Effectiveness of Security Export Control Ontology for Predicting Answer Type and Regulation Categories

Rafal Rzepka[*]
Hokkaido University
Sapporo, Japan
rzepka@ist.hokudai.ac.jp

Akihiko Obayashi
Hokkaido University
Sapporo, Japan
obayashi@eng.hokudai.ac.jp

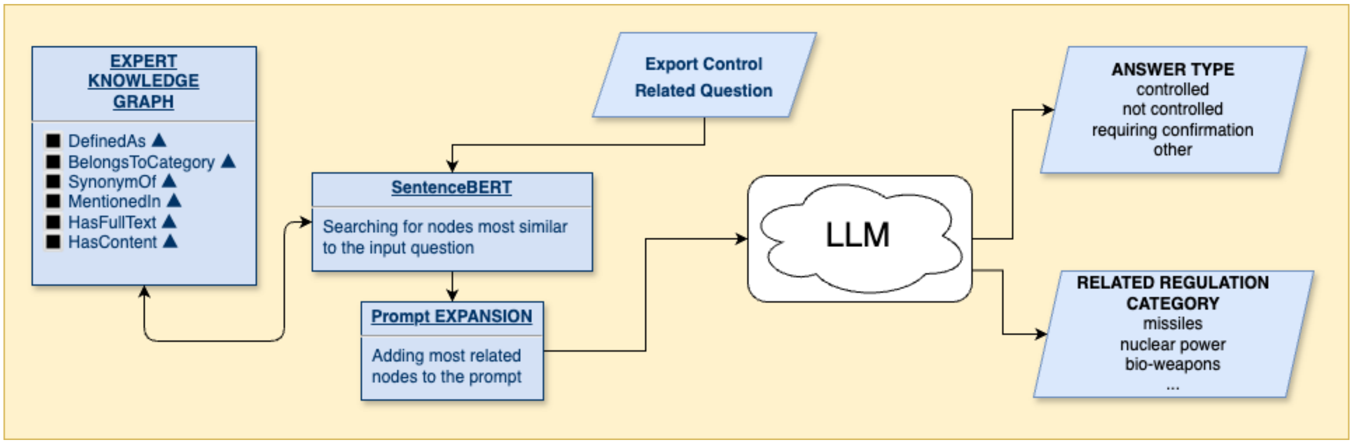


Figure 1: Overview of the experiment examining effectiveness of a graph-based prompt expansion

Abstract

In this paper we present results of our experiments investigating if an expert knowledge graph can improve Large Language Models accuracy in predicting correct answer labels and regulations related to the topic of security export control. As the lack of related data prevents machine-learning or fine-tuning approaches, we implement prompt expansion by searching most relevant nodes of the graph and adding the expanded context to the prompt. Results of our experiments show that the addition improved answer type selection but clearly hamper the capability of finding a correct regulation category.

CCS Concepts

• **Information systems** → **Information retrieval**; **Expert systems**; • **Applied computing** → **Law**.

Keywords

Expert Systems, GraphRAG, Large Language Models, Knowledge Graph, Security Export Control

ACM Reference Format:
Rafal Rzepka and Akihiko Obayashi. 2024. Effectiveness of Security Export Control Ontology for Predicting Answer Type and Regulation Categories. In *Proceedings of 8th International Conference on Advances in Artificial Intelligence (ICAAI '24).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3704137.3704180

1 Introduction

Export control regulations play a crucial role in international trade and national security. These regulations govern the transfer of sensitive goods, technologies, and information across borders, ensuring that potentially dangerous items do not fall into the wrong hands. However, navigating the complex landscape of security export control regulations presents significant challenges for businesses, researchers, and government agencies alike. One of the primary difficulties in understanding and complying with export control regulations stems from the distributed nature of controlled item descriptions. Often, a single controlled item or technology may be described across multiple paragraphs in different categories within the regulatory documents. This fragmentation makes it challenging for stakeholders to gain a comprehensive understanding of the regulations applicable to their specific items or activities. Consequently, there is a pressing need for more effective and efficient methods to search, interpret, and apply these regulations.

[*]Corresponding author

ICAAI '24, October 17–19, 2024, London, UK

ACM ISBN 979-8-4007-1801-4/24/10
https://doi.org/10.1145/3704137.3704180

As we confirmed in our previous experiments [11] Large Language Models (LLMs) are prone to various problems such as hallucinations, also in very specific areas of expert knowledge. Retrieval Augmented Generation (RAG)[2], originally proposed in 2001 [6], is one of the methods which aims to mitigate these issues by grounding LLM outputs in relevant, factual information. RAG combines the power of large language models with information retrieval systems, allowing for more accurate and contextually relevant responses. In the context of export control regulations, RAG can significantly enhance the search and interpretation process by retrieving pertinent regulatory information and using it to generate accurate and comprehensive answers to specific queries.

