原　著

Towards Efficient Retrieval of Similar Medication Incident Reports: An Exploration of Lexical and Semantic-based Search Techniques

Zoie SY WONG[1)]*

薬剤の医療事故事例の効率的な探索に向けて：自然言語処理技術による探索的研究

ウォン スイー[1)]*

〔Abstract〕

Background: This study evaluates two efficient information retrieval (IR) methods, namely the keyword-based Rank BM 25 and the semantic-based SBERT. Method: Drawn from publicly available resources in Japan, the experiment is conducted using a set of synthetically generated and machine translated incident reports in English. Result: In the context of rephrase sentence search, the mean reciprocal rank (MRR) values for Rank BM 25, SBERT-Document and SBERT-Sentence are 100%, 80% and 80%. In terms of semantic-based queries, mean average precision (MAP) values stand at 73% for Rank BM25 (a lexical method), 22% for SBERT-Document, and 2% for SBERT-Sentence (a semantic method). Discussion: The study outcome reveals that applied into incident report retrieval application, Rank BM 25 outperforms SBERT on various search queries, encompassing root-cause-based, phrase-based, characteristic-based, and lightly rephrase sentence searches. Future directions to semantic-based search methods include their refinement through the scaling of models with clinically-focused information resources and the integration of named entities as knowledge graphs, while ensuring efficiency in digital implementation.

〔Key words〕 Incident reports, medication errors, information retrieval (IR), Rank BM 25, sentence-BERT

〔要　旨〕

本研究では，キーワードベースのRank BM 25と意味ベースのSBERTという２つの効率的な情報検索方法の評価を行う。Rephraseタスクでは，Rank BM 25，SBERT-Document，SBERT-Sentenceのmean reciprocal rank（MRR）値はそれぞれ100％，80％，80％となった。意味ベースのクエリに関して，mean average precision（MAP）値はキーワードベースの検索方法であるRank BM25が73％，SBERT-Documentが22％，意味ベースのSBERT-Sentenceが２％となった。本研究の結果から，インシデントレポートの検索タスクにおいて，Rank BM 25がルート原因ベース，フレーズベース，特徴ベース，軽く再表現された文章の検索Lightly-Rephraseタスクなど，さまざまな検索語句でSBERTより優れた結果を示した。意味ベースの検索方法の今後の展望としては，プログラムの実装の効率性を担保しつつ，臨床情報資源を活用したモデルのスケーリングや，知識グラフによる固有表現の統合による改善などが考えられる。

[1] Graduate School of Public Health, St. Luke's International University・聖路加国際大学公衆衛生学研究科
ORCiD ID: Zoie SY Wong https://orcid.org/0000-0003-4499-9779
*Corresponding author.

受付　2023年９月27日　　受理　2023年11月24日

〔キーワーズ〕インシデントレポート，薬剤，情報検索（IR），Rank BM 25，Sentence-BERT

1．Introduction

Globally, various nations, regions, and healthcare organizations have implemented distinct incident reporting systems to enhance hospital safety. An illustrative example is the Japan Council of Quality Health Care (JQ)[1], which has established a comprehensive national incident reporting system to aggregate incident data from over a thousand participating hospitals within Japan. Furthermore, numerous other countries and territories, including the United Kingdom[2], Australia[3], and Hong Kong[4, 5], similarly have instituted systems for the collection of incident reports at the national or regional level. Guided by the World Health Organization's Minimal Information Model for Patient Safety (MIM PS)[6], conventional incident reporting systems commonly employ unstructured free text input for documenting incident details, including background, causative factors, consequences, and ameliorating actions. However, harnessing the wealth of incident reports presents inherent challenges, stemming from the inability to efficiently retrieving pertinent incidents when needed[7].

Information retrieval (IR) has long been a complex and pivotal area within the domain of natural language processing (NLP). Among the established techniques, keyword match search[8], which entails searching for an exact match with the input provided, have reached a level of maturity and serve as commonplace IR methods. However, when dealing with content-rich medical notes, reliance solely on simplistic keyword matching is unrealistic for effectively retrieving relevant records. Taking incident reports as an example, which encompass medical concepts, relations, time dependencies, contributory factors, and underlying root causes of incidents occurred in the hospital systems, a more nuanced approach is imperative. Furthermore, there are inherent variability in the manner and tone used to describe similar incidents from different frontline reporters, all of which make simplistic keyword matching insufficient.

Recent advancements in NLP have opened new opportunities for enhancing similarity search, transitioning from lexical-level search methods to more advanced semantic search approaches. In the context of lexical search engines, notable models such as Rank BM25[9] have proven highly efficient in implementing keyword-based ranking. This efficiency makes them well-suited for instantaneously querying large databases, also potentially suitable for retrieval of incident reports. Furthermore, the recent transformer-based Large Language Model (LLM) architectures, has presented significant potential within the NLP domain. Bidirectional Encoder Representations from Transformers (BERT)[10] has exhibited exceptional performance across various NLP challenges. By adapting pre-trained BERT architecture, sentence embedding method, known as Sentence-BERT (SBERT)[11], has become a new benchmark in semantic textual similarity tasks with an efficient computational advantage. These advancements offer possibilities for practical, instantaneous and context-sensitive IR which is particularly relevant in scenarios involving complex medical textual data, like incident reports.

