Vol 5, Issue 4, April 2025 E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

Automatic Generation of Concept Maps from Academic Texts Using the Concept Map Miner Framework

Bramhankar Smita Gulabrao [1], **Dr. Nidhi Mishra** [2]

[1, 2] Department of Computer Science, Kalinga University, Naya Raipur, Chhattisgarh, India.

ABSTRACT

The Concept Map Miner (CMM) approach discussed in this article presents a prototype system for the automated creation of concept maps from unstructured academic texts. The framework's goal is to improve how we represent information by finding conceptual patterns in English-language academic publications and putting them into organized, semantically coherent hierarchies. The CMM system can find important propositions, which are triples of concept–relation–concept, and turn them into meaningful visual concept structures by using linguistic pattern recognition, syntactic analysis, and domain-sensitive filtering. In one experiment, a comprehensive concept map was made from all the identified propositions. In the other, summary maps were made using a ranking and filtering strategy to test the prototype. We compared the maps that were developed automatically to those prepared by 10 domain experts to assess how well they worked. The concept analysis had a recall of 0.67 and a precision of 0.47, while the link identification had a recall of 0.44 and a precision of 0.29.

Keywords: *Automatic, Text, Domain, Relationship, Concept.*

I. INTRODUCTION

The automatic creation of concept maps using the Concept Map Miner (CMM) framework is a big step forward in knowledge engineering, text mining, educational technology, and intelligent information processing. This is especially true now that digital information is growing quickly and people need structured knowledge representations to understand it better. Novak and Gowin first came up with concept maps, which are visual tools that show information in terms of ideas and the connections between them. They are great cognitive aids for learning, summarizing, making decisions, organizing a curriculum, and transferring knowledge. Making concept maps has always been a manual, time-consuming task that involves knowledge of the subject, the ability to analyze information, and a lot of time.

As academic, scientific, and technological literature continues to develop at an exponential pace, the demand for automated, scalable, and accurate techniques for idea map generation has grown more critical. It is no longer possible to make concept maps by hand for huge, complicated, and constantly changing material, particularly in fields like health, engineering, law, social sciences, and information technology. This problem has led to the creation of computational frameworks like the Concept Map Miner (CMM), which try to automatically pull out, organize, and show conceptual structures from text that isn't organized. The CMM framework uses natural language processing (NLP), domain knowledge resources, corpus-based linguistic analysis, machine learning techniques, semantic parsing, and rule-based reasoning to find important ideas and connections in academic literature. Its main strength is that it can turn unstructured text into hierarchical, well-structured conceptual pictures that people can comprehend without a lot of manual work.

The CMM framework's main goal is to automate two important tasks: finding relationships and extracting concepts. Identifying important words, multiword expressions, and phrases that are related to the domain that together make up the essential knowledge units of a document is what concept extraction is all about.

This approach is complicated since terminology that are specialized to a field may not always be found in generic dictionaries, may rely on the context, and may have quite different forms in other languages. Medical domain writings, for instance, include particular words like "angiogenesis," "synaptic pruning," and "cerebrovascular regulation," which need to be processed in a way that is specific to the field. Engineering publications also include complex technical terms and algorithmic jargon that may not work the same way as everyday English.

The CMM framework uses a number of tactics to deal with these problems. These include grammar-based rules, depth-first search for finding syntactic patterns, domain-specific corpus mining, frequency distribution analysis, and semantic similarity metrics to find and filter candidate ideas. The framework also includes techniques like anaphora resolution, which makes sure that pronouns and other referring phrases are appropriately linked to their antecedents. This makes it easier to identify concepts. After extracting ideas, the next big job is to figure out how these concepts are related to each other. The CMM framework employs syntactic dependency parsing, semantic role labeling, co-occurrence clustering, and relation pattern templates to find meaningful links like "causes," "leads to," "is a type of," "is part of," "affects," "depends on," or "results in." These linkages make the linking sentences that turn separate ideas into organized networks of knowledge.

