프롬프트 엔지니어링을 이용한 자동화된 지식 유닛 처리

미라 마울레쇼바 [1], 밀란 후슈카 [1], 시바니 산제이 콜레카르 [2], 김경백 [2]
[1] 체코 생명과학대학교 시스템 공학과 (체코, 프라하)
[2] 전남대학교 인공지능학과
mauleshova@pef.czu.cz, houska@pef.czu.cz, shivanikolekar@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

Prompt Engineering automated Knowledge Unit processing

Mira Mauleshova[1], Milan Houška[1], Shivani Kolekar[2], Kyungbaek Kim[2]
[1]System Engineering Department
Czech University of Life Sciences
Prague, Czech Republic

[2]Dept. of Artificial Intelligence Convergence,
Chonnam National University
Gwangju, South Korea

Abstract

This paper is dedicated to automated Knowledge Unit extraction by employing Prompt Engineering within the Generative AI models. The experiment is conducted analyzing the ability of GPT-4 to extract Knowledge Units' elements in an automated manner by the created prompts. Verification is proceeded by comparison among results from GPT-4 and the elements of Knowledge Units coming from the original methodology of knowledge-structured texts' creation. The results of the research prove that thanks to the created prompts, Knowledge Units are successfully extracted from unstructured text.

1. Introduction

In the era of "Big Data", when the amount of data is growing exponentially, the decision–making process is a complex issue since humanity deals not only with a tremendous volume of structured data but also with unstructured ones [1]. To tackle this immersive volume of data, automation of the processes is the way, how to effectively acquire, formalize, code, test and maintain knowledge, which are the main phases of Knowledge Engineering [2].

Concerning the present report, the focus is on automated knowledge extraction. This step is associated with the NLP task, namely, information extraction, which is implemented to acquire structured information from unstructured text. The task proceeded in order to transform this information into the organized, structured form, which can be afterwards used for analysis as well as further processing [3].

The aim of the research is associated with automated knowledge extraction, namely, to procedural knowledge extraction represented in the form of Knowledge Units. According to [4], analytical form of Knowledge Unit is represented accordingly, Problem situation (*X*), Elementary problem (*Y*), Goal of solving elementary problem (*Z*), Solution of elementary problem (*Q*).

2. Related Works

Firstly, within the research, the source of knowledge is text, to be precise, unstructured text, which is characterized by high content variation and complexity due to the absence of a predefined structure [5]. Currently, information extraction methods from unstructured texts are used in various fields, mainly in the field of medicine [6, 7, 8] but also in consulting [9], geology [10].

The second point of automated knowledge extraction is related to the target knowledge to be acquired. The target knowledge to be obtained is procedural knowledge that can be represented by production rules as well as Knowledge Units. However, the focus of the research is on procedural knowledge in the form of Knowledge Units.

Nowadays Knowledge Units are extracted manually [11]. There is room for improvement, as the current process of manual Knowledge Unit extraction is time-consuming. Nevertheless, by dint of AI technologies, there is a possibility to manipulate knowledge automatically. Thanks to Prompt Engineering, the mentioned bottleneck of ineffective procedural knowledge extraction is eliminated. Prompt engineering plays a vital role in designing queries, which serves in obtaining high-quality output from generative AI models [12]. The generative AI model applied in the current paper is ChatGPT.

Advantages of the mentioned Generative Pre-trained Transformer have been already taken in a bunch of the papers, for instance, within the paper, where entity-centric constructed prompts guide the GPT models to recognize as well as to tag

materials-related entities in science texts [3]. Another paper is dedicated to demonstration of the potential of generative AI, such as GPT-4 to extract information from research articles about microbial performance [13]. Moreover, in the paper on automation information extraction and knowledge synthesis from ecological papers, the model performance is optimized not by model parameters fine-tuning, but thanks to prompt fine-tuning [14].

In this paper, we intend to automatically extract procedural knowledge in form of Knowledge Unit from unstructured text by means of designed prompts levering in ChatGPT.

3. Prompt creation for automated Knowledge Unit extraction:

Knowledge Unit is extracted from the unstructured text thanks to the prompts utilized in the LLM [15]. In order to set a general precondition for queries, the customization is set within the configuration of ChatGPT, to be concrete, set a role for it, explaining the domain of the Knowledge Unit in general, emphasizing 4 elements of Knowledge Unit.

