RUC Team at SemEval-2025 Task 5: Fast Automated Subject Indexing via Similar Records Matching and Related Subject Ranking

Tian Xia[1], Xin Yang[1], Wenjing Wu[1], Yueheng Xiu[1], Xin Zhang[2], Jinyu Li[1]
Tong Gao[1], Zhuoxi Tan[2], Rundong Hu[1], Tao Chen[2], Junzhi Jia[1]*

[1]School of Information Resource Management, Renmin University of China

[2]School of Information Management, Sun Yat-sen University

Abstract

This paper presents MaRSI, an automatic subject indexing method designed to address the limitations of traditional manual indexing and emerging GenAI technologies. Focusing on improving indexing accuracy in cross-lingual contexts and balancing efficiency and accuracy in large-scale datasets, MaRSI mimics human reference learning behavior by constructing semantic indexes from pre-indexed documents. It calculates similarity to retrieve relevant references, merges, and reorders their subjects to generate index results. Experiments demonstrate that MaRSI outperforms supervised fine-tuning of LLMs on the same dataset, offering advantages in speed, effectiveness, and interpretability.

1 Introduction

With the increasing diversification of academic disciplines and the continuous influx of research outputs, the volume of documents in libraries has grown rapidly, raising the demand for efficient and accurate subject indexing techniques(Zhang et al.). Libraries use controlled vocabularies such as thesaurus and name authorities to assign subject terms for kinds of documents. However, manual indexing has high cost and low processing efficiency. With the use of Artificial Intelligence(AI), the automated subject indexing technology improves indexing efficiency. By virtue of its strong semantic processing and generalization capacity, generative AI, represented by large language models(LLMs), provides an automated path for multi-lingual subject indexing in large-scale data.

In this context, the TIB Leibniz Information Centre for Science and Technology launched the LLMs4Subjects shared task at SemEval-2025, one of the ACL 2025 Semantic Evaluation challenges(D'Souza et al., 2025). The task encourages exploring LLMs-based automated subject indexing through fine-tuning, retrieval-augmented generation (RAG), chain-of-thought prompting, and few-shot learning. This initiative serves as the motivation for our study.

Through comparative experiments with multiple AI indexing approaches English and German documents, we propose a deep learning-based embedded indexing solution that enhances automatic indexing by matching similar records and ranking related subjects. The study focuses on two key issues: 1) How to improve indexing accuracy in cross-lingual environments? 2) How to maximize accuracy while ensuring data processing efficiency in large-scale datasets?

2 Related Work

Subject indexing uses controlled vocabularies to assign subject terms to bibliographic resources, enhancing knowledge organization and retrieval. This intellectual process remains predominantly dependent on human expertise for accurate conceptual analysis. Since the 1970s, the library science and information retrieval communities have systematically investigated automated subject indexing, which primarily employ: multi-label classification, controlled term extraction and machine-assisted subject assignment. The primary methods currently in use include:

(1) Rule-based matching uses predefined syntactic, semantic, and domain-specific rules to align text terms with controlled vocabulary(Fernandez-Llimos et al., 2024), aiding subject suggestion. Methods like Maui(Medelyan, 2009) and STM(Gusfield, 1997) filter potential matches, but this approach struggles with diverse terminology and partial mismatches, leading to under-indexing.

(2) Statistical analysis-based indexing applies word frequency, associations, and co-occurrence patterns to categorize large-scale text data, facil-

*Corresponding author: JIA JUNZHI, E-mail: Junzhij@163.com, ORCID: 0000-0003-1486-673X

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2437–2442
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

itating rapid retrieval in news, academia, and e-commerce(Janssens et al., 2009). Bibliometric analysis enhances visualization, reducing manual effort and rule dependence. However, lacking semantic understanding, these methods struggle to capture deep meaning.

(3) Machine learning-based automatic indexing leverages algorithms trained on annotated datasets to classify and index subjects across domains. Clustering and text mapping methods effectively classify journal articles, while algorithmic advancements have improved accuracy, driving broader adoption.

(4) Semantic computation with vector embeddings, as seen in Word2Vec and BERT(Sharma and Kumar, 2023), improves retrieval, clustering, and classification by generating semantic indexes. By mapping document features into vector space, these models enable precise indexing beyond keyword matching but require large corpora for high-quality representation(Nentidis et al., 2023).