Denecke (2017)[12] evaluated the retrieval of incident reports by comparing conventional lexical methods with semantic approaches using medical ontologies in the German language. Since the advent of BERT in 2018, to the best of our knowledge, there is a gap in the literature regarding the assessment of text similarity search methods for incident reports. We undertake this study using incident reports of medication errors as a focal point, as medication errors are a significant concern in hospital settings, representing one of the primary drivers of avoidable harm within hospitals. The primary goal of this experimental study is to assess whether modern advancements in language model-powered search method outperform traditional keyword-based search approach. To accomplish this, we systematically analyze the search performance of two distinct methods: lexical-based using Rank BM25 and semantic-based using SBERT. By taking into consideration of the specific properties and context of incident reports, as well as the unique demands of conducting incident report searches within healthcare settings, we attempt to suggest strategies for the integration of future semantic search methods into the clinical search process. This study contributes to the optimal approaches for effectively searching and retrieving incident reports in the healthcare domain. The capability to effectively execute search tasks is pivotal, as it not only facilitates subsequent analyses and the generation of insights based on incident reports but also supports subsequently analysis and presentation of similar cases occurred in the past.

2. Methods

We describe the design of the experimental study here. In the first chapter, we address the process of creating synthetic incident reports in the English language, drawing from open-source incident reports originally documented in the Japanese language. In Chapter 2.2, we provide a concise introduction to the two distinct search methods that were employed in this research. Subsequently, in Chapter 2.3, we address the details of the methods used for the evaluation of search performance. This study does not require ethical approval because the incident report data sources utilized in the generation of synthetic reports were publicly accessible on the JQ platform[1], which is accessible freely to all interested parties. Python frameworks included panda, nltk, rank_bm25, sentence-transformers were used.

2.1. Synthetic data generation by machine translation and data argumentation

We created incident reports of medication errors in English using the ethical and open-source repository of incident reports collected by JQ. All the original reports underwent deidentification by JQ before being shared on the open platform, ensuring that they do not raise any ethical concerns regarding the potential tracking of patient identities. We randomly selected 40 incident reports of medication errors from the openly accessible JQ incident report database from the years 2010 to 2020. The randomization process was to guarantee that the selected samples constituted a representative subset of the incident reports. (Note: This set of 40 reports had previously been annotated based on medication-incident report ontologies and had been utilized in other research studies). We only utilized the free-text portions of this dataset, and these Japanese reports were initially translated into English using Google Translation, a widely employed machine translation engine. We referenced the recent data argumentation advancement using large language models[13] and engineered prompts via ChatGPT to enhance the quality of the translated document, ensuring it reads smoothly in English. To evaluate the quality of the translated content compared with the original Japanese version, one manual reviewer with research experience in Japanese medication errors carefully reviewed every English synthetic report and comparing it with the original Japanese content. If any discrepancies/errors were found, the reviewer provided the correct English reference based on the written style in which the synthetic data was provided. The Bilingual Evaluation Understudy Score (BLEU) was computed to evaluate the match between the synthetic set and the reference set. **Table 1** displays examples of the original incident reports of medication errors in Japanese, alongside the synthetic and reference incident reports in English.

2.2. Corpus level summary and preprocessing

A free-text summary is provided, which includes information about the number of sentences and unique tokens found across the entire corpus. We follow standard NLP preprocessing methods as described in[14], which involve tokenization, removal of English stopwords, and the utilization of stemming functions available in the Python nltk framework. We also present length distribution and frequency plots for the top 20 tokens. Furthermore, a Word Cloud is used to visually depict the relative importance of term frequency within the corpus.

2.3. Search Methods

Traditional keyword search aims to match specific search query strings within the search corpus. Typically, the order of search strings and case sensitivity do not matter, and these search engines return exact matches or close variations of the user-input query. For comparison, we selected the state-of-the-art sentence-based BERT as a vector-based method with keyword-based BM25, a custom scoring method considering term and document frequencies.

2.3.1. Keyword-based search engine: BM25

Tortman, Puurula and Burgess (2014)[9] evaluated the precision improvement of nine state-of-the-art relevance ranking and document retrieval ranking functions and concluded that the improvements among the BM25-like algorithms are insignificant. It was

Table 1: Examples of original Japanese reports, synthetic incident reports in English, and their manually reviewed reference in English.

Original Incident Reports (in Japanese)	Synthetic incident reports (in English)	Reference translation (in English)
レベミル皮下注を自己注射の患者。8 単位から 4 単位へ減量の指示が出たので口頭での説明をしたが、減量していなかった。	Patients were handling their own Levemir injections under the skin. I was told to cut the dose from 8 units down to 4 units, and I explained this verbally. However, the dose wasn't actually reduced.	The patient had self-administrated Levemir subcutaneous injection. I was told to cut the dose from 8 units down to 4 units, and I explained this verbally. However, the dose wasn't actually reduced.
帝王切開術後 4 日目のクレキサン注射はないと思い込み、指示票を確認しなかった。16 時 40 分、点滴の確認をしていた準夜の看護師が、9 時のクレキサンが施行されずに残っているのを発見した。	I mistakenly assumed there wouldn't be a Clexane injection on the 4th day following the caesarean section and didn't verify the instructions. It was only at 4:40 p.m. when a night nurse, during an IV drip check, noticed that the Clexane scheduled for 9 o'clock had not been administered.	I mistakenly assumed a Clexane injection wouldn't be administrated on the 4th day following the cesarean section and didn't verify the instructions. It was only at 4:40 p.m. when a semi-night nurse, during an IV drip check, noticed that the Clexane scheduled for 9 o'clock had not been administered.
患者 a のベッドサイドにトラゼンタが落ちていたが、患者 a は元々トラゼンタを内服していなかった。同室内でトラゼンタを内服しているのはこの患者 b しかおらず、朝の分を内服できていない可能性が高い。	We discovered Trazenta at the bedside of patient A, even though patient A was not originally prescribed Trazenta. Patient B, who shares the room with patient A, is the only individual in the room taking Trazenta, and it's highly probable that they missed their morning dose.	(no correction made)
注射指示はシプロキサン 300mg/150ml を 100ml 投与の指示だったが、125ml 投与した時点で間違えに気付いた。	The injection instructions called for the administration of 100ml of ciproxan 300mg/150ml, but I realized the mistake after administering 125ml.	The injection instructions called for the administration of 100ml of Ciproxan 300mg/150ml, but I realized the mistake after administering 125ml.