One significant part of the CMM framework is that it uses domain knowledge to make things more accurate and relevant. Resources for domain knowledge, such as ontologies, lexicons, thesauri, taxonomies, and specialized domain corpora, are very important for improving extracted ideas, clearing up terminology, and checking the correctness of semantic linkages. Domain corpora, especially, include a lot of unique language patterns and co-occurrence structures that are only seen in certain disciplines. They assist the CMM framework tell the difference between generic and domain-specific terms, figure out the contextual meanings of polysemous words, and find multiword units that are important to the domain. The presence of "blood" and "vessels" together in a medical corpus, for example, shows that these are two separate generic terms and not a single domain-specific idea.

The term "neural network pruning" in machine learning literature denotes a specific conceptual process that ordinary corpora fail to reflect well. So, by using domain knowledge, the CMM framework can generate concept maps that are not only structurally correct but also true to the meaning of the text's domain. Furthermore, domain knowledge improves the CMM's capacity to identify hierarchical relationships, including superordinate and subordinate structures, which are crucial for developing idea maps that demonstrate expert-level comprehension.

The CMM framework's ability to automatically create concept maps is especially useful in schools, where concept maps are great for helping students learn, organizing their thoughts, and becoming more conscious of their own thinking. Automated concept maps let instructors rapidly summarize long academic texts, find the main learning goals, and create lesson plans that fit with the aims of the curriculum. Students learn better when they can see complicated ideas, especially in subjects like biology, chemistry, data science, mathematics, or environmental science that have a lot of complicated connections between concepts. The CMM framework helps students look at how topics are structured, follow conceptual paths, and see how ideas build on each other.

Automatic concept map development may also help adaptive learning systems and smart tutoring environments by letting them create unique learning materials based on how well each student does and how well they comprehend the content. In research settings, the CMM framework is a very useful tool for bringing together different pieces of information. It helps researchers look at large amounts of literature, find gaps in their knowledge, and see how study areas have changed over time. For example, in systematic literature reviews, concept maps that are automatically generated from research papers might show theme clusters, new trends, and connections between important ideas.

Automated concept map production is useful for more than just education and research. It may also be used in decision support systems, knowledge management, corporate training, and massive digital libraries. Concept maps may help organizations that have a lot of text data, including hospitals, law firms, government agencies, and scientific research organizations, organize their information, sort their documents, summarize their expertise, and improve their semantic search. In the realm of artificial intelligence and machine learning, concept maps produced by the CMM framework can function as interpretable representations that augment data-driven models, thereby improving transparency, explainability, and reliability. As AI systems get more complicated, the demand for explainable AI (XAI) has grown. Concept maps are a way to make machine thinking easier to grasp for people.

The automated creation of idea maps has several problems that the CMM framework tries to fix, even if it has some good points. These include language ambiguity, domain heterogeneity, extraction mistakes, difficulties in managing long-distance dependencies, and challenges in recognizing implicit links not expressly articulated in the text. Texts could also include metaphors or rhetorical devices that are hard for automated systems to understand.

Because of this, the CMM framework has several levels of validation, filtering, and refinement to make sure that the idea maps it generates are both structurally and semantically sound. Also, automatic concept map generating systems like the CMM are becoming better because to ongoing updates and advances in NLP approaches like transformer-based models, contextual embeddings, and neural semantic comprehension.

II. DOMAIN KNOWLEDGE FOR CMM

Concept Map Mining (CMM) relies on domain expertise to gain insight into a particular field of study, identify relevant ideas and relationships, and then organize and classify them. On the one hand, domain knowledge resources excel at understanding unstructured language and on the other, in extracting and retrieving domain-relevant information. Ontologies, thesaurus, and hierarchical taxonomies are examples of more complex systems, whereas simple dictionaries and lexicons compiled by humans are also examples. When it comes to enhancing semantic comprehension, each of these resources is crucial. Ontologies give more in-depth semantic frameworks that describe entities, classes, and connections within a domain, whereas dictionaries aid in clarifying meanings at the term level.