Within the experiment, there are input texts, which are characterized as unstructured text. The experiment aims to extract 4 Knowledge Unit elements in analytical form automatically from unstructured text thanks to the designed prompt.

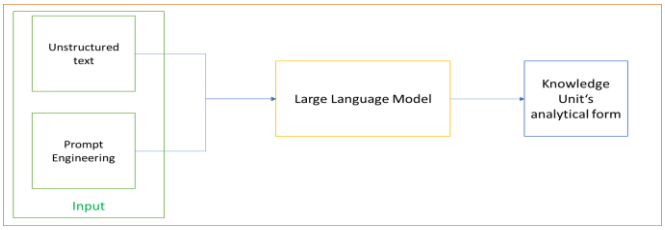

Figure 1: Knowledge Unit extraction in analytical form

Prompt: s The prompt is designed to extract elements of Knowledge Unit one by one into analytical form. In order to proceed with extraction, the prompt focuses on each element separately, emphasizing the main characteristic of each part, and afterwards all elements are put together to obtain the whole Knowledge Unit via one request. It is crucial to point out that the first prompt is created based on the description from "*Methodology for the creation of knowledge-structured texts"* [11]. Prompt is highlighted in the **Appendix**; however, the main points of the prompt are mentioned in **Figure 2** below:

Problem situation	specific circumstances or conditions
Elementary problem	state specific needs or requirements
Goal	distinguish goals from other aspects of elementary knowledge
Solution	• expressed using a main clause or compound sentences. • avoid the use of excluding conjunctions

Figure 2: Main points of 1st prompt

Afterwords, the mentioned prompt is improved by additional analysis of phrases associated with each Knowledge Unit element. The improved prompt is included into **Appendix.**

Originally, [11] has developed the methodology of Knowledge Unit extraction from unstructured text in a manual way. The acquired Knowledge Unit is then represented in analytical form, which breaks down Knowledge Unit into elements, such as Problem situation (*X*), Elementary problem (*Y*), Goal of solving elementary problem (*Z*), Solution of elementary problem (*Q*). Once all elements of Knowledge

Unit are extracted, unstructured text can be transformed into structured text containing Knowledge Unit based on a defined structure. Comparing to the initial research, our report is an enhancement of the current process in the manner of a new approach to tackle the issue of Knowledge Unit extraction. In the current report, we suggest the automated way of Knowledge Unit extraction via designed prompts leveraging in ChatGPT.

4. Evaluation/Case Study

Description: The case study demonstrates the ability of GPT-4 to extract Knowledge Unit from unstructured text by use of the created prompt. The input text originates from "*Methodology for the creation of knowledge-structured texts"* [11]. The instance of **input unstructured text** is provided below:

"*The goal is to ensure that a candidate whose health suitability does not correspond to the performance of the anticipated work is not assigned to a job under conditions with expected health demands*".

Compare results: The results obtained from GPT-4 are compared with the results originated from "*Methodology for the creation of knowledge-structured texts*".

Comparison of Knowledge Unit elements

	From the original source	Initial prompt	Improved prompt
X	In conditions with expected health demands	Is not assigned to a job under conditions with expected health demands	Under conditions with expected health demands
Y	For the performance of work	A candidate whose health suitability does not correspond to the performance of the anticipated work	A candidate whose health suitability does not correspond to the performance of the anticipated work
Z	Ensure that no candidate/employee whose health suitability does not correspond to the performance of the anticipated work is included	To ensure that a candidate whose health suitability does not correspond to the performance of the anticipated work is not assigned	To ensure that a candidate whose health suitability does not correspond to the performance of the anticipated work
Q	Every job applicant and employee in an employment relationship will attend the initial medical examination	ensure that... is not assigned	is not assigned to a job

Figure 3: Comparison of Knowledge Unit's elements (manual/automated ways)

The comparison presented in **Figure 3** reveals that the phrases extracted by means of prompts (initial prompt and improved prompt) match the analytical form of Knowledge Units stated in the examples from the original source. Nevertheless, Solution of elementary problem (*Q*) of the initial and improved prompts does not correspond to Solution of elementary problem (*Q*) from the original source. The reason lies in the fact that in the initial unstructured text, Solution of elementary problem (*Q*) is not a part of the mentioned text. In other words, it is added by human Expert. Therefore, GPT-4 generated the Solution of elementary problem (*Q*) in case of initial and improved prompts based on context information, which makes sense. It is essential to note that the phrases representing elements of Knowledge Unit may not correspond verbatim to those in the original examples. The underlying meaning of GPT-4 obtained elements should match the original ones in case of extraction of information included in input unstructured text. However, in case, where element is not mentioned in the text, then GPT-4 should be able to generate it. The experiment is conducted with other instances of unstructured texts, in order to verify preciseness of results. Therefore, it is wrapped up that GPT-4 is able to extract Knowledge Units from unstructured text.