concluded that BM25 has performed at near human level on the standard TREC evaluation collections. Therefore, in our study we use the baseline BM25 (Okapi) to demonstrate the performance of lexical-based search. The algorithm considers term frequency and document length normalization to address the bias from different document lengths across the corpus. For a given query, *q*, the retrieval status value, rsv_q, is expressed as below:

$$rsv_q = \sum_{t \in q} log \frac{N}{df_t} \left(\frac{(k_1 + 1) tf_{td}}{k_1 (1 - b + b \frac{L_d}{L_{avg}}) + tf_{td}} \right)$$

rsv_q equals to the sum of individual term, *t*, scores. The number of documents in the collections is denoted as *N*. df_t is the document frequency, whereas tf_{td} is the number of times term t occurs in document *d*. The length of the document (in terms) and the mean of the document lengths are denoted as L_d and L_{avg}. The Python code implementation of algorithm can be found at[15] and we adopted the default tuning parameter values, *b* and k_1 as 0.75 and 1.5. The input corpus was the original synthetic full text ones without NLP pre-processing.

2.3.2. Vector-based search engine: SBERT

Vector-based search engines are designed to generate numerical representations of text queries using embedders or pre-trained language models. These representations are then indexed within a high-dimensional vector space, and the relevance of results is determined by ranking how closely a query vector aligns with vectors across the corpus. Some of these methods have demonstrated the ability to provide faster and more effective search results, making them potentially valuable for large document corpora within medical contexts.

Traditional BERT construction renders it unsuitable for efficient semantic textual similarity search due to heavy inference computation requirements among sentence pairs. Introduced in 2019, SBERT[11] has proven its ability to significantly decrease the computational effort required to identify the most similar pairs compared to BERT or RoBERTa models. SBERT incorporates a pooling operation into the BERT output. It subsequently employs Siamese and triplet networks to generate sentence embeddings, facilitating the comparison of semantically meaningful embeddings. Sentences with semantic similarity are located closely within the vector space and can be compared using cosine similarity functions. We utilized document-level and sentence-level inputs, creating two scenarios for experiment SBERT.

2.4. Performance Evaluation

Adopted by TREC as a benchmark development method, an ad hoc information retrieval approach and its relevant evaluation metrics[16), which assess a one-time or batch-based retrieval of documents relevant to a query, are used in this study. Similar to most realistic IR scenarios, it is too labor and time-intensive to review all the ground truth labels across the entire corpus for all search queries. Therefore, the evaluation is conducted in a precision-based rather than a recall-oriented manner. We adopted the pooling method that requires the document pool to be manually evaluated. Essentially, reviewers evaluate all the documents in the pool, and documents outside the pool are automatically considered irrelevant. Mean (un-interpolated) average precision (MAP), which refers to the mean precision at seen relevant documents, is adopted to measure the search methods. MAP is denoted as follows:

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(rel = i)$$

where Q_j is the number of relevant documents for query j and N is the number of queries and $P(rel=i)$ is the precision at ith relevant document. MAP is applied to the following semantic-based searches:

- Root cause search: Three queries are designed based on root causes of incidents that occurred.
- Phrase search: Five queries are designed based on a combination of concepts of incident type and drug generic or specific names.
- Characteristic search: Three queries were used to describe the characteristics of medication errors.

As a reference, plain keyword search (which is the mature string matching task) is also conducted across concepts of drug name specificity (5 items) and condition specificity (5 items). We also report the performance using MAP.

In addition to the above categories, we also perform a set of similar sentence searches. Five randomly selected sentences were lightly rephrased by generative AI to serve as the inputs for the search queries. (Note: In a lightly rephrased sentence, the same content is presented using synonyms and/or a different sentence structure. As an example, “I verified the information” has been rephrased as “I double-checked the details”). Since we conduct searches using complete sentences, we capitalize the first letter of each sentence. In this design, there is one relevant document that matches each query. Mean reciprocal rank (MRR) is therefore used to measure performance, denoted as:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}RR_i$$

considering rank position k of the first relevant document, each query we will retrieve a reciprocal rank (RR) $=1/k$. MRR is denoted as the mean RR across multiple queries (n=5). MRR is bounded between 0 and 1.

3. Results

3.1. Text descriptive summary

In this chapter, we provide a descriptive summary of the synthetic incident reports in English at the corpus level. From the synthetic incident report corpus, the total number of sentences is 143, and the number of unique tokens is 617. The average BLEU score, obtained by comparing the reference and synthetic versions, is reported as 0.700, indicating a high-quality match. This ensures that these synthetic English reports closely mimic the context of the real-life hospital incident reports. Figure 1 displays the distribution of report lengths, which range from 25 to 150 words.