Despite their value, domain resources aren't always easy to get by and are often changing to reflect new information. Unfortunately, complete, current, machine-readable knowledge bases are lacking in many sectors, particularly in specialized scientific or technological fields. Expert input, manual annotation, or massive crowdsourcing are usually necessary for the labor-intensive production of domain knowledge resources. As fresh ideas arise, it becomes more difficult to keep up with this process, which is also time-consuming. Furthermore, the competence, reliability, and correctness of interpretation of the people engaged in their creation greatly affect the resources' quality. The final product shows any prejudice or carelessness on the part of the makers.

Due to the scarcity and maintenance challenges of high-quality domain resources, automated or semi-automatic approaches are being used more and more in current CMM systems to generate domain knowledge. These methods encompass a wide range of approaches, such as word extraction from corpora, automatic ontology synthesis, knowledge graph construction, statistical co-occurrence analysis, and sophisticated machine learning algorithms that can detect domain ideas autonomously. Processing massive amounts of domain-specific texts efficiently, flexibly, and scalable is made possible by automated domain knowledge production, which also lessens the need for human specialists.

Domain Corpus

A domain corpus is a collection of books, articles, and other resources that have been marked up to show what they are about in relation to a given academic field, industry, or issue. Domain corpora often possess a restricted vocabulary, unlike collections of general-purpose writings. This implies that they only have a small number of terminologies that are regularly used and have a clear meaning in the field. The statistical patterns, co-occurrence distributions, and semantic links of these domain-specific terms differ from their usage in general English.

Take the term "blood vessels," which is a compound word that means a specific physiological structure. This is because the words "blood" and "vessels" are used together in medical writing, even though they have different meanings in everyday English. To get exact concept extraction, domain corpora keep track of domain-specific phraseology, technical jargon, hard words, phrases with more than one word, and so forth.

Domain corpora make it feasible to identify morphological, syntactic, and semantic features that are peculiar to a certain field. This is especially useful for tasks like these.

- **Term Extraction:** detecting frequently occurring domain terms.
- **Relation Extraction:** discovering conceptual links based on co-occurrence patterns or syntactic dependencies.
- **Ontology Learning:** constructing hierarchical relationships between concepts.
- **Semantic Disambiguation:** correctly interpreting domain-specific meanings of ambiguous words.
- **Information Retrieval and Classification:** improving search accuracy within domain literature.

Large, representative domain databases can fulfill double duty:

1) resource development through the automated derivation of new domain dictionaries, taxonomies, and knowledge bases from the corpus; furthermore
2) enhancing resources by updating, validating, or supplementing existing subject knowledge using insights generated from corpora.

Because they give abundant contextual data for idea identification and connection modeling, domain corpora are particularly useful in the context of idea map mining. Inferring the importance of key ideas, the nature of relevant linkages, and the clustering of sub-concepts within the domain's overall knowledge structure are all possible through pattern extraction from a domain corpus. So, idea maps built with domain corpora are more likely to be consistent, correct in terms of semantics, and in line with what experts understand.

Because of their ability to automate the extraction of domain information that would normally need substantial expert input, domain corpora are crucial tools for improving the accuracy and breadth of CMM systems.

III. REVIEW OF LITERATURE

Kusumadewi, Rida & Kusmaryono, Imam. (2022) Students' cognitive development and capacity for original thought are intimately tied to the use of concept maps in the classroom. The purpose of this research is to find out how well students use concept maps to enhance their knowledge retention and critical thinking abilities while they are studying. Descriptive analysis is the study approach that was selected. Interviews, idea map evaluation rubrics, and questionnaires were used to gather research data. The percentage approach was used to examine the questionnaire data. A descriptive analysis is performed on the students' free-form responses to questions on the benefits and drawbacks of using concept maps. Most people who took the survey thought that using concept maps to learn was a great idea. Results from surveys and free-form questions show that students' conceptual maps help those better grasp complex ideas and develop their critical thinking abilities, which in turn boost their motivation, outlook, and performance in the classroom.