(5) Automatic indexing using LLMs benefits from their advanced semantic understanding and strong performance in tasks like multi-label classification and keyword extraction. With prompt engineering, models like ChatGPT can infer relevant labels from few-shot examples. However, directly applying LLMs to document indexing(Kasprzik, 2024), which adheres to specific classifications and controlled vocabularies, often results in suboptimal performance.

In general, each automatic subject indexing method has its own strengths and limitations, with no universally perfect solution. Therefore, the study proposes a comparative experiment across three representative approaches-semantic embedding, supervised fine-tuning (SFT), and retrieval-augmented generation (RAG) to explore automatic such solutions from both qualitative and quantitative perspectives

3 Methods and Design

3.1 Research Question

This study frames subject indexing as the process of finding a specific function or model, f, where the title t_r and abstract a_r of a document r are input, and the model outputs the expected subject list, i.e.

$$f(t_r, a_r) = subjects(r) \tag{1}$$

The function f can internalize relevant knowledge by learning from existing data, typically through supervised fine-tuning (SFT) of LLMs.Using a labeled corpus, the model learns to map titles and abstracts to subject lists. However, due to the large number of subjects in the LLMs4Subjects task compared to the available training samples, SFT proves ineffective, as confirmed through experiments.

To address this, we propose MaRSI (Match and Rank Subject Indexing), a fast indexing method inspired by human imitation learning. MaRSI utilizes expert-generated indexing results to automatically index new documents. For a given document r,, we first identify pre-indexed samples similar to its content. The indexing results of these similar samples are treated as a candidate set, which is then ranked, and the top k, results are selected as the subject terms.

Let the indexed document set be D $= \{d_1, d_2, \ldots, d_n\}$, where each document d_i has an associated set of indexed subjects S$_i$ $= \{s_{i1}, s_{i2}, \ldots, s_{im_i}\}$, with m_i representing the number of subjects for the document d_i. The function sim(d, r) measures the similarity between document d and the target document r for indexing. By calculating the similarity between all documents in d and r, a similarity ranking is obtained, and the top k documents are selected as reference results, forming the set C $= \{c_1, c_2, \ldots, c_k\}$, where $c_i \in$ D, and sim$(c_i, r) \geq$ sim$(c_j, r) \geq$ sim(d, r) (for any $i < j$, $d \in$ D $\setminus$ C).

Next, the subjects from the reference set C are merged into a set T $= \{t_1, t_2, \ldots, t_n\}$, containing m distinct subjects. The subjects in T are then ordered using a sorting algorithm, and the sorted result is used as the final prediction, defined as:

$$f(t_r, a_r) = rank(T|r) \tag{2}$$

In this process, the similarity function sim(d, r) and the ranking function rank$(d|r)$ are critical to the prediction outcome. The following section will outline the specific approaches for these functions.

3.2 Main Framework of MaRSI

The overall workflow is illustrated in Figure 1. After discovering similar documents, two processing approaches are possible: direct sorting and output through the subject ranking module, or enhancing retrieval using LLMs, where the indexing results of similar documents serve as reference inputs for few-shot learning, as shown in the gray background of the figure. After manual analysis, RAG methods did not yield better results, so the first approach was adopted in practice. Note that LLM filtering (gray area) not used in final system.

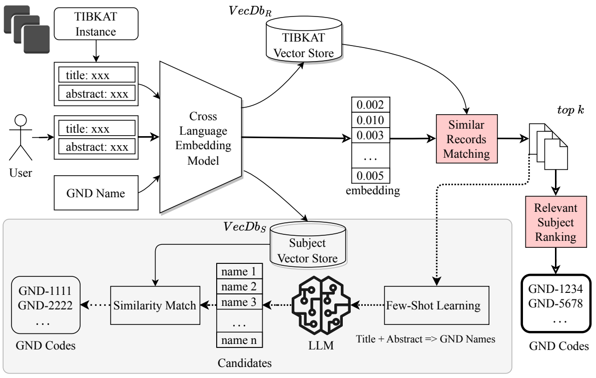

Figure 1: MaRSI Workflow.