After preprocessing, which includes tokenization, removal of English stopwords, and the use of stemming functions provided by nltk framework, the top 3 most frequently occurring terms are “patient,” “medication,” and “day,” as depicted in Figure 3. In addition, Figure 3 displays the relative importance of each word that appeared in the corpus.

3.2. Precision-based Performance

As depicted in Figure 1, for similar sentence search, the reported MRR values are 100% (Rank BM 25), 80% (SBERT-Document), and 80% (SBERT-

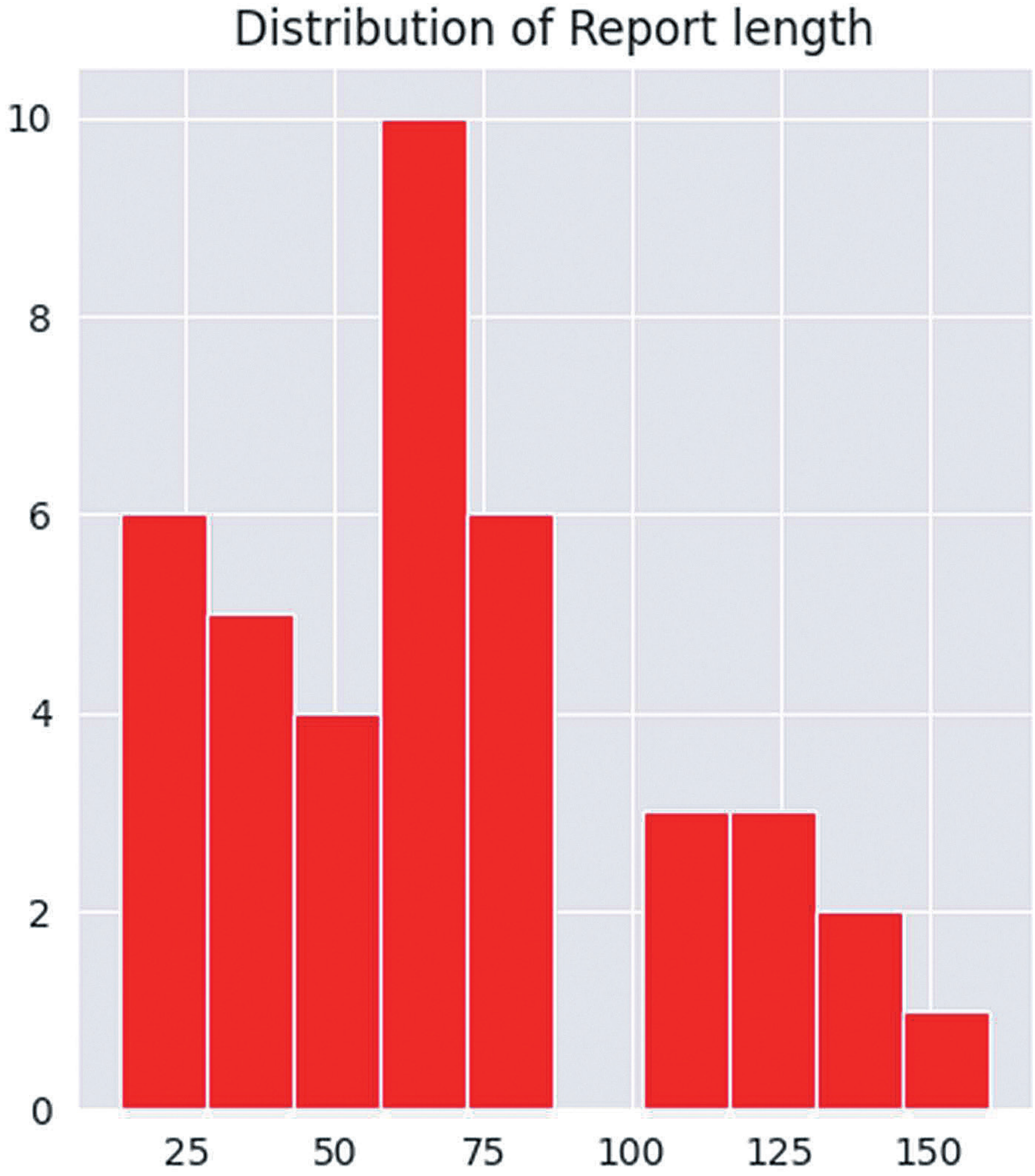

Figure 1: Report Length Distribution. X-axis: number of words. Y-axis; number of reports

Sentence). The semantic-based MAP for Rank BM25 and SBERT-Document and SBERT-Sentence are 73%, 22%, and 2% respectively.

Figure 5 presents the MAP results for different search categories. Under the root cause search, we performed searches using three queries based on the listed contributory factors of medication errors concluded by content analysis[17]. These queries include "interruption or distraction when preparing or administering medication," "unclear communication or orders," and "lack of adequate access to guidelines or unclear organisational routines". The root-cause based MAP for Rank BM25, SBERT (document level), and SBERT (sentence level) are reported as 67%, 40%, and 8%, respectively.