Vinothini Kasinathan et al., (2022) Despite the linear nature of PowerPoint presentations in lecture halls and classrooms, the brain mostly operates with interconnected and integrated core ideas. In contrast to the usual practise of students reading and reciting the slides, this article introduces a novel technique that may automatically construct a mind map using a collection of keywords and phrases retrieved from PowerPoint presentations. As part of its text mining strategy, the system uses an algorithm for extracting keywords and keyphrases. The survey findings show that there is a growing need for this kind of application in educational technology, and we look at the possible advantages and talk about them.

Alomari, Sara & Abdullah, Salha. (2019). Students have found that concept maps help them better understand classes that cover several topics or themes by outlining the relationships between various bits of information. But creating a concept map from scratch is an arduous and time-consuming process. It takes a lot of time and effort to read the whole book and figure out how the concepts relate to one another. Because of this inefficiency, there has been a flood of research into developing intelligent algorithms using different data mining techniques. By using the weighted K-nearest neighbors (KNN) algorithm in place of the traditional KNN method, this work aims to improve the TA-ARM technique. By optimizing the classification accuracy using the weighted KNN, we hope to produce a higher-quality idea map.

Li, Yancong et al., (2018) as a way to visualize information, idea maps are used by adaptive learning systems. Students' weak concepts may be identified via the analysis of concept maps, which enables us to assist their learning and provide instructional advice. The current research shows that it gets difficult to derive valuable information from concept maps when there are many concepts and complicated interactions among them. It seldom displays the learning outcomes of different student groups and fails to portray the characteristics of adaptive learning systems. Our LPG program automatically creates learning routes for adaptive learning systems. The learning performance of different student groups is carefully considered by the LPG algorithm. The topological sorting algorithm, clustering, and association rules mining all work together to generate idea maps with different learning properties; the latter two algorithms also provide a plethora of simpler learning paths. The experimental results show that the LPG algorithm is able to differentiate between different groups of students quite well and generate unique learning routes depending on their performance.

Pipitone, Arianna et al., (2014) in a virtual learning environment (VLE), the three main participants are the student (in the role of instructional designer), the tutor, and the student themselves. To control how well the learning materials, adhere to the course domain structure and to provide the system the structure of the course, instructional designers need simple interaction. Tutors want resources that allow them to see the big picture of how individual students or groups are progressing in the virtual learning environment (VLE) as they study. Finally, students need to be able to self-regulate when it comes to managing their own learning pace, reflecting on the methods they use to study, and making comparisons to their classmates. Our claim in this work is that all of these actors may satisfy the criteria by sharing an implicit representation of the information about the course domain. To back up our claim, we show a tool that was built as part of the I-TUTOR project. The program examines a relevant document corpus that defines the course domain, then produces a semantic space. This space is then shown as a 2D zoomable map. The map illustrates all the important domain ideas, and the application makes it easy to access instructional resources. Students' social interactions inside the virtual learning environment (VLE) and the texts they create while studying may also be included in the map. In addition to reporting the work's goals and underlying AI approaches, this paper provides a detailed explanation of the full system, including all of the evaluations conducted throughout the project.

Chen, Shyi-Ming & Bai, Shih-Ming. (2010). An essential area of study for adaptive learning systems is, of course, the construction of concept maps to guide learners' learning. The current approach to building concept maps is flawed because it ignores information on questions that learners have properly answered in favor of focusing on association rules that pertain to questions that have been incorrectly answered. The current approach also has the problem of creating concept maps with missing or superfluous linkages between ideas. This research introduces a novel approach to adaptive learning system data mining for the autonomous construction of idea maps. The shortcomings of the current technique may be remedied by implementing the suggested approach. In adaptive learning systems, it gives us a practical technique to automatically build idea maps.

