Chatbot Program for Proposed Requirements in Korean Problem Specification Document

Young Yun Baek[1], Soojin Park[2], Young B. Park[3]
Dept. of Software Science, Dankook University, Seoul, South Korea[1, 3]
Graduate School of Management of Technology, Sogang University, Seoul, South Korea[2]

***Abstract*—In software engineering, requirement analysis is a crucial task throughout the entire process and holds significant importance. However, factors contributing to the failure of requirement analysis include communication breakdowns, divergent interpretations of requirements, and inadequate execution of requirements. To address these issues, a proposed approach involves utilizing NLP machine learning within Korean requirement documents to generate knowledge-based data and deduce actors and actions using natural language processing knowledge-based information. Actors and actions derived are then structured into a hierarchy of sentences through clustering, establishing a conceptual hierarchy between sentences. This is transformed into ontology data, resulting in the ultimate requirement list. A chatbot system provides users with the derived system event list, generating requirement diagrams and specification documents. Users can refer to the chatbot system's outputs to extract requirements. In this paper, the feasibility of this approach is demonstrated by applying it to a case involving Korean-language requirements for course enrollment.**

Keywords—Requirement engineering; NLP machine learning; clustering; Korean document; chatbot

I. Introduction

In software engineering, requirements analysis is the process of eliciting, analyzing, specifying, and validating the requirements that must be satisfied in software development. It is a crucial activity in the early stages of development and has a significant impact on the design and other phases of the software development lifecycle. Insufficient requirements analysis can lead to project failures, while a proper requirements elicitation process serves as the foundation for overall software product quality. Therefore, it is necessary to perform requirements analysis in a thorough and specific manner, considering its importance and high impact on the overall success or failure of a project [1].

Furthermore, requirements analysis represents the client's needs, contractual obligations, standard specifications, and documented information in software development. It expresses the conditions, functionalities, or capabilities that the development program must perform. It helps identify the desired features, goals, constraints, and other necessary information from the users for system development. Thus, requirements analysis is crucial in creating software that meets user expectations and is considered the most critical process in the software development life cycle [2],[3].

However, extracting clear requirements from problem statements and capturing requirements that encompass the overall software can still be challenging and complex due to the diverse values, attitudes, behavioral norms, beliefs, and communication of stakeholders, who have different perspectives [4]. Another reason is that problem statements are written in natural language. Natural language descriptions can be interpreted differently by individuals, and the same word can have different meanings. Furthermore, as the field of software expands and becomes more complex, analyzing clear requirements becomes difficult, and it requires fundamental knowledge to understand the system. Additionally, since humans are involved in the process of extracting these requirements, there are limitations in consistency and analysis due to the subjective nature of human work.

In this paper, to address these challenges, natural language processing and NLP artificial intelligence analysis are employed to extract key words from problem specification documents in the document phase. By extracting actor-behavior relationships and hierarchical data through clustering, sequential analysis of sentences and hierarchical analysis data are obtained, which are then structured into ontology to be used as foundational data for a chatbot. This approach allows for the extraction of consistent requirements from complex problem specification documents and proposes a method to resolve human effort by recommending and providing a list of requirements to users through question-and-answer interactions with the chatbot.