To investigate if RAG can help in our task which tackles a narrow domain and Japanese language, we created an ontology of over 7,000 nodes and performed a preliminary experiments to determine if a graph-based addition could help in improving performance of our security export control-related dialog system which algorithm could be supported by LLMs without losing credibility due to incorrectly generated information.

2 Related Work

As mentioned in the introduction, traditional RAG approaches have became popular over last few years)[2] but may still struggle with the interconnected and distributed nature of export control regulations. This is where graph-based RAG systems come into play [1, 3, 5, 7, 12, 14]. By representing the regulatory information as a graph, these systems can capture the complex relationships between different items, technologies, and regulatory categories. The literature suggests that graph-based RAG offers several advantages in this domain.

- Contextual understanding: Graphs can represent the hierarchical and cross-referential structure of regulations, allowing for a more nuanced understanding of how different parts of the regulations relate to each other.
- Improved retrieval: Graph-based retrieval can leverage the connections between different regulatory elements to find relevant information that might be missed by traditional keyword-based searches.
- Holistic interpretation: By traversing the graph, these systems can provide a more comprehensive view of the regulations applicable to a specific item or technology, even when the relevant information is scattered across multiple sections.
- Dynamic updates: As regulations change, graph structures can be easily updated to reflect new connections or modifications, ensuring that the system remains current and accurate.

When it comes to security export control as a topic of natural language processing, our work is rather isolated [9–11]. However, there is a vast range of studies dealing with RAGs, graph knowledge and highly specialized domains like medicine [8] or law [13]. The main difference between our task and others, is that the graph data is relatively small and there is no data that could be used by classical machine-learning data (especially in Japanese language).

3 Graph Data Creation

To ensure the quality of our ontology, we have created the nodes half manually by collecting data from openly available sources and the following relation categories were used. Statistical information about the graph are given in Table 1. and Figure 2.

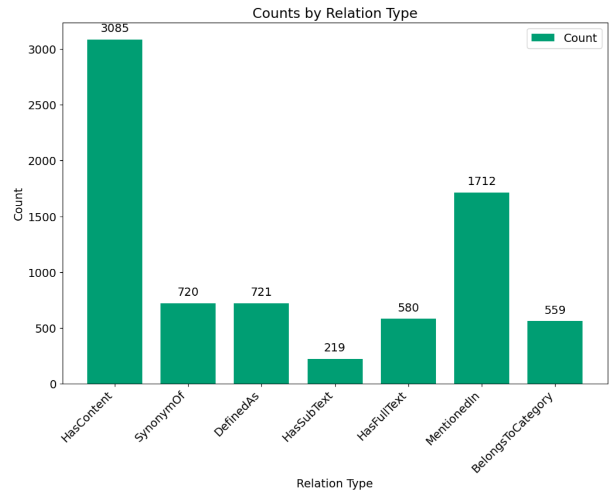
- **HasContent**: This relation indicates that a regulation specified by number (subject) contains specific text (object) describing details related to certain controlled items or technologies. It is used to identify the contents or specifications detailed in the regulatory text and linking knowledge via connections to the same regulation. The regulation texts are provided by Japanese government in XML format. An example entry is "Article 8, Paragraph 1, Item 9 (iv) of the Ordinance on Cargo and Related Matters" *HasContent* "Shortest vector or closest vector problem related to lattices" (translated from Japanese).
- **HasFullText**: This relation indicates the main part of a given regulation. Both *HasFullText* and *HasSubText* are combined in the *HasContent* category. The division was made to allow more precise linking for cases when a broader or more detailed explanations are required by our dialog system.
- **HasSubText**: This relation denotes the additional text or supplementary information related to the main content. It might include elaborations, explanations, or examples that provide further context to the primary regulations.
- **SynonymOf**: This relation is used to link terms or phrases that are considered equivalent in the context of export control regulations. It helps in mapping different terminologies that refer to the same concept or item. The glossary we utilized is provided by the government as well. An example is "CPU", *SynonymOf*, "Multi Chip Module (MCM)".
- **DefinedAs**: This relation specifies that a term or item is explicitly defined in the context of the regulations. It provides a clarification as standard definitions are often different form the ones in export control context. A translated example is "Closure part" *DefinedAs* "A part that is movable to control the fluid and functions as a closure during valve closure.". This part is not included into XML files and was extracted from the official Excel file.
- **MentionedIn**: This relation shows that a particular term from glossary is referenced within a regulation. It helps in identifying where specific items or concepts are discussed or cited across different regulatory texts.
- **BelongsToCategory**: This relation categorizes a term or item into a specific classification within security export control regulations. There are 15 topic categories under security export control [9]: (1) Arms, (2) Nuclear Power, (3) Chemical Weapons, (3-2), Biological Weapons (4) Missiles, (5) Advanced Mate- rials, (6) Material Processing, (7) Electronics, (8) High Performance Computers, (9) Telecommunication, (10) Sensors, (11) Navigation Systems, (12) Oceanic Technologies, (13) Propulsion Devices, (14) Other Technologies and (15) Sensitive Items. The glossary terms were searched through the regulation text and the category was decided by