Novak, Joseph & Cañas, Alberto. (2007). According to our definition, concept maps are visual aids for organizing and depicting connections between ideas, with a connecting line indicating a relationship between two concepts. The connection between the two ideas is specified by the words on the line, which are called connecting words or linking phrases. A common organizational pattern for ideas and claims is a hierarchical one, beginning with the broadest and most inclusive and working their way down to the most narrowly focused. We have found that building idea maps in relation to a specific topic we want to answer—a focus question—is the most effective way to do so. As a means of organizing information, a concept map may be useful in gaining a better understanding of a particular situation or event. This serves as the idea map's context. Our New Model for Education is supported by concept map-based learning environments, which are made possible through the integration of technology in software like CmapTools. This paper provides a brief history and theoretical background of concept maps, explains their construction, and goes on to show how this integration works. As a last step, we show how concept maps may be used to three different domains to arrange material according to domain experts' expertise, making it easier for learners to traverse.

IV. RESEARCH METHODOLOGY

In this research, we build an automated method for creating idea maps by using the Concept Map Miner (CMM) procedure. The Spring Framework is now being used to create the suggested system as a Web service that is based on Java. Unstructured material is converted into a structured, hierarchical idea map via a series of interrelated processes in the approach.

Data Source of Description

In terms of the Data Source Description phase, our method is defined by the use of English-language unstructured text, namely academic publications of a normal size. The publishing process guarantees high-quality work, and the language is formal, befitting an academic article. In order to achieve the aim of representing the information derived from text as a summary, the data source is restricted to just one document. The User selects the data source, which serves as input for the whole procedure.

Domain Definition

Regarding the Domain Definition stage, our method is defined by using a human User to aid in domain identification; this User is responsible for selecting the data source in step 4.1. Our method suggests using clustering and classification as supervised learning approaches for semi-automatic domain identification of a given text, but we'll save that for a later project. At this point, we recommend using an automatically-built domain thesaurus to identify components that belong to a certain domain. Concept maps are constructed by processing fresh texts from the same domain, which progressively builds the thesaurus, which provides a conceptual vocabulary for that domain.

Map Visualization

One distinguishing feature of this method is the creation of an appropriate user interface for perusing the produced idea map. Among its capabilities is the ability to show the hierarchy of ideas and to distinguish between sub-concepts. Because the method relies on automatically creating concept maps from texts—ideas that students don't always know—we think it's crucial to identify such concepts so students can better grasp and absorb the material.

Experimental Analysis

Here we show the first model of the method. With 26 sentences and 617 words, the material is written in English. Since the method is in its early stages of development, we are evaluating it primarily via the Element Identification step. First, we generated a concept map with all of the propositions that were collected from the data source. Second, we used the Ranking and Summarization process to filter the propositions.

We compared the automatically produced idea map to one that was hand-built by 10 domain experts so that we could examine the map's faithfulness to the text. We were given the following directions: (1) In order to get the experts started, we gave them some background on concept maps and what the experiment was all about. (2) We told them to keep the labels short, meaningful, and taken directly from the text. (3) We told them that nouns should be in the labels of concepts and verbs of relations. (4) We told them to change the labels that had named entities or prepositions to something more suitable.

V. RESULTS AND DISCUSSION

Tables showing accuracy and recall computed by comparing the approach's map construction with expert-generated maps follow. Table 1 displays the results of the identified ideas' analysis, which achieved a Precision of 0.52 and a Recall of 0.70. We neglected grammatical gender, number, and degree as examples of label flexion in expert-built concept maps in this investigation.

Table 1: Performance Metrics for Concept Identification Based on Expert Comparison

Concept Analysis											
Expert	**Exp.1**	**Exp.2**	**Exp.3**	**Exp.4**	**Exp.5**	**Exp.6**	**Exp.7**	**Exp.8**	**Exp.9**	**Exp.10**	**AVG**
Precision	0.58	0.41	0.46	0.57	0.49	0.51	0.48	0.40	0.39	0.45	0.52
Recall	0.60	0.74	0.71	0.77	0.62	0.70	0.59	0.74	0.64	0.81	0.70