Regarding the phrase search tasks, five queries were designed based on a combination of concepts involving incident type and drug generic or specific names. These queries are: "wrong patient taking Trazenta", "self administration of eye drop", "missed dose of opsumit", "oral medication at wrong time", and "IV

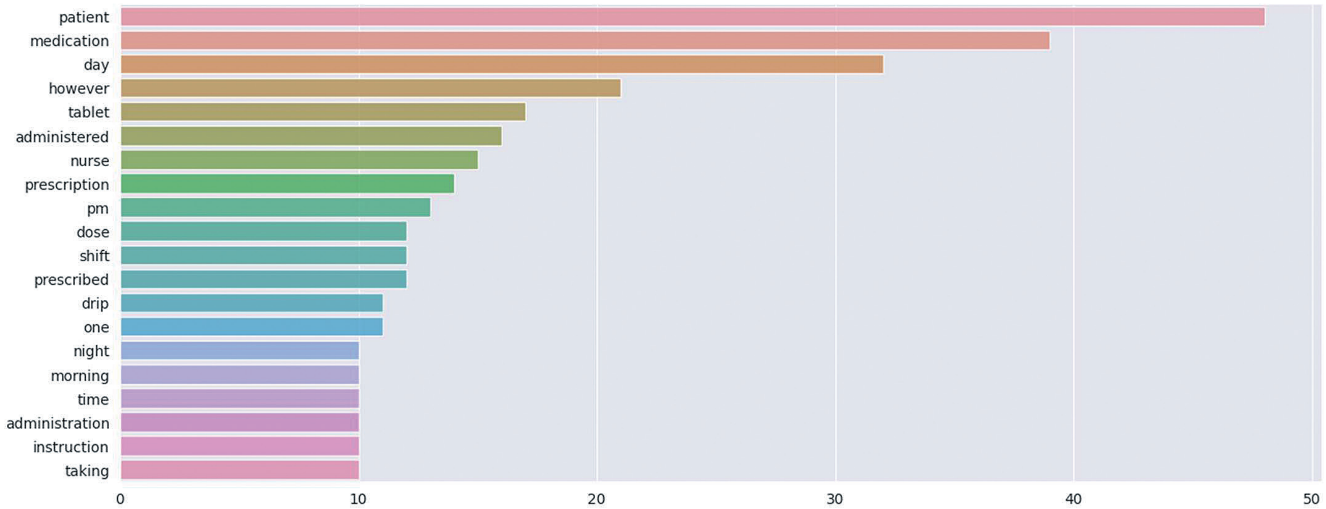

Figure 2: Top 20 Token Frequency

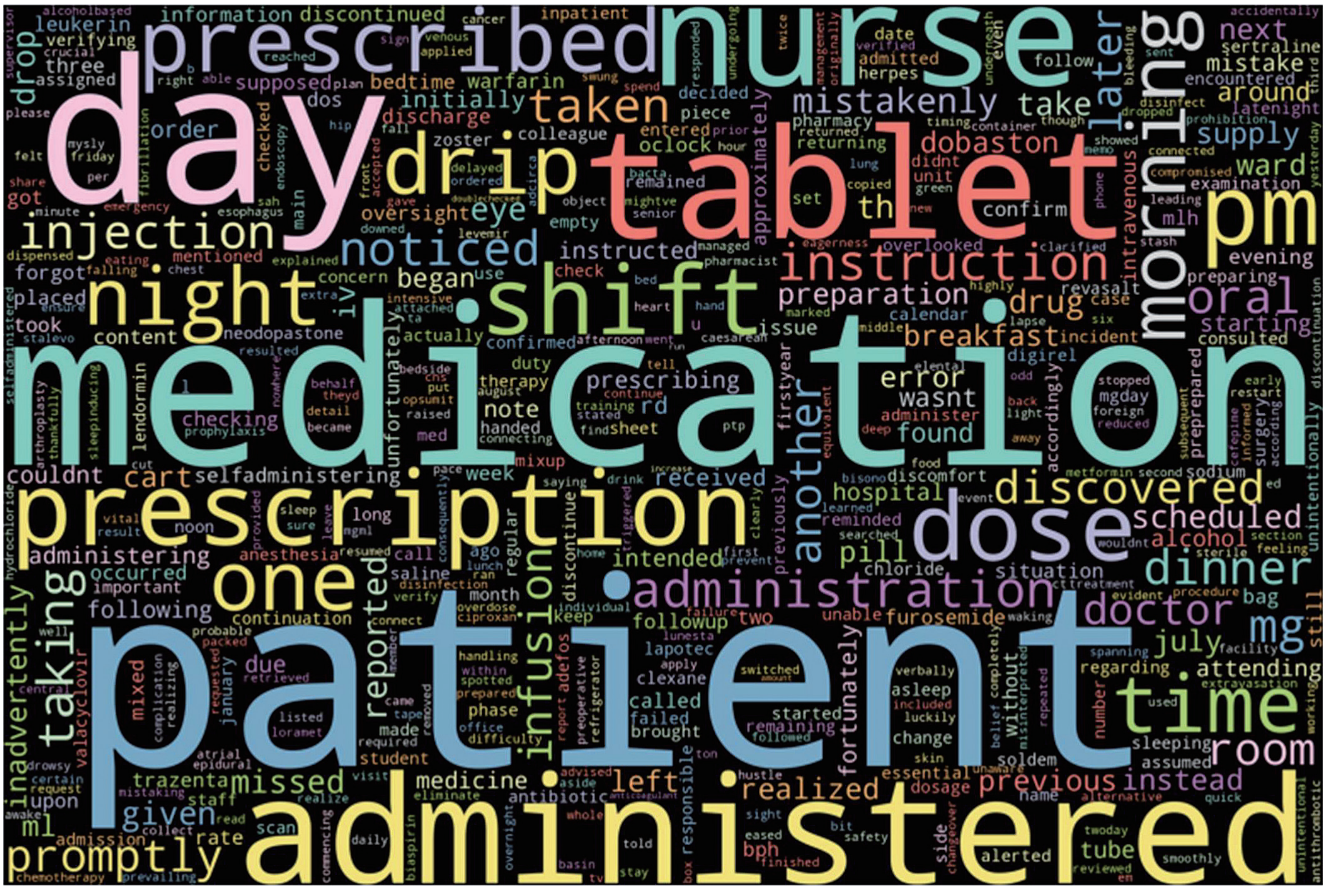

Figure 3: Word Cloud showcasing the importance of term frequency within the corpus