the related regulation number provided by XML file. "Triisopropyl phosphite" *BelongsToCategory* "Chemical Weapons" is an example entry.

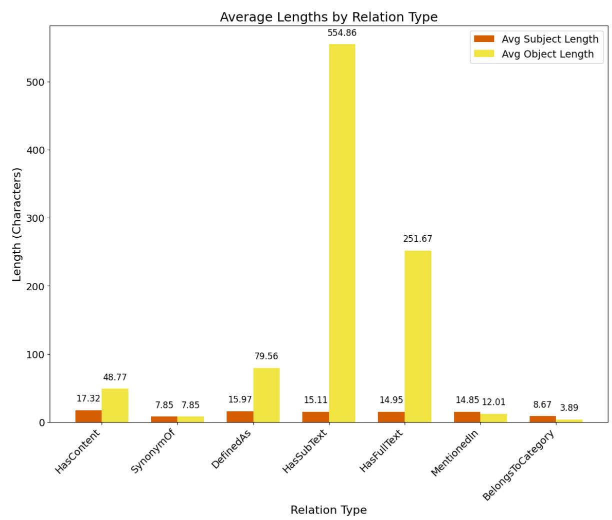
It should be noted that *BelongsToCategory* numbers play important role in the regulation category prediction task.

Relation Type	Count	Subjects	Objects
HasContent	3,085	17.32	48.77
SynonymOf	720	7.85	7.85
DefinedAs	721	15.97	79.56
HasSubText	219	15.11	554.86
HasFullText	580	14.95	251.67
MentionedIn	1,712	14.85	12.01
BelongsToCategory	559	8.67	3.89

Table 1: Number of relations and average lengths of subjects and objects of graph triples



(a) Counts by Relation Type



(b) Average Lengths by Relation Type

Figure 2: Statistical Data Visualization

4 Method Used in Experiments

We developed a simple graph-based Retrieval-Augmented Generation (RAG) algorithm inspired by GraphRAG to perform experiments with classification of security export control questions from a previously developed QA dataset [10]. This system combines a knowledge graph representation of export control regulations with state-of-the-art language models to provide accurate and context-aware responses to classification queries.

4.1 Knowledge Graph Construction

The system begins by constructing a directed graph from a set of triples representing export control concepts and their relationships described in the previous section. Each triple consists of a subject, predicate, and object. The graph is built using the NetworkX library[1], with nodes representing concepts and edges representing relationships between them. To ensure comprehensive coverage of relationships, we implement bidirectional connections by adding both the original edge and its inverse for each triple.

4.2 Semantic Embedding and Indexing

To enable efficient semantic search, we employ the Sentence Transformers model *all-MiniLM-L6-v2*[2] to generate embeddings for each node in the graph. These embeddings should capture the semantic meaning of the concepts represented by the nodes. We then use the FAISS [4] library to create an index of these embeddings, allowing for rapid similarity-based retrieval.

4.3 Query Processing and Context Retrieval

When a query is received, the system performs the following steps:

(1) Encodes the query using the same Sentence Transformer model.
(2) Uses the FAISS index to find the k most semantically similar nodes to the query.
(3) For each similar node, retrieves its neighbors and the relationships (edges) connecting them.
(4) Constructs a context string that includes the relevant nodes, their neighbors, and the relationships between them.