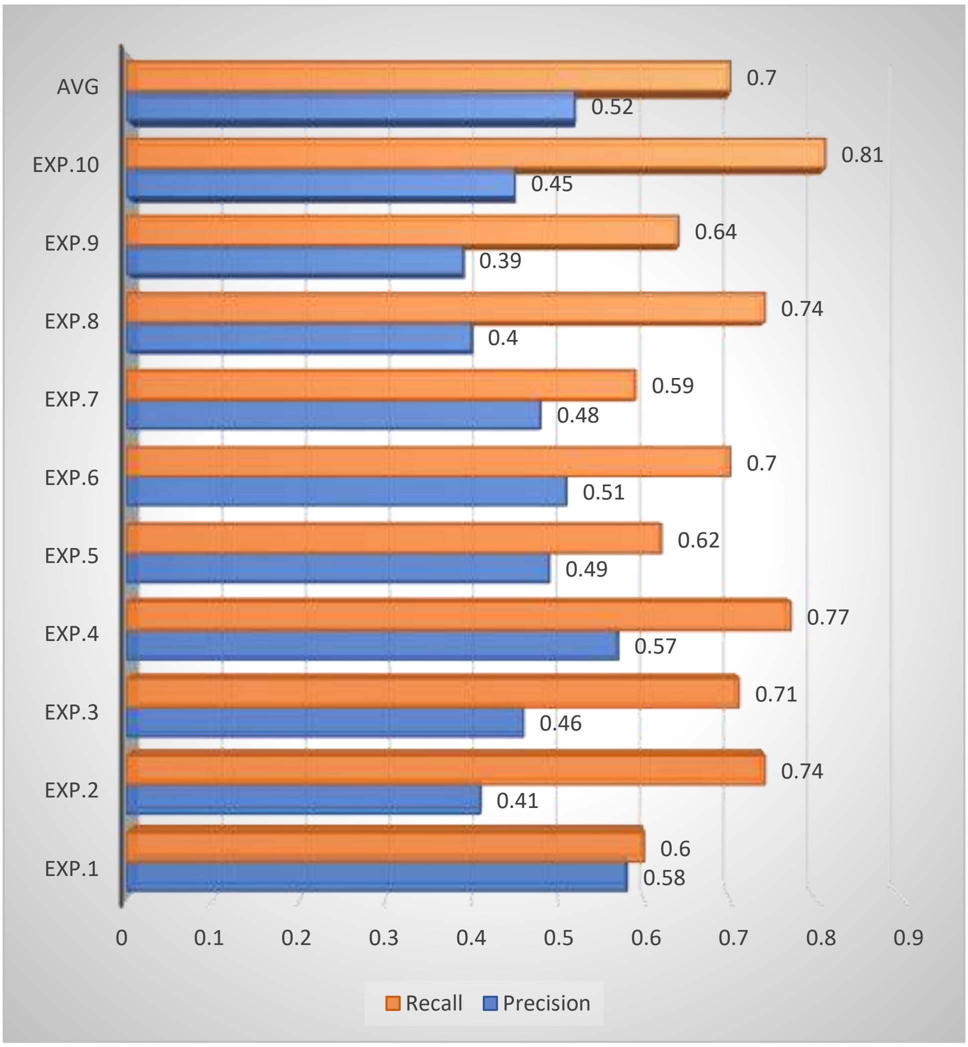

Figure 1: Performance Metrics for Concept Identification Based on Expert Comparison

Table 2 displays the results of the identified association's study, which yielded a Precision of 0.31 and a Recall of 0.47. If the linkages created by the expert's link the same ideas identically and have the same meaning, then we consider them comparable in this assessment. This is a shortcoming of our assessment that has to be addressed in further studies by looking at the labels of relations.