II. Related Work

A. Sequential Analysis for Requirement Classification and Word Entity Extraction

In Wang et al. [5], they addressed the problem of modeling the interaction between the physical environment and users in model generation for testing and validation in NLR (Natural Language Requirements) by using NLP techniques and model mapping rules to identify model elements. Jahan et al. [4] recognized the importance of automation in behavior modeling and proposed an automated approach for constructing SD (System Design) from natural language-written use case scenarios to bridge some of the gaps in the literature. Limaylla-Lunarejo et al. [6] applied machine learning algorithms combined with natural language processing to classify software requirements into functional and non-functional categories. Koscinski et al. [7] automated the formalization of NL (Natural Language) requirements by using IE (Information Extraction) techniques to extract structured information from NL SysRS (Natural Language System Requirements) data. Güneş et al. [8] automatically generated and visualized goal models based on user stories using a natural language processing pipeline and heuristics. Tobias et al. [9] proposed an approach called NoBERT, which utilizes the fine-tuning mechanism of BERT for classifying requirements. Saini et al. [10] proposed an automated approach that combines NLP and ML techniques to extract domain models. Tiwari et al. [11] proposed an approach using entity recognition NLP techniques to identify use case names and actor names in text-based requirements specifications. Imam et al. [12] proposed a method using SVM (Support Vector Machine) for recognizing use case entities in unstructured sentences.

These methods have strengths in sequential analysis for requirement classification and actor behavior extraction. However, they have limitations in establishing the relationships between related requirements. While it is possible to classify and extract requirements within a single sentence during requirement analysis, it is challenging to understand the overall flow and structure of the entire set of requirements, which is a drawback in grasping the flow of requirements.

B. *Dependency Analysis of Requirements through Hierarchical Analysis*

In Deshpande et al. [13], requirement dependencies were extracted using domain ontology and labeling learning. Zhang et al. [14] proposed an automated requirement term extraction and ranking framework based on a graph-based ranking algorithm. Wardhana et al. [15] suggested using ontology to represent the semantic context in system design.

These hierarchical analysis methods have the advantage of capturing the dependencies and hierarchical relationships between requirements by identifying the parent-child relationships among them. However, when performing hierarchical analysis alone, it is limited to simple comparisons between requirements, and it cannot analyze the relationships between words within sentences, which is a drawback.

III. RE CHATBOT SYSTEM CONFIGURATION

A. *System Configuration*

In [1] framework, clustering was added to enable sequential analysis and hierarchical analysis of sentences, which were then applied to ontology construction. By utilizing hierarchical analysis in ontology construction, it becomes possible to identify higher-level concepts of requirements between sentences. This allows for both word-centric sequential analysis and sentence-centric hierarchical analysis. By understanding the relationships between sentences, beyond simple sequential analysis for classification, it is possible to provide requirements that are prioritized and classified based on their inter-sentence dependencies.

Each sentence in the problem specification document is subjected to morphological analysis and BERT Q&A operations to generate the underlying knowledge data in advance. The system possesses data for sequential analysis through the aforementioned process, which is then combined with the results of clustered sentence data to construct ontology. Through the ontology, entities such as actors, actions, events, and systems are extracted. The extracted relationship data and ontology data exist as base data in the chatbot program. Users go through the process of final requirement specification by interacting with the chatbot through requirement Q&A sessions. Users can review the lists provided by the chatbot Q&A and use them as a reference to compose their requirements. The new system structure is depicted in Fig. 1.

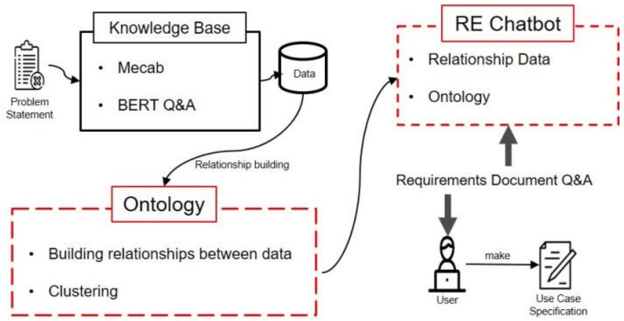


Fig. 1. System overview of process of creating components for writing the use case specification in the problem description document