4.4 Utilized Language Models

First, we have experimented with various open-source models[3] but the lack of efficient GPUs or small size of some of the models prevented us from achieving comparable results. For that reason, before further experiments with free models, we integrated three

[1]https://networkx.org
[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3]elyza/Llama-3-ELYZA-JP-8B, cyberagent/open-calm-small, tokyotech-llm/Swallow-7b-instruct-hf, matsuo-lab/weblab-10b-instruction-sft, stabilityai/japanese-stablelm-instruct-alpha-7b, stabilityai/japanese-stablelm-instruct-gamma-7b, llm-jp/llm-jp-13b-instruct-full-jaster-v1.0llm-jp/llm-jp-13b-v1.0, elyza/ELYZA-japanese-Llama-2-7b-instruct, cyberagent/open-calm-7b, cyberagent/calm3-22b-chat

advanced language models provided by OpenAI into our system to generate responses based on the retrieved context (and without it): *gpt-3.5-turbo-0125*, *gpt-4-0125-preview* and *gpt-4o-mini*. These models are accessed via OpenAI API, the code is written in Python.

4.5 Prompts and Question Answering

We investigate two capabilities of the models, given a question, one is to guess a type of answer and second is to list all related regulation categories. Two key prompts are used in the system, translated here from Japanese to English:

(1) Answer Type Classification Prompt: "The following is a question about export control. Please choose one of the following options that best fits the answer: Applicable, Not Applicable, Requires Confirmation, Other. Question: 'question' Answer:"
(2) Regulation Identification Prompt: "The following is a question about export control. Please display all relevant items from the Combined Matrix[4] in Python list format. If not directly related to the Combined Matrix, display 'NA'. For example, ['5'], ['1', '3-2', '15'], or ['NA']. The Combined Matrix table divides regulated goods and technologies into item numbers (1-15) of the Export Trade Control Order Appendix 1 and Foreign Exchange Order Appendix. It is a list of provisions from government ordinances, ministerial ordinances, notices, etc.: 1 (Weapons), 2 (Nuclear), 3 (Chemical Weapons), 3-2 (Biological Weapons), 4 (Missiles), 5 (Advanced Materials), 6 (Material Processing), 7 (Electronics), 8 (Computers), 9 (Communications), 10 (Sensors), 11 (Navigation Equipment), 12 (Marine-related), 13 (Propulsion Devices), 14 (Miscellaneous), 15 (Sensitive Items). Please create a list of item numbers only. Using category name words is prohibited. Question: 'question' Answer:"

For each query, the system generates responses to both prompts using the chosen language model. The responses are then processed to extract the predicted answer type and the relevant regulation items. There are two types of prompts – one is just a question, one is a combination of context retrieved by SentenceBERT with the question.

4.6 Evaluation Metrics

To assess the performance of our system, we implement two primary evaluation metrics:

(1) **Answer Type Prediction** Accuracy: This measures the percentage of correctly predicted answer types compared to human-labeled gold standard answers.
(2) **Regulation Prediction** Agreement: We calculate an agreement score based on the F1 score between the predicted regulation items and the human-annotated gold standard. This score considers both precision and recall in the prediction of relevant regulation items. We have noticed that prompting model to output a python list of regulation numbers not only simplifies the calculation but also forces it to keep the required, mostly numeric, format.

[4]Combined Matrix (*gattai matorikusu*) is a popular expression describing the set of security export control regulations in Japan.

These metrics allow us to quantitatively compare the performance of different language models and assess the overall effectiveness of our graph-based RAG system.

5 Experiments and Results

We have used all for models to answer 116 questions from the dataset. An example question is (translated) "There is a laser module consisting of three stacked single semiconductor laser bars, each with an oscillation wavelength of 680 nanometers and an average output of 60 watts, resulting in a total average output of 180 watts and an average output density of 600 watts per square centimeter. According to Item 10, Paragraph (8) <Ordinance Article 9, Item 10, Sub-item (ii) 3>, which regulates devices with an average output exceeding 100 watts, does this laser module fall under the regulation?". All questions are labelled with all regulation numbers (if not applicable "NA" label is given) and four answer types: *Applicable*, *Not Applicable*, *Requires Confirmation* and *Other*. As mentioned before, all questions are repeated with graph-based knowledge and without it. The results are shown in Tables 2 and 3.

Table 2: Answer Type Prediction (accuracy)

Model	With Graph	Without Graph
gpt-3.5-turbo-0125	**33.62**	11.21
gpt-4o-mini	11.21	5.17
gpt-4-0125-preview	17.24	12.93

Table 3: Regulation Number Prediction (f-score)

Model	With Graph	Without Graph
gpt-3.5-turbo-0125	26.14	36.77
gpt-4o-mini	33.44	39.31
gpt-4-0125-preview	40.56	**43.01**

6 Discussion

Overall accuracy of all models was low for both models. Context addition for Answer Type prediction task substantially improved the results but it had an opposite effect on Regulation Number prediction task. All models were confused by additional information which raise a question about correctness of context generation, which will be our future work.