infusion with sodium chloride". The reported MAP values for these tasks are 65%, 20%, and 0%.

Three queries were used to test the characteristic-based search ability, which allows for the description of medication error characteristics. These queries include "drug incidents involving anticoagulant", "misuse of antibiotics", and "look-alike sound-alike". As a reference, plain keyword search (which tests string matching ability) was also conducted, though high performance was not anticipated in this type of search, as matching keyword is a mature process and regarded as a different IR task. We used concepts related to 5 drug names and 5 conditions for this search, and the MAP values are recorded as 75%, 16%, and 18%.

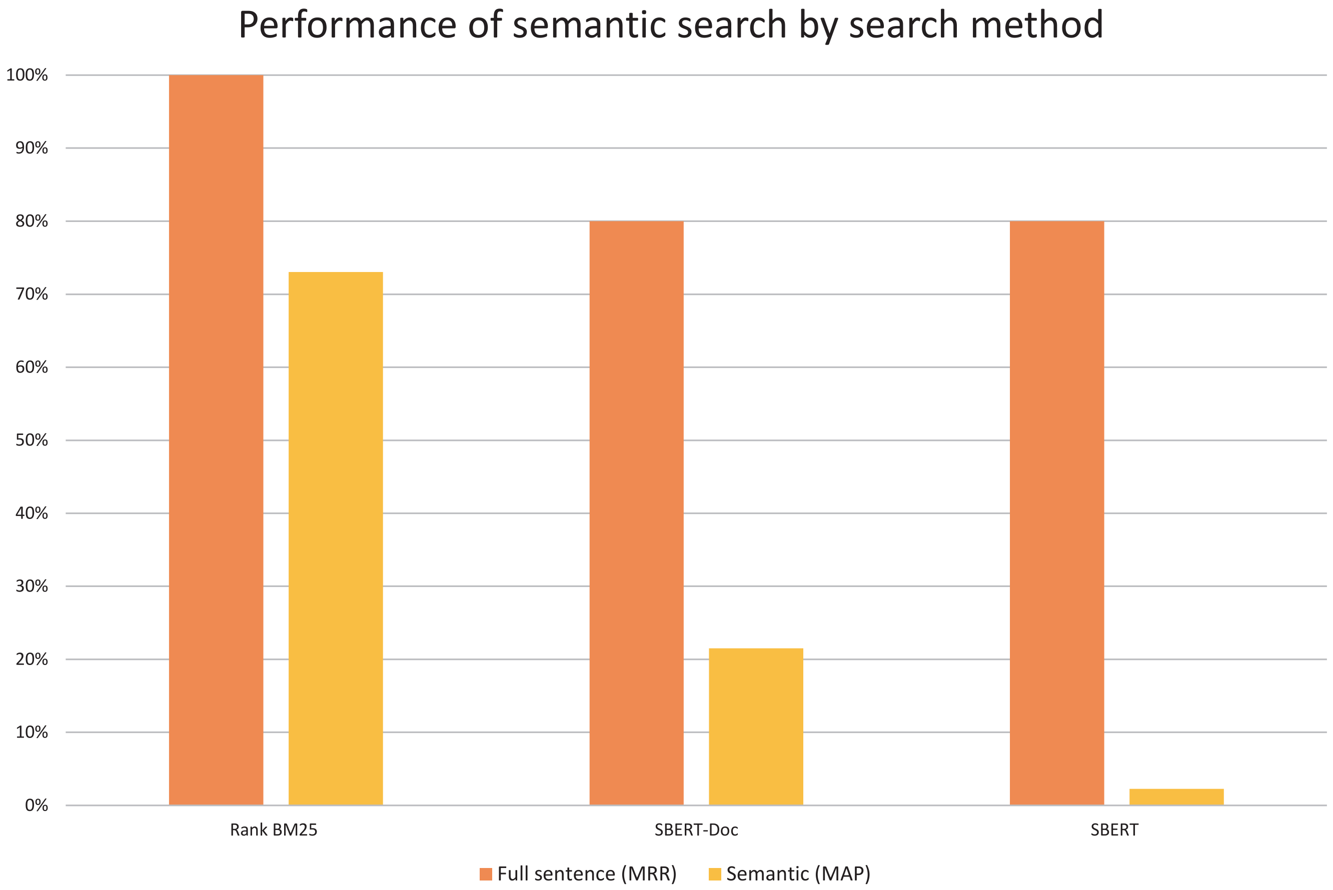

Figure 4: Performance of semantic search by search method

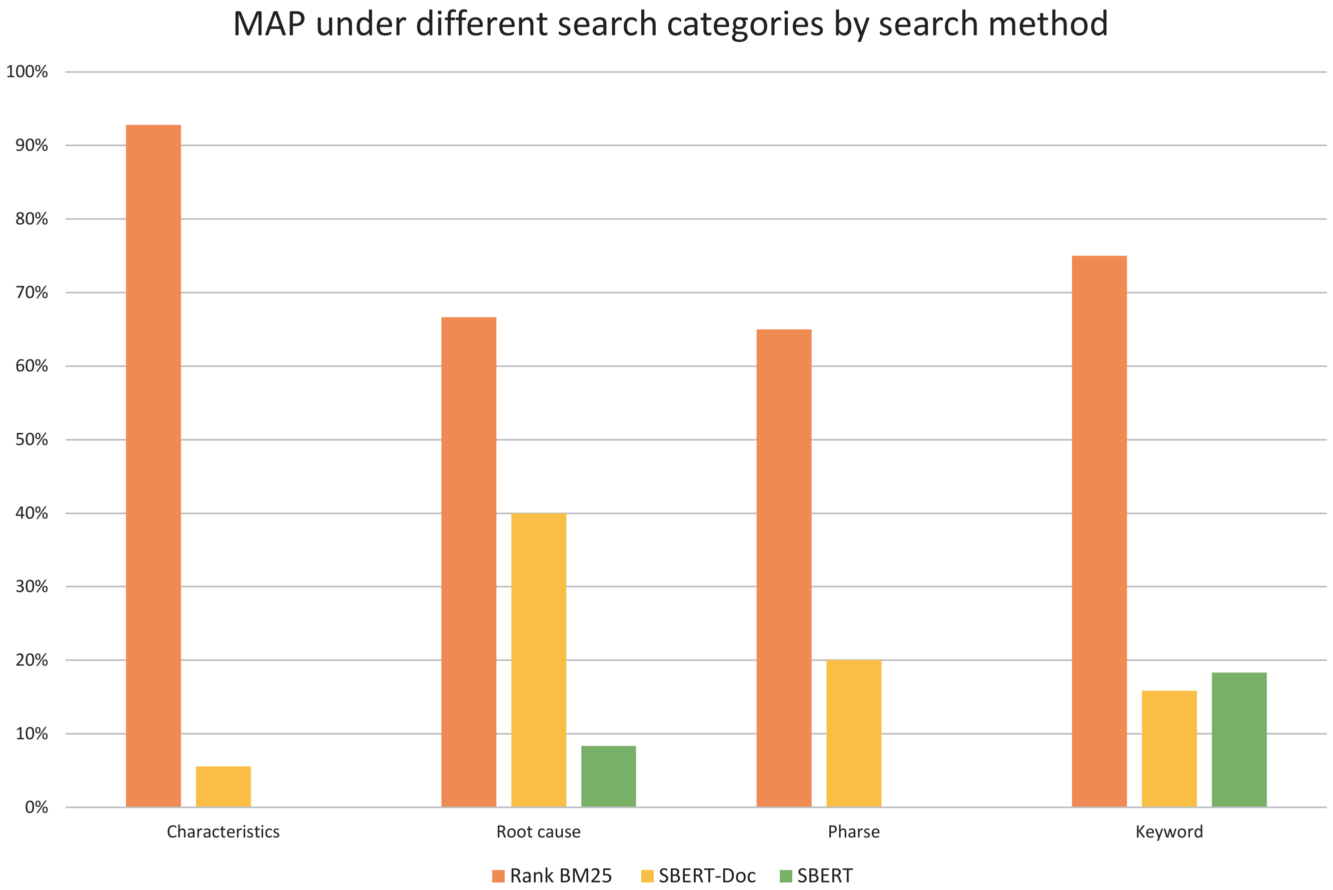

Figure 5: MAP under different search categories by search method

4. Discussion

Lexical search is indeed a fast and efficient method for finding exact matches in large databases. However, it has limitations when it comes to understanding the context of the query and the documents. These ranking algorithms rely on language usage patterns without the ability to comprehend the semantic context of the words used. Semantic search, if implemented effectively, is designed to interpret the meaning of a query and identify relevant results, which can be particularly valuable in medical IR, where many searches carry semantic significance.