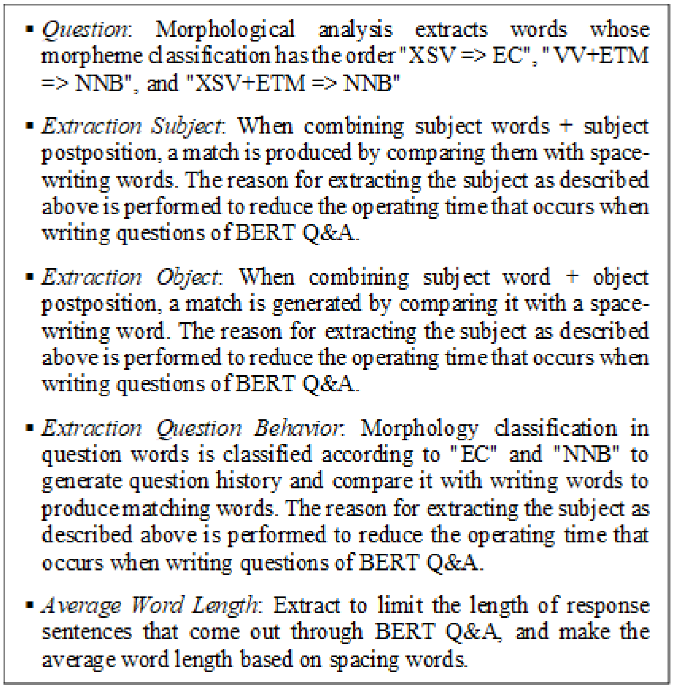
B. *Mecab Morphological Analysis, BERT Q&A Knowledge Base*

The process of constructing the knowledge base data is described in [1] following the approach. The problem statement is preprocessed using Mecab and BERT Q&A to perform morphological analysis and sequential analysis of the sentences, respectively. The Mecab morphological analysis method for constructing the knowledge base data from the problem statement is shown in Fig. 2, while the BERT Q&A analysis method is illustrated in Fig. 3.

C. *Clustering Configuration*

For clustering, two types of clustering were performed to cluster sentences. First, a tf-idf analysis was performed using only the subject words within the sentences. By excluding unnecessary words such as particles, adverbs, adjectives, etc., it becomes possible to analyze the core actors and actions in Korean sentences. Furthermore, by analyzing tf-idf based on the subject words, which focuses on the subject words, clustering can be performed. Since it is subject-word-centered clustering, it is possible to perform sequential analysis with the results of morphological analysis and sequential analysis of the sentences, allowing for clustered analysis.

Next, the results of tf-idf analysis are used to perform DBSCAN clustering and Agglomerative clustering. DBSCAN clustering is applied to obtain clustered results of the problem statement text based on subject words. This allows for the analysis of similar sentences that include a specific subject word and the analysis of similar subject words. Agglomerative clustering is performed to obtain hierarchical clustering results among sentences. This enables the analysis of the hierarchical relationships of a specific sentence with other sentences. Through these two clustering methods, the sentences are analyzed in a hierarchical structure, allowing for the extraction of sentences with hierarchical relationships.



- *Question*: Morphological analysis extracts words whose morpheme classification has the order "XSV => EC", "VV+ETM => NNB", and "XSV+ETM => NNB"
- *Extraction Subject*: When combining subject words + subject postposition, a match is produced by comparing them with space-writing words. The reason for extracting the subject as described above is performed to reduce the operating time that occurs when writing questions of BERT Q&A.
- *Extraction Object*: When combining subject word + object postposition, a match is generated by comparing it with a space-writing word. The reason for extracting the subject as described above is performed to reduce the operating time that occurs when writing questions of BERT Q&A.
- *Extraction Question Behavior*: Morphology classification in question words is classified according to "EC" and "NNB" to generate question history and compare it with writing words to produce matching words. The reason for extracting the subject as described above is performed to reduce the operating time that occurs when writing questions of BERT Q&A.
- *Average Word Length*: Extract to limit the length of response sentences that come out through BERT Q&A, and make the average word length based on spacing words.

Fig. 2. Morphological analysis algorithm of the RE chatbot system.