Table 2: Performance Metrics for Relationship Identification Based on Expert Comparison

Relationship Analysis											
Expert	**Exp.1**	**Exp.2**	**Exp.3**	**Exp.4**	**Exp.5**	**Exp.6**	**Exp.7**	**Exp.8**	**Exp.9**	**Exp.10**	**AVG**
Precision	0.35	0.26	0.21	0.35	0.29	0.37	0.20	0.21	0.32	0.27	0.31
Recall	0.43	0.49	0.48	0.41	0.39	0.45	0.35	0.54	0.47	0.49	0.47

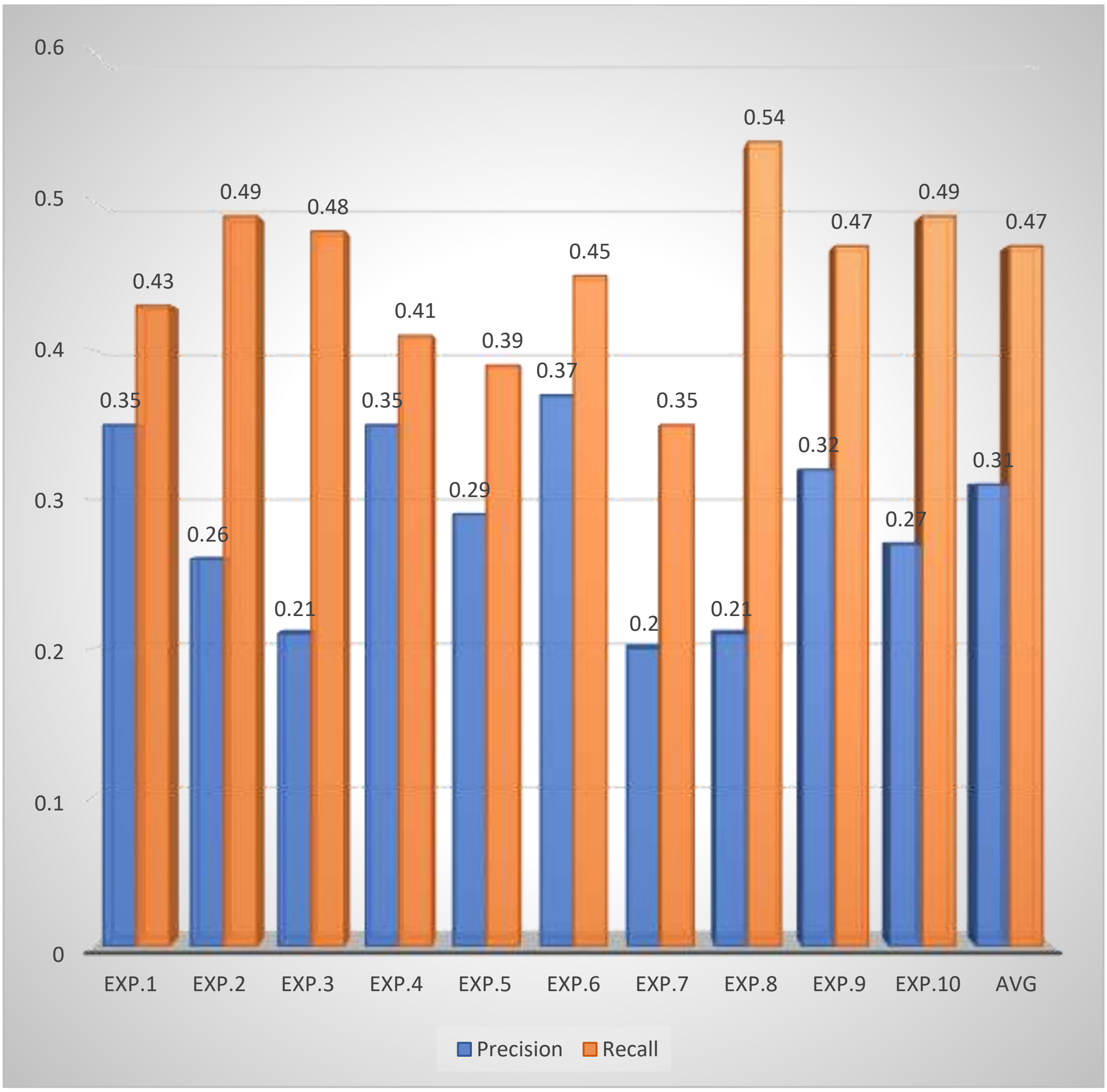

Figure 2: Performance Metrics for Relationship Identification Based on Expert Comparison

Even for domain specialists, creating a map from text is a challenging undertaking, as we can see from the experiment. To the contrary, it was instructed that the specialists should not incorporate the ideas into their mental models. They were instead told to stick to the ideas presented in the book, which they struggled with.