Based on the preliminary results presented in this experiment, it appears that the lexical-based search method (Rank BM25, a widely used ranking algorithm) consistently outperforms SBERT (at both document and sentence levels) in incident report search tasks, including root-cause-based search, phrase search, characteristic-based search, and full sentence (rephrased) search. SBERT is considered the existing state-of-the-art method for semantic search in a general context, as indicated in relevant papers[10, 11]. The substandard result demonstrated by SBERT suggests that there are numerous potential directions for further improvement in semantic-based methods to make them practical for clinical incident report retrieval applications.

The current SBERT model is built upon the BERT architecture, which was originally designed for generic contexts. Without undergoing specialized training on clinical text, word embeddings trained on generic corpora represent the meaning of a word based on the context in which they were trained, therefore, they may not comprehensively cover clinical knowledge and medical meanings. To enhance embedding representation, it may be possible to integrate clinical concepts into the model through a fine-tuning process[18, 19]. Currently, there are clinical data pretrained BERT models available that have contributed to advancements in clinical NLP tasks, such as BioBERT[20] etc. SBERT remains an efficient BERT-based training architecture which is particularly suitable for IR tasks in generic context[11]. Researchers may explore how to incorporate an efficient IR method using these clinically-oriented BERT models to enhance clinical IR. Furthermore, a knowledge graph-based search approach should be investigated in the future, especially with recent advancement we can now incorporate medical ontologies into the analysis for incident reports[21].