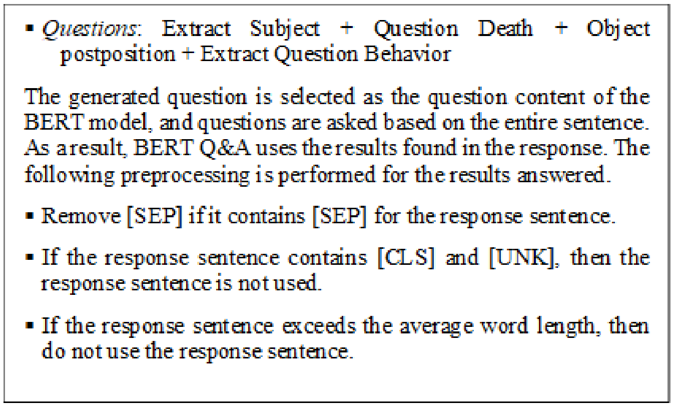
D. Ontology Configuration

Based on the extracted subject words from the pre-constructed knowledge base data, the subject part of the ontology is formed. This subject part consists of words that are used as actors in the requirements. Subsequently, the predicate part and the object part are written with the subject words as the focus. Next, the necessary steps are taken to construct the predicate part and the object part. The following is the method for constructing the predicate part and the object part, which are essential for the operation of the entire system.

To express the actions of actor words, the predicate part of the ontology is constructed as "Action". Next, to construct the object part of this predicate, the actions that actors can perform within the system are extracted from the knowledge base data and used as objects. This object part consists of sentences that are used as actions in the requirements analysis.

To construct associated words for actor words, the predicate part of the ontology is structured as "Associated Words." Then, to construct the object part of this predicate, words associated with the subject part are extracted from the clustering results and used as objects. This object part is used in requirements analysis to analyze similar words centered around a given word.

To create associated sentences for the word "actor," the ontology's descriptors are structured as "associated sentences." Next, to form the purpose of these descriptors, sentences containing subject words are extracted from the clustering results and used as the purpose. This purpose is employed to provide similar requirement sentences centered around the main sentence when inputting key sentences during requirement analysis.



- *Questions*: Extract Subject + Question Death + Object postposition + Extract Question Behavior

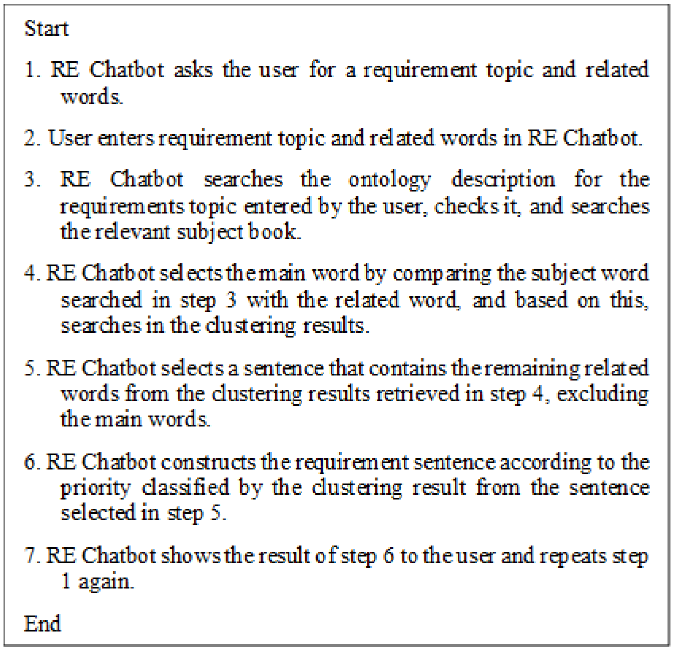
The generated question is selected as the question content of the BERT model, and questions are asked based on the entire sentence. As a result, BERT Q&A uses the results found in the response. The following preprocessing is performed for the results answered.