(1) Vector Database Construction. We first embed the indexed documents records and subject sets of TIBKAT to construct the documents vector database $VecDb_R$ and the subject vector database $VecDb_S$. Given that the TIBKAT dataset contains both English and German data, the multilingual embedding model Arctic-Embed 2.0(Yu et al., 2024) was used. Based on the Transformer architecture and pretrained on a large multilingual corpus, this model is specifically optimized for semantic representation in both German and English, enabling it to efficiently capture semantic features and perform well in cross-lingual tasks.

When computing embedding vectors, the title and abstract of each document are concatenated as follows:

```
"""
title: {title}
abstract: {abstract}
"""
```

The subject calculation template is as follows:

```
"""
Subject: {name}
Related subjects: {related_subjects}
Classification Name: {classification_name}
"""
```

By iterating through TIBKAT's training set and subject list, two vector databases are generated. The TIBKAT document vector database is used to compute similar documents, while the subject vector database maps output subject terms from large language models to the most similar subject codes.

(2) Processing Workflow. Once the vector database is constructed, for a given document r, the following steps are performed: the title and abstract of document r are concatenated and input into a cross-lingual embedding model to generate its semantic vector. Using a vector-based nearest neighbor search algorithm, similar documents are retrieved from $VecDb_R$. The indexed results of these documents are merged to form a candidate set of subjects, which are then ranked to obtain the final results.

3.3 Similar Document Retrieval

To identify similar documents, this study employs an inner product-based similarity search algorithm. Given two documents d1 and d2, their semantic vectors d_1 and d_2, are derived from the embedding model, and their similarity is calculated as follows:

$$sim(\mathbf{d}_1, \mathbf{d}_2) = \sum_{i=1}^{N} v_i(\mathbf{d}_1) \cdot v_i(\mathbf{d}_2). \qquad (3)$$

Where N is the vector length (1024 in this study), and $v_i(d)$ epresents the i-th component of the semantic vector for document d.

For efficient computation, the semantic vectors of indexed documents are stored in a Faiss index using the IndexFlatIP structure. This structure stores vectors in a flat data structure, calculates the inner product between the query vector and all indexed vectors, sorts them based on the inner product values, and returns the Top k most similar vectors, ensuring globally optimal results.

3.4 Candidate Subjects Ranking

The subject ranking is based on three assumptions: (1)The more similar a document d is to the target document r, the more important the subjects in d are; (2) A subject s in document d is more important the earlier it appears in the document's list of subjects; (3) A subject's name or a candidate name appearing in the title or abstract of d increases its importance. Based on these assumptions, we propose the following ranking algorithm for candidate subjects.

Algorithm 1: Candidate Subject Ranking Algorithm

```
Input: $C, title, abstract$
scores ← defaultdict$(0.0)$
for $i \leftarrow 1$ to len$(C)$ do
    $d \leftarrow C[i-1]$
    for $j \leftarrow 1$ to len$(d$.gnd_codes$)$ do
        score $\leftarrow 1.0 + \frac{1.0}{j}$
        if $i \leq$ PARAM1 then
            names ← $d$.alt_names $\cup \{d$.name$\}$
            for name in names do
                if name $\in$ title then
                    score ← score + PARAM2
                end if
                if name $\in$ abstract then
                    score ← score + PARAM3
                end if
            end for
        end if
        value $\leftarrow \frac{\text{len}(C)}{i}$
        score ← score × $(1 + \log(\text{value}))$
        scores$[d$.gnd_codes$[j-1]]$ ← score
    end for
end for=0
```

In Algorithm 1, C is the set of similar documents to the target document r, with title and abstract representing $r's$ title and abstract. The algorithm iterates over each related document and its subjects, accumulating scores in the scores dictionary. After sorting the scores, the ranked list of subjects is returned as the indexing result. Three hyperparameters are used in the algorithm: PARAM1 determines how many top documents are weighted for subject occurrence, defaulting to 5; PARAM2 and PARAM3 provide additional scores when subject names appear in the title and abstract of the target document, set to 2 and 1, respectively, in the experiment.