Conclusion and Future Work

The research is dedicated to improvement of the current Knowledge Unit manipulation since nowadays Knowledge Unit processing is conducted manually. Thanks to new AI technologies there is a possibility to manipulate with Knowledge Units in an automated manner.

The current paper deals with processes of Knowledge Unit extraction in analytical form. The extraction is currently being implemented manually. However, our research proves that Prompt Engineering assets an ability to extract knowledge in an automatic manner by the designed prompt. The results of extraction align with the general description of Knowledge Unit parts. The built prompts can be utilized in various domains of unstructured texts since the mentioned prompts are generalized.

The potential limitations could be related to the length of unstructured texts, as in the conducted experiments there are texts with short passages. While, in the case of longer unstructured texts, for instance, scientific papers, the prompt should be more specific, as the range of information is higher. The further work will be focused on prompt optimization by key words, results' comparison of optimized prompts, and work with different quality level of input texts.

Acknowledgement

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2022-00156287, 50%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, 50%) grant funded by the Korea government(MSIT).

References

[1] Vaidya, Madhavi, Katkar, Shweta. 2022. Exploring performance and predictive analytics of agriculture data. AI, Edge and IoT-based Smart Agriculture. pp. 409-436. ISBN 9780128236949

[2] Mařík, Vladimír. 1997. Umělá inteligence 2. Prague: Academia. ISBN 80-200-0504-8. 11

[3] Choi, Jaewoong, Lee, Byungj. 2024. "Accelerating Materials Language Processing with Large Language Models." Communications Materials 5(1). doi: 10.1038/s43246-024-00449-9.

[4] Brožová, Helena, Houška, Milan et al. 2011. Modelování znalostí. Prague: Professional Publishing. ISBN 978-80-7431-069-0

[5] CORTICAL.IO. ©2022. Business. Cortical.io. [online]. [cit. 2023-02-07]. Available from: https://www.cortical.io/blog/4-things-you-should-know-about-unstructured-text/

[6] Jouffroy, Jordan, Feldman, Sarah, Lerner, Ivan, Rance, Bastien, Burgun, Anita, Nauraz, Antoine. 2021. Hybrid Deep Learning for Medication – Related Information Extraction from Clinical Texts in French: MedExt Algorithm Development Study. JMIR Medical Informatics. Vol. 9 (3). DOI: 10.2196/17934

[7] Steinkamp, Jackson, Bala, Wasif, Sharma, Abhinav, Kantrowitz, Jacob. 2020. Task definition, annotated dataset, and supervised natural language processing models from symptom extraction from unstructured clinical notes. Journal of Biomedical Informatics. Vol. 102. DOI: 10.1016/j.jbi.2019.103354.

[8] Manesh, Batta. 2018. Machine learning algorithms – A Review. International Journal of Science and Research (IJSR). Vol. 18, Issue 8. DOI: 10.21275/ART20203995

[9] Maitra, Anutosh, Garg, Shivam, Sengupta, Shubhashis. 2020. Enabling Interactive Answering of Procedural Questions. Natural Language Processing and Information Systems. Lecture Notes in Computer Science(), vol 12089. Springer, Cham. DOI: 10.1007/978-3-030- 51310-8_7

[10] Qiu, Qinjun, Xie, Zhong, Wu, Liang, Tao, Liufeng. 2020. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. Earth Science Informatics. Vom. 13, 1393-1410. DOI: https://doi.org/10.1007/s12145-020-00527-9

[11] Houšková Beránková, Martina, Mudrychová, Kristýna, Peták, Michal, Horáková, Tereza, Houška, Milan. 2021. Metodika tvorby znalostně strukturovaných textů [Methodology of creating knowledge-structured texts]. E – Approved Methodology (NmetS), ISBN 978-80-213-3120-4, Agreement on the application of the methodology with Česká pošta s.p., closing date

4/1/2021. C – The result is use without restrictions on the range of users, Ministry of Labor and Social Affairs, 14/06/2021.