- Remove [SEP] if it contains [SEP] for the response sentence.
- If the response sentence contains [CLS] and [UNK], then the response sentence is not used.
- If the response sentence exceeds the average word length, then do not use the response sentence.

Fig. 3. BERT Q&A analysis algorithm of the RE chatbot system.

E. RE Chatbot Configuration

The chatbot possesses the constructed ontology file from the previous steps, along with the clustered sentence content and hierarchy information as its data. The chatbot reads the constructed ontology file and clustering information to generate actor and action information based on the user's input of key sentences and words. It provides a prioritized list of requirements as a result. Users can utilize the requirement list provided by the chatbot for their analysis. The chatbot algorithm used by the user is illustrated in Fig. 4.



```
Start
1. RE Chatbot asks the user for a requirement topic and related
   words.
2. User enters requirement topic and related words in RE Chatbot.
3. RE Chatbot searches the ontology description for the
   requirements topic entered by the user, checks it, and searches
   the relevant subject book.
4. RE Chatbot selects the main word by comparing the subject word
   searched in step 3 with the related word, and based on this,
   searches in the clustering results.
5. RE Chatbot selects a sentence that contains the remaining related
   words from the clustering results retrieved in step 4, excluding
   the main words.
6. RE Chatbot constructs the requirement sentence according to the
   priority classified by the clustering result from the sentence
   selected in step 5.
7. RE Chatbot shows the result of step 6 to the user and repeats step
   1 again.
End
```

Fig. 4. Chatbot operation algorithm of the RE Chatbot system.

IV. Case Study Applying Requirements Document

To validate the proposed system, a selection of requirements documents, specifically those related to course enrollment, was chosen. These requirements were used to conduct the validation of the system by applying the suggested requirement diagram and requirement specification automatic generation system outlined in the paper.

A. Clustering

Clustering was performed using two approaches to cluster the sentences. First, the sentences were analyzed using tf-idf based on the extracted subject words within each sentence. DBSCAN clustering was then applied to this dataset, resulting in sentence clustering based on subject words. Similarly, the subject words within the sentences were extracted and tf-idf analysis was performed. Agglomerative clustering was applied to this dataset to achieve sentence clustering based on subject words.

B. Ontology Contruction

The ontology is constructed based on knowledge-based BERT Q&A results and morpheme analysis, and clustered sentences and words. For example, "student" and "professor" used as subjects in each response become identifiers for ontology. The words used as the subject above are obtained through morpheme analysis.

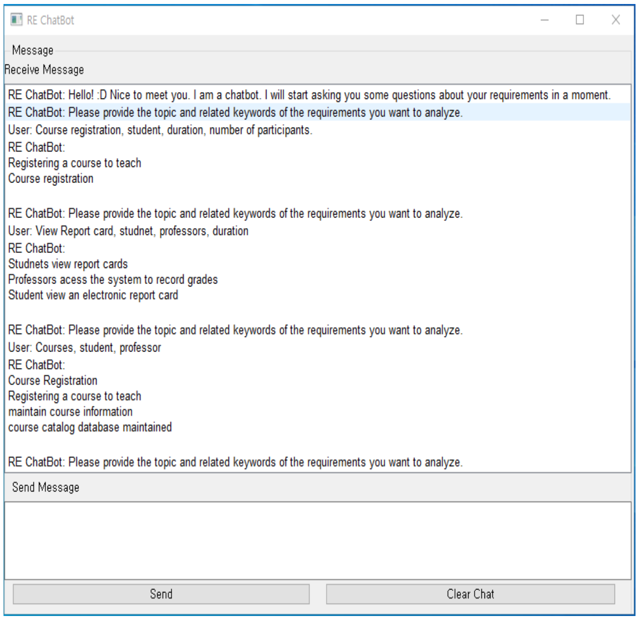
Create a question sentence with the given subject as the focus: "What can be viewed through the system by a student?" Using BERT Q&A, the response word "transcript" is obtained. Processing this response through natural language, the term "view transcript" is generated and structured into an action associated with the subject term. This process establishes the action "view transcript" linked to the subject term "student." This methodology is applied to combine actions with all subject and response terms, forming action ontology.