We could not confirm higher scores achieved for Answer Type by [11] – 31.0% for gpt-3.5 and 50.0% for gpt-4, however the direct comparison cannot be made as we used slightly different versions of both models and asked more questions. Nevertheless, this discrepancy suggest a need for repeating experiments few times and experimenting with parameters[5].

Although *BelongsToCategory* should link the keyword terms to the correct regulations, it is surprising that the knowledge is not utilized well in this case suggesting that the conversion from the

[5]In our experiments we used max_tokens=62, n=1, stop=None, temperature=0.01, and system role was set "you are an export control specialist".

regulation numbers to category numbers was difficult for all tested models.

7 Conclusion and Future Work

Our study on graph-based Retrieval-Augmented Generation (RAG) systems for export control classification revealed several key insights. Overall accuracy for both Answer Type and Regulation Number prediction tasks was lower than expected across all tested models. Notably, context addition improved Answer Type prediction but negatively impacted Regulation Number prediction, highlighting the complexity of knowledge integration in this domain.

We observed challenges in model utilization of structured knowledge, particularly in converting regulation numbers to category numbers. These findings, along with discrepancies from previous studies, underscore the need for robust, reproducible experiments in this field.

Future work should focus on a) optimizing context generation and testing other knowledge integration techniques; b) enhancing experimental robustness and reproducibility; c) developing task-specific model adaptations and more nuanced evaluation metrics.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23K11757.

References

[1] Chen Chen, Yufei Wang, Bing Li, and Kwok-Yan Lam. 2022. Knowledge is flat: A seq2seq generative framework for various knowledge graph completion. *arXiv preprint arXiv:2209.07299* (2022).

[2] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[3] Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. 2023. Think and Retrieval: A Hypothesis Knowledge Graph Enhanced Medical Large Language Models. *arXiv preprint arXiv:2312.15883* (2023).

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[5] Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846* (2023).

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[7] Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6418–6425.

[8] Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics* 40, 6 (2024).

[9] Akihiko Obayashi and Rafal Rzepka. 2019. Towards Interactive Advisory System for Security Export Control. In *Proceedings of IJCAI Workshop on Language Sense on Computer, Macau*.

[10] Akihiko Obayashi and Rafal Rzepka. 2022. Expanding Export Control-related Data for Expert System. *Procedia Computer Science* 207 (2022), 3065–3072. https://doi.org/10.1016/j.procs.2022.09.364 Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 26th International Conference KES2022.

[11] R. Rzepka, S. Muraji, and A. Obayashi. 2023. Expert Evaluation of Export Control-Related Question Answering Capabilities of LLMs. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–6. https://doi.org/10.1109/CSDE59766.2023.10487735

[12] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. *arXiv preprint arXiv:2203.10321* (2022).

[13] Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning*. Springer, 445–460.

[14] Donghan Yu and Yiming Yang. 2023. Retrieval-enhanced generative model for large-scale knowledge graph completion. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2334–2338.

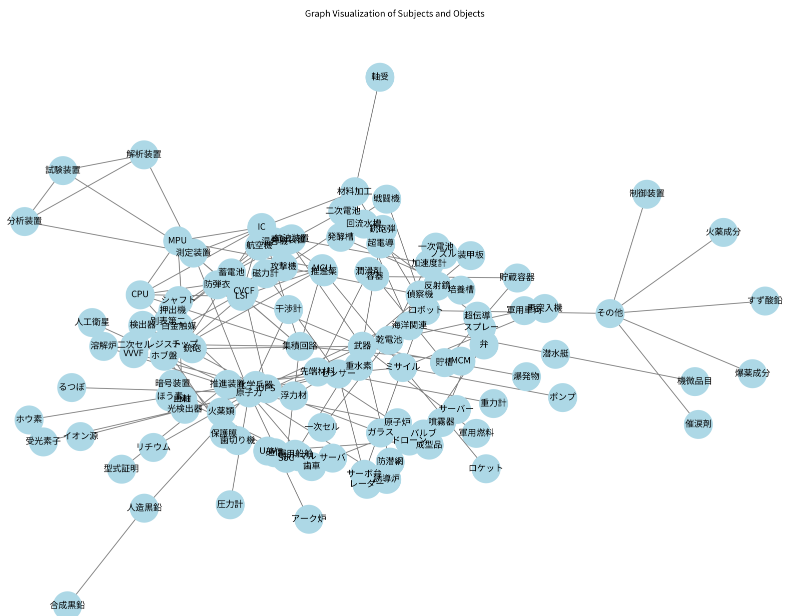


Figure 3: Part of the created ontology (limited to 4-character words) in original Japanese