4 Experiment and Results

4.1 Datasets and Evaluation Methods

The dataset used in this study consists of three parts: the Train, Dev, and Test sets. For the full subject indexing task in TIBKAT, the training, validation and test sets contain 163,874, 27,332, and 55,972 items, respectively. For the core subject indexing task, the sets contain 83,804, 13,960, and 12,348 items, respectively. The datasets include five types of literature: articles, books, conference papers, reports, and theses-written in German or English.

Evaluation is carried out using three metrics: Recall, Precision, and the harmonic mean of both (F1 Score). Quantitative evaluation: Indexing results and average recall are calculated every five output subject terms. Qualitative evaluation: Documents are indexed by subjects, and random sampling is performed for expert evaluation to assess the indexing quality across different subjects.

4.2 Indexing Methods Compared

For the cross-lingual subject indexing task, three approaches-Supervised Fine-Tuning (SFT), Retrieval-Augmented Generation (RAG), and Vector Semantic Embedding-were compared, as follows:

LoRA-based Supervised Fine-Tuning with default parameters: Using the LlaMa 8b model, the training datasets for both tasks were combined for fine-tuning. Retrieval-Augmented Generation (RAG): The Chat GLM 4 (130b) model, known for strong text comprehension, knowledge reasoning, and multilingual support, was used to build an external knowledge base from the GND vocabulary and the training set's indexed documents. MaRSI: The final method used, based on document similarity matching and relevant subjects ranking, termed MaRSI.

4.3 Results and Analysis

(1) Results and Analysis on Dev set. Three approaches were used to generate about 50 subject terms for each document in the development set (Dev). Initial random sampling checks revealed that, due to factors such as model performance and external knowledge base construction, the RAG indexing approach performed poorly. Therefore, only two approaches-LLMs fine-tuning (SFT) and semantic embedding (MaRSI)-were compared on the merged validation set. The results are shown below:

As shown in Table 1, MaRSI consistently outperforms SFT in the top 30 results, with a significant decline in SFT's performance from the 10th result onward, possibly due to the factors mentioned earlier. In addition to better indexing performance, MaRSI follows a human-like processing approach, similar to manual subject indexing, where relevant documents are matched to the target subjects through semantic computation. Furthermore, MaRSI is a fast and scalable method, well-suited for large document datasets.

(2) Results and Analysis on Test Set. The test set

P@K	SFT	MaRSI
P_1	38.03%	52.43%
P_2	21.10%	38.65%
P_3	14.09%	30.50%
P_4	10.57%	25.00%
P_5	8.46%	21.25%
P_6	7.05%	18.69%
P_7	6.04%	16.28%
P_8	5.23%	15.12%
P_9	4.67%	13.86%
P_10	4.23%	12.78%
Avg_MAP	20.90%	41.19%

Table 1: SFT & MaRSI Dev Comparison (Top 10).

Task	**@5**			**@10**		
	P	R	F1	P	R	F1
Core	24.94%	47.51%	32.71%	15.81%	57.49%	24.80%
All	22.99%	43.81%	30.15%	44.21%	52.20%	22.41%
Task	**@15**			**@20**		
	P	R	F1	P	R	F1
Core	11.70%	62.29%	19.70%	9.34%	65.39%	16.35%
All	10.43%	56.12%	17.59%	8.25%	58.41%	14.45%

Table 2: Quantitative Analysis of MaRSI Indexing.

consists of two parts: Core and All, corresponding to two distinct indexing tasks. For each part, the MaRSI method outputs 50 subjects per document. Quantitatively, the core subject indexing achieved an average recall rate of 65.68%, ranking 1st in the task, while the full subject indexing had an average recall rate of 58.56%, ranking 3rd. Generally, precision and F1 scores decline as the number of subjects increases (Table 2). The overall low accuracy may be attributed to the fact that the number of subjects in the Train and Dev sets was much smaller than the entire GND vocabulary, making it difficult for some unused subjects to be matched through semantic embedding.