Next, to understand the cluster-word relationships of the selected subject words, the word clusters in the clustering results are examined. Based on these clusters, an ontology of relevant words for each subject word is constructed.

Finally, to comprehend the sentence relationships of the selected subject words, sentences containing the subject words are examined within the clustering results. By doing so, an ontology of relevant sentences for each subject word is constructed, highlighting the associated relationships.

C. Providing a List of Requirements based on the Responses from the RE Chatbot

Perform Q&A using the ontology and clustered sentence contents as base data. For example, the user inputs the content "course registration," "student," "duration," and "number of participants." Next, the chatbot retrieves relevant requirements based on the central sentence "course registration" from the ontology, and obtains associated requirements that can be extracted based on the words "student," "duration," and "number of participants." Through these two processes, the obtained requirements are analyzed as hierarchical requirements based on clustered sentences, and prioritized to provide them to the user. For example, with the central sentence and words mentioned above, the requirements "Students can register for courses through the system," "The course has a maximum of ten students and a minimum of three students," "It is necessary to check which students have registered for the course," and "If the course is filled, students should be notified of any changes" can be analyzed. Among them, the requirement "Students can register for courses through the system" has the highest relevance to the central sentence and words, so it is provided as the top-level requirement. The result of the chatbot is shown in Fig. 5.



```
RE ChatBot: Hello! :D Nice to meet you. I am a chatbot. I will start asking you some questions about your requirements in a moment.
RE ChatBot: Please provide the topic and related keywords of the requirements you want to analyze.
User: Course registration, student, duration, number of participants.
RE ChatBot:
Registering a course to teach
Course registration

RE ChatBot: Please provide the topic and related keywords of the requirements you want to analyze.
User: View Report card, studnet, professors, duration
RE ChatBot:
Studnets view report cards
Professors acess the system to record grades
Student view an electronic report card

RE ChatBot: Please provide the topic and related keywords of the requirements you want to analyze.
User: Courses, student, professor
RE ChatBot:
Course Registration
Registering a course to teach
maintain course information
course catalog database maintained

RE ChatBot: Please provide the topic and related keywords of the requirements you want to analyze.
```

Fig. 5. An example of the final outcome through the execution of the RE Chatbot system

V. Conclusion and Future Research Directions

Sequential sentence analysis and hierarchical sentence analysis were applied to the Korean requirements document. Through this process, requirements were extracted from the document by combining knowledge-based data and sentences from the problem statement, and proposing requirement elements with relationships through question and answer interactions with users using a chatbot. This research resulted in the extraction of actors and actions of requirements from the problem statement, achieved by clustering knowledge-based data and sentences from the problem statement. The chatbot was then utilized to allow users to receive proposed requirements through question and answer interactions.

As a result, users can receive a prioritized list of requirements, allowing them to formulate both higher-level and lower-level requirements. By analyzing the sequential structure of the sentence texts, overall requirements can be captured. Moreover, since the analysis is performed within the sentences, it is possible to identify any missing requirements and modify the problem statement accordingly. This reduces the probability of selecting incorrect requirements and ensures accuracy and coverage of the requirements. Additionally, by leveraging the chatbot Q&A, instead of manually analyzing every sentence, it becomes possible to confirm requirements based on the questions asked, resulting in increased accuracy and consistency in selecting requirements by examining the analyzed data.