Currently, the search queries have been grouped into root-cause-based search, phrase-based search, and characteristic-based search, guided by our understanding of how these IR needs related to incidents can be addressed. In the future it would be important to cocreate with frontline incident reporters to design more realistic search queries that would be useful in understanding specific causes and incident concepts, or for conducting root cause analysis.

Furthermore, from a digital tool design perspective, the IR system could adopt multiple methods, including keyword matching, lexical-based, and semantic-based search, to cater to the diverse needs of users within the hospital. Furthermore, clustering and ensemble approaches could potentially be employed to explore groupings and weight different rankings generated by various algorithms, enhancing the robustness of retrieval results. To evaluate the algorithms' robustness across various languages, additional research can involve utilizing the original Japanese reports and integrating Japanese tokenization methods into the algorithms.

There are some limitations to this experimental study and potential biases and issues that require further attention. Firstly, we did not tune the hyperparameters of both search models. Secondly, synthetically generated incident reports were created by generative AI data augmentation methods. The way we engineered the prompts might create different synthetic reports as input, potentially influencing the subsequent search outcomes. Thirdly, the quality of the data was assessed based on its ability to preserve the original Japanese content accurately without introducing discrepancies. We did not specifically analyze errors related to content, tone, nuances, grammar, or parts of speech. In the future, we anticipate comprehensively evaluating the synthetic data generation ability to facilitate the exploration of a graph-based search approach, which requires named entities, relations, and incident types identified from existing Japanese data to be retained in the translation.

5. Conclusion

This study serves as a preliminary examination that assesses the performance of two efficient state-of-the-art IR methods: keyword-based Rank BM 25 and semantic-based SBERT. At present, Rank BM 25 outperforms all search queries, including root-cause-based search, phrase search, characteristic-based search, and rephrased sentence search, using our synthetically generated incident reports (originally sourced from public materials without any patient identifiers). We anticipate future enhancements in semantic-based search methods by scaling models with clinically-oriented resources and incorporating named entities as knowledge graphs in an efficient manner.

References

1) Project to Collect Medical Near-Miss/Adverse Event Information. Project details and how to participate [Iryō jiko jōhō shūshū-tō jigyō jigyō no naiyō to sanka hōhō]. The Japan Council for Quality Health Care (JQ); 2022.

2) Franklin BD, Panesar SS, Vincent C, Donaldson LJ. Identifying systems failures in the pathway to a catastrophic event: an analysis of national incident report data relating to vinca alkaloids. BMJ quality & safety. 2014:bmjqs-2013-002572.

3) Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. Journal of the American Medical Informatics Association. 2019;26(12):1600-8.

4) Liu J, Wong ZSY, So HY, et al. Evaluating resampling methods and structured features to improve fall incident report identification by the severity level. Journal of the American Medical Informatics Association. 2021;28(8):1756-64.

5) Wong ZS, So HY, Kwok BS, et al. Medication-rights detection using incident reports: A natural language processing and deep neural network approach. Health informatics journal. 2019:1460458219889798.

6) Minimal information model for patient safety incident reporting and learning systems: user guide. WHO; 2016. Contract No.: 14 Feb.

7) WHO. Patient safety incident reporting and learning systems: technical report and guidance. 2020.

8) Full-Text Search Functions: Oracle; 2023 [Available from: https://dev.mysql.com/doc/refman/8.0/en/fulltext-search.html.

9) Trotman A, Puurula A, Burgess B. Improvements to BM25 and Language Models Examined. Proceedings of the 19th Australasian Document Computing Symposium. 2014.

10) Devlin J, Chang MW, Lee K, et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

11) Reimers N, Gurevych I, editors. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks2019 November; Hong Kong, China: Association for Computational Linguistics.

12) Denecke K. Concept-Based Retrieval from Critical Incident Reports. Studies in health technology and informatics. 2017;236:1-7.

13) Ubani S, Polat SO, Nielsen R. ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT. 2023.

14) Wong ZS. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. Health informatics journal. 2014 (Nov 11):1-17.

15) Brown D. Rank-BM25: A Collection of BM25 Algorithms in Python: Zenodo; 2020 [Available from: https://doi.org/10.5281/zenodo.4520057.

16) Teufel S. An Overview of Evaluation Methods in TREC Ad Hoc Information Retrieval and TREC Question Answering. 372007. p. 163-86.

17) jörkstén KS, Bergqvist M, Andersén-Karlsson E, et al. Medication errors as malpractice-a qualitative content analysis of 585 medication errors by nurses in Sweden. BMC Health Serv Res. 2016;16(1):431.

18) Khattak FK, Jeblee S, Pou-Prom C, et al. A survey of word embeddings for clinical text. J Biomed Inform. 2019;100s:100057.

19) Ji Z, Wei Q, Xu H. BERT-based Ranking for Biomedical Entity Normalization. AMIA Jt Summits Transl Sci Proc. 2020;2020:269-77.

20) Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (Oxford, England). 2020;36(4):1234-40.

21) Zhang HK, Sasano R, Takeda K, et al. Development of a Medical Incident Report Corpus with Intention and Factuality Annotation. LERC 2020; 2020; Marseille.

Acknowledgement

The author extends gratitude to Ms. Miwa Kanematsu for her review of the appropriateness of

the English synthetic reports in comparison to the original Japanese reports.