Qualitatively, LLMs4Subjects task incorporated expert analysis. This explanation references Case 1 results(D'Souza et al., 2025), which assess only the technical correctness of indexing without strict subject relevance requirements. MaRSI achieved an average precision of 52.13%, ranking first in the task. The precision of subject assignmentvaried significantly across disciplines (Table 3). Architecture, social sciences and economics showed the highest indexing accuracy (approximately 78%, 78%, and 74% respectively), respectively, demonstrating the method's robustsemantic capture capability inthese fields, likely due to the semantic richness and accuracy of subject terms of them. Semantic embedding methods also performed well in engineering, literature studies and physics, where subject terms are typically more specialized and precise. However, the results were more moderate inmaterial science , mathematics, computer science and chemistry , and poorer inelectrical engineering, linguistics, traffic engineering medicine, and history. This can be attributed to the semantic complexity and ambiguity in these fields, where subject terms tend to be more nuanced and multifaceted, and the subject terms are often highly specialized and complex, making semantic similarity matching less effective.

CLS	**P_5**	**P_10**	**P_15**	**P_20**	**MAP**
Architecture	94.29%	75.71%	73.33%	72.14%	78.87%
Social Sciences	88.00%	79.00%	74.67%	71.00%	78.17%
Economics	84.00%	79.00%	68.67%	64.50%	74.04%
Engineering	93.33%	63.33%	52.22%	43.33%	63.05%
Literature Studies	77.50%	66.25%	55.00%	48.75%	61.88%
Physics	72.00%	65.00%	56.67%	50.00%	60.92%
Material Science	66.00%	56.00%	54.00%	51.50%	56.88%
Mathematics	65.71%	57.14%	52.38%	48.57%	55.95%
Computer Science	77.14%	55.71%	47.62%	40.00%	55.12%
Chemistry	62.86%	54.29%	47.62%	45.71%	52.62%
Electrical Engineering	70.00%	55.00%	44.67%	38.50%	52.04%
History	60.00%	53.00%	44.00%	43.00%	50.00%
Linguistics	56.00%	48.00%	39.33%	32.00%	43.83%
Traffic Engineering	34.00%	36.00%	39.33%	45.00%	38.58%

Table 3: Subject Indexing Results for Different Classifications.

5 Conclusion and Future Research

This study compares three automated subject indexing methods: RAG, SFT, and MaRSI. It finds that MaRSI, which emulates the cognitive patterns of human indexers , is an efficient and high-quality method. Its performance improves with the richness and diversity of subject terms in the training set. However, indexing results exhibit clear disciplinary variations due to different semantic characteristics across fields. These methods are not mutually exclusive, and future work will explore the synergistic use of semantic embedding, LLMs, and rule-based reasoning. By adjusting fusion output weights according to disciplinary semantic traits, we aim to enhance the accuracy, efficiency, and consistency of automated subject indexing in libraries.

6 Acknowledgements

This research is supported by the National Social Science Fund of China (No.22BTQ068).

References

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. Semeval-2025 task 5: Llms4subjects – llm-based automated subject tagging for a national technical library's open-access catalog.

Fernando Fernandez-Llimos, Luciana G. Negrão, Christine Bond, and Derek Stewart. 2024. Influence of automated indexing in medical subject headings (mesh) selection for pharmacy practice journals. *Research in Social and Administrative Pharmacy*, 20(9):911–917.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences*, volume 2025. Cambridge University Press.

Frizo Janssens, Lin Zhang, Bart De Moor, and Wolfgang Glänzel. 2009. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, 45(6):683–702.

Anna Kasprzik. 2024. The automation of subject indexing at zbw and the role of metadata in times of large language models. *Procedia Computer Science*, 249:160–166.

Olena Medelyan. 2009. *Human-competitive automatic topic indexing*. Ph.D. thesis, The University of Waikato.

Anastasios Nentidis, Thomas Chatzopoulos, Anastasia Krithara, Grigorios Tsoumakas, and Georgios Paliouras. 2023. Large-scale investigation of weakly-supervised deep learning for the fine-grained semantic indexing of biomedical literature. *Journal of Biomedical Informatics*, 146:104499.

Anil Sharma and Suresh Kumar. 2023. Machine learning and ontology-based novel semantic document indexing for information retrieval. *Computers Industrial Engineering*, 176:108940.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise.

Shiwei Zhang, Ming-Lun Wu, and Xiuzhen Zhang. Utilising a large language model to annotate subject metadata: A case study in an australian national research data catalogue.