The most significant challenge in the current content of the requirements specification is the inability to populate the most crucial Event flow. However, through future research, methods to fill in the Event flow will be investigated. Additionally, if a user desires to include actors or actions that cannot be extracted through morphological analysis or are not present in the requirement sentences, methods to add such information will be studied. Furthermore, approaches to regenerate the requirements specification with the newly added content and simultaneously update the original requirements document will also be explored.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP), grant funded by the Korean Government (MSIT) (2021-0-00177, High Assurance of Smart Contract for Secure Software Development Life Cycle).

REFERENCES

[1] M. Muqeem, S. Ahmad, J. Nazeer, M. F. Farooqui, and A Alam, “Selection of requirement elicitation techniques: a neural network based approach,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 351-359, 2022.

[2] Y. Y. Baek, Y. B. Park, “Suggested automatic creation of use case diagrams through machine learning analysis in Korean requirements documents,” *Journal of the Institute of Electronics Engineers of Korea*, vol. 2022, no. 6, pp. 2737-2739, 2022.

[3] H. Noor, M. Tariq, A. Yousaf, H. W. Ali, A. A. Moqeet, A. B. Hamid, and O. Naseer, "Emerging Requirement Engineering Models: Identifying Challenges is Important and Providing Solutions is Even Better," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 646-656, 2021.

[4] M. K. Hanif, M. R. Talib, N. U. Haq, A. Mansoor, M. U. Sarwar, and N. Ayub, "A collaborative approach for effective requirement elicitation in oblivious client environment," *International Journal of Advanced Computer Science And Applications*, vol. 8, no. 6, pp. 179-186, 2017.

[5] C. Wang, L. Hou, and X. Chen, "Extracting Requirements Models from Natural-Language Document for Embedded Systems," in *Proceedings of the 30th International Requirements Engineering Conference Workshops (REW)*, 2022, pp. 18-21.

[6] M. Jahan, Z. S. H. Abad, and B. Far, "Generating sequence diagram from natural language requirements," in *Proceedings of the 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 39-48.

[7] M. I. Limaylla-Lunarejo, N. Condori-Fernandez, and M. R. Luaces, "Towards an automatic requirements classification in a new Spanish dataset," in *Proceedings of the 30th International Requirements Engineering Conference (RE)*, 2022, pp. 270-271.

[8] V. Koscinski, C. Gambardella, E. Gerstner, M. Zappavigna, J. Cassetti, and M. Mirakhorli, "A Natural Language Processing Technique for Formalization of Systems Requirement Specifications," in *Proceedings of the 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 350-356.

[9] T. Güneș, and F. B. Aydemir, "Automated goal model extraction from user stories using NLP," in *Proceedings of the 28th International Requirements Engineering Conference (RE)*, 2020, pp. 382-387.

[10] T. Hey, J. Keim, A. Koziolek, and W. F. Tichy, "Norbert: Transfer learning for requirements classification," in *Proceedings of the 28th International Requirements Engineering Conference (RE)*, 2020, pp. 169-179.

[11] R. Saini, G. Mussbacher, J. L. Guo, and J. Kienzle, "Towards queryable and traceable domain models," in *Proceedings of the 28th International Requirements Engineering Conference (RE)*, 2020, pp. 334-339.

[12] A. T. Imam, A. Alhroob, and W. Alzyadat, "SVM machine learning classifier to automate the extraction of SRS elements," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 179-186, 2021.

[13] S. Tiwari, S. S. Rathore, S. Sagar, and Y. Mirani, "Identifying use case elements from textual specification: A preliminary study," in *Proceedings of the 28th International Requirements Engineering Conference (RE)*, 2020, pp. 410-411.

[14] G. Deshpande, Q. Motger, C. Palomares, I. Kamra, K. Biesialska, X. Franch, and J. Ho, "Requirements dependency extraction by integrating active learning with ontology-based retrieval," in *Proceedings of the 28th International Requirements Engineering Conference (RE)*, 2020, pp. 78-89.

[15] H. Wardhana, A. Ashari, and A. K. Sari, "Transformation of sysml requirement diagram into owl ontologies," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 106-114, 2020.