[12] IBM.com ©2024. What is prompt engineering?. IBM.com. [online]. [cit. 2024-06-13]. Available from: What Is Prompt Engineering? | IBM

[13] Xiao, Zhengyang, Li, Weny, Moon, Hannah, Roell, Garrett, Chen, Yixin, Tang, Yinjie. 2023. "Generative Artificial Intelligence GPT-4 Accelerates Knowledge Mining and Machine Learning for Synthetic Biology." ACS Synthetic Biology 12(10):2973 – 2982. doi: 10.1021/acssynbio.3c00310

[14] Scheepens, Daan, Milliard, Josep, Farrell, Maxwel, Newbold, Tim. 2024. "Large Language Models Help Facilitate the Automated Synthesis of Information on Potential Pest Controllers." Methods in Ecology and Evolution. doi: 10.1111/2041-210X.14341.

[15] OpenAI.com. 2023. GPT-4. Openai.com. [online]. [cit. 2024-06-13]. Available from: GPT-4 | OpenAI

Appendix

Initial prompt: *"Which phrases from the text are related to the parts of the Knowledge Unit, representing the Problem Situation, Elementary Problem, Goal of Solving the Elementary Problem, and Solution to the Elementary Problem?*

To determine the Problem Situation:

• Look for expressions that are typically associated with the description of specific circumstances or conditions in which the problem occurs.

To determine the Elementary Problem:

• Concentrate on specific expressions that usually state specific needs or requirements.

To identify the goal of solving the elementary problem for each specific problem:

• To distinguish goals from other aspects of elementary knowledge, it is useful to pay attention to the use of subordinate purpose clauses.

To determine a clear solution to the elementary problem in the context of the goal:

• Pay attention to the sentence structure in the text, where the solution should be expressed using a main clause or compound sentences.

• Avoid the use of excluding conjunctions, such as the use of 'OR'.

The result should include the Problem Situation, Elementary Problem, Goal of Solving the Elementary Problem, Solution to the Elementary Problem. Put the results into a table".

Improved prompt: "Improved prompt

"*Which phrases from the text are related to the parts of the Knowledge Unit, representing the Problem Situation, Elementary Problem, Goal of Solving the Elementary Problem, and Solution to the Elementary Problem?*

To determine the Problem Situation:

• Look for expressions that are typically associated with the description of specific circumstances or conditions in which the problem occurs.

• Focus on phrases like '...within...', '...during...', '...at...', 'when', 'as soon as', 'around', 'throughout', 'under the condition that', 'provided that', 'because of', 'due to', 'leading to', 'resulting in', 'in relation to', 'compared with', which can indicate the context in which the Problem Situation is placed.

To determine the Elementary Problem:

• Concentrate on specific expressions that usually state specific needs or requirements.

• Look for phrases like '...if you want...', '...if you must...', '...if you need...', 'should you encounter', 'in the event of', 'requires that', 'necessitates', 'demands', 'calls for', 'is crucial', 'is vital', 'presents the challenge of', 'poses the question of', 'to achive', 'to resolve', which indicate direct conditions or situations that the user needs to address.

To identify the goal of solving the elementary problem for each specific problem:

• Search the text for expressions and phrases like '...for...', '...to...', '...for the purpose of...', '...with the goal of...', 'aimed at', intended to', 'with the objective of', 'in order to', 'designed to', 'geared towards', 'resulting in', 'leading to', 'for achieving', 'with the purpose of', 'to finalize', 'to culminate in', which signal a purpose or intent. These expressions often follow the description of the problem and serve to clearly define it within the context of its solution.

• To distinguish goals from other aspects of elementary knowledge, it is useful to pay attention to the use of subordinate purpose clauses.

To determine a clear solution to the elementary problem in the context of the goal:

• Pay attention to the sentence structure in the text, where the solution should be expressed using a main clause or compound sentences.

• Focus on phrases like 'implementing', 'applying', 'by following', 'through the use of', 'via', 'utilizing', 'developing', 'constructing', 'and', 'additionally', 'resulting in', 'leading to', 'completing', 'finalizing'.

• Avoid the use of excluding conjunctions, such as the use of 'OR'.

The result should include the Problem Situation, Elementary Problem, Goal of Solving the Elementary Problem, Solution to the Elementary Problem. Put the results into a table"