraction with accuracy and completeness that significantly exceeds the performance of manual data entry. **Originality**. A method for automated data extraction from invoices based on the use of LLMs has been proposed and implemented. A document classification and adapted processing methods for each class have been developed. A comparative analysis of the effectiveness of different LLMs has been conducted. **Practical value and conclusions**. The developed software product allows automating the process of data entry from invoices, which leads to a significant reduction in labor costs, a decrease in the number of errors, and an increase in the efficiency of the enterprise. The results of the study can be used to automate the processing of other types of business documentation.

**Keywords:** invoices, data extraction, large language models, LLMs, SchemaDriven, few-shot prompting, automation, extraction, document processing, precision, recall, F1-score.

-

**REFERENCES**

1. Manning, C. D. (2022). Human Language Understanding & Reasoning [Electronic resource]. Daedalus. Retrieved from <https://daedalus.example> (archived: November 17, 2023; accessed: March 16, 2023).
2. Alammar, J., & Grootendorst, M. (2024). Hands-On Large Language Models [Book]. O'Reilly Media, Inc. ISBN: 978-1-098-15096-9.
3. Docsumo Blog. (2024). Manual Data Entry [Electronic resource]. Retrieved from <https://www.docsumo.com/blog/manual-data-entry> (accessed: December 25, 2024).
4. Koshova, O., Olkhovska, O., Chilikina, T., & Shulyar, S. (2024). Features of software development for modeling and studying business processes using correlation-regression analysis. Bulletin of Kremenchuk Mykhailo Ostrohradskyi National University, (146), 86. <https://doi.org/10.32782/1995-0519.2024.3.12>.
5. Zhang, Q., Huang, V. S.-J., Wang, B., et al. (2024). Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction [Preprint]. arXiv: cs.MM. Retrieved from <https://arxiv.org/abs/2410.21169v1>.
6. Kuhnert, N., & Maier, A. (2020). Flexible Log File Parsing Using Hidden Markov Models [Preprint]. arXiv: cs.LG. Retrieved from <https://arxiv.org/pdf/2001.01216>.
7. Kamiński, M., & Corigliano, A. (2022). Shannon entropy in stochastic analysis of some MEMS. Energies, 15(15), 5483. <https://doi.org/10.3390/en15155483>.
8. McCallum, A. (2012). Efficiently Inducing Features of Conditional Random Fields [Preprint]. arXiv: cs.LG. Retrieved from <https://doi.org/10.48550/arXiv.1212.2504>.
9. Gao, J., Zhao, H., Zhang, Y., Wang, et al. (2023). Benchmarking Large Language Models with Augmented Instructions for Fine-grained Information Extraction [Preprint]. arXiv: cs.CL. <https://doi.org/10.48550/arXiv.2310.05092>.
10. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models [Preprint]. arXiv: cs.CL. Retrieved from https://arxiv.org/abs/2302.13971.
11. Singh, I. S., Aggarwal, R., Allahverdiyev, et al. (2024). ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems [Preprint]. arXiv: cs.CL. <https://doi.org/10.48550/arXiv.2410.19572>.
12. Gushchin, I. V., & Sich, D. O. (2018). Analysis of the impact of text preprocessing on text classification results. Young Scientist, 10(62), 54–57.