VI. CONCLUSION

The prototype efficiently retrieves a considerable number of relevant ideas and connections by combining the Concept Map Miner approach with domain-driven thesaurus building and automated element recognition methods. The results of the comparison with expert-generated maps are promising, especially in the area of concept recognition, but they also demonstrate where there is room for improvement, including in the areas of relationship extraction accuracy and semantic alignment between labels. These results show the system's strengths and weaknesses, indicating that future research should concentrate on improving relation label assessment, developing supervised learning for domain identification, and using more advanced natural language processing techniques.

REFERENCES

1. C. Dong *et al.*, “A survey of natural language generation,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2022.
2. V. Kasinathan, A. Mustapha, T. Appalasamy, and V. Thiruchelvam, “Otomind: A text mining approach to automatic concept mapping,” *Mathematical Statistician and Engineering Applications*, vol. 71, no. 2, pp. 192–207, 2022.
3. R. Kusumadewi and I. Kusmaryono, “Concept maps as dynamic tools to increase students' understanding of knowledge and creative thinking,” *Premiere Educandum: Jurnal Pendidikan Dasar dan Pembelajaran*, vol. 12, no. 1, pp. 1–12, 2022.
4. V. Dos Santos *et al.*, “Conceptual map creation from natural language processing: A systematic mapping study,” *Revista Brasileira de Informática na Educação*, vol. 27, no. 3, pp. 150–176, 2019.
5. S. Alomari and S. Abdullah, “Improving an AI-based algorithm to automatically generate concept maps,” *Computer and Information Science*, vol. 12, no. 4, pp. 72–72, 2019.
6. Y. Li, Z. Shao, X. Wang, X. Zhao, and Y. Guo, “A concept map-based learning paths automatic generation algorithm for adaptive learning systems,” *IEEE Access*, vol. 6, pp. 1–15, 2018.
7. A. Erdem, “Mind maps as a lifelong learning tool,” *Universal Journal of Educational Research*, vol. 5, no. 12A, pp. 1–7, 2017.
8. A. Pipitone, V. Cannella, and R. Pirrone, “Automatic concept maps generation in support of educational processes,” *Journal of E-Learning and Knowledge Society*, vol. 10, no. 1, pp. 85–103, 2014.
9. I. Qasim, J. W. Jeong, J. U. Heu, and D. H. Lee, “Concept map construction from text documents using affinity propagation,” *Journal of Information Science*, vol. 39, no. 6, pp. 719–736, 2013.
10. J. J. Villalón and R. A. Calvo, “Concept maps as cognitive visualizations of writing assignments,” *Educational Technology & Society*, vol. 14, no. 3, 2011.
11. S.-M. Chen and S.-M. Bai, “Using data mining techniques to automatically construct concept maps for adaptive learning systems,” *Expert Systems with Applications*, vol. 37, no. 6, pp. 4496–4503, 2010.
12. N.-S. Chen, C.-W. Wei, and H.-J. Chen, “Mining e-learning domain concept map from academic articles,” *Computers & Education*, vol. 50, no. 3, pp. 1009–1021, 2008.
13. W. M. Wang, C. F. Cheung, W. B. Lee, and S. K. Kwok, “Mining knowledge from natural language texts using fuzzy associated concept mapping,” *Information Processing & Management*, vol. 44, no. 5, pp. 1707–1719, 2008.
14. J. Novak and A. Cañas, “Theoretical origins of concept maps, how to construct them, and uses in education,” *Reflecting Education*, vol. 3, 2007.