Prediction-Based NLP System by Boyer-Moore Algorithm for Requirements Elicitation

Dr A.Sumithra¹, K.Poongothai², Dr S.Gavaskar³

¹Associate Professor, Dept of Computer Science & Engineering, VSB College of Engineering Technical Campus
²Assistant Professor, Dept of Computer Science & Engineering, VSB College of Engineering Technical Campus
³Assistant Professor, School of Computer Science and Engineering, Bharthiyar Univeristy, Coimbatore, India

Abstract—Natural Language Processing (NLP) has been around for a number of decades. In this paper, NLP is used to gather the information from the problem declaration by some of the techniques in requirement engineering. In the existing system, NLP-KAOS in text mining is used to extract the goals from the engineers for research publications. NLP-KAOS is a technique which is used to produce goal specification for any kind of system or application and used to represent system specification through different concepts. In the proposed system, Prediction-Based with Boyer-Moore string search algorithm is an impression which is used in data mining technique for predicting the requirements for complex system where the requirement engineer can be used it frequently by the term ontologies learning environment and improves the performance of the system in terms of space and time complexity. By this algorithm, the perfect goals can be framed by gathering the collected requirements and reduce the ambiguity in meaning in gathered requirements.

Keywords—Requirement Engineering, NLP, KAOS, Ontologies, Prediction-Based, Boyer-Moore algorithm.

I. INTRODUCTION

Software Engineering is one of the expanses which play the important role with Data Mining. Software Engineering mainly concerned with tools, methods, practices for emerging a specialized software product. Software Engineering has many of the Requirement Engineering (RE) Process. Requirement Engineering (RE) Process depending on the presentation territory, the persons involved and the organization developing the requirements.

Requirement Engineering refers to the process of framing, documenting and preserving the software requirement. Requirement Engineering included some of the activities like Possibility study, elicitation and examination, description and documentation [9]. Requirement engineering highlights the use of systematic and repeatable methods that confirm the comprehensiveness, reliability and consequence of the system requirements.

Requirement Engineering is a branch of Software Engineering concerned with the real world goals, functions and constraints on software systems. The processes used for RE vary widely depending on the application domain, the people involved and the organization developing the requirements. Requirements Engineering involves two main processes are Requirements Development and Requirements Management.
Requirement Elicitation is the part of Requirement Engineering Process. To complete the requirement engineering process, the requirements are succeeded using a distinct change control process which always residues consistent. The major objective of this process is to avoid the misunderstanding between the participants and the predictors.

The Requirement Engineering (RE) is an approach which is used to understand what a system will perform in various technology, methods, tools, issues that need to be considered as critical [6]. The requirement engineering is used to communicate between the participants or stakeholders and the engineer.

II. MOTIVATION

In this paper, Natural Language Processing in Requirement Elicitation process is used to gather the list of requirements for the given problem statement or system. The main motivation for using the NLP system is computerization the requirement engineering process is too complex and difficult which addresses the requirement engineering elicitation and area understanding. The software requirement elicitation addresses the excluded performance of the system and restrictions under which it must work. Requirements are need to be clearly quantified or specified and standard for all system execution or system application. Ontologies which present here are non-categorized relationships i.e., the non-categorized relationship normally denoted about the features of heavyweight or formal ontologies. Ontologies have subclass of taxonomies where it describes about the easier problem to solve rather than learning ontologies where the semantic information is disqualified from the taxonomy algorithm, but they are comprised in the goal extraction algorithm.

To have the perfect goals from the collected requirements hereby we proposed Prediction-Based NLP technique. In addition to this technique, Boyer-Moore Search Algorithm is used for the performance of the system which can be improved by time and space complexity.

III. EXISTING METHODOLOGY

A. NATURAL LANGUAGE PROCESSING (NLP)

The methodology used to describe about the requirement elicitation is Natural Language Processing (NLP). Natural language processing system takes strings of words (sentences) as their input and
produce structured representations capturing the meaning of those strings as their output. A natural language understanding system helping as an interface to a database that might accept questions in English which relate to the kind of data held by the database. In this case, the meaning of the input (the output of the system) might be expressed in terms of structured SQL queries which can be directly submitted to the database.

The major aspects that are implemented in Natural Language Processing are:

**Syntax:** The syntax represents in the form of language. Normally it is specified by a grammar. Natural language is much more difficult than the formal languages and second hand for the artificial languages of logics and computer programs.

**Semantics:** The semantics which provides the meaning of the statements or sentences of the language. This semantic theory occurs when we build a natural language understanding system for a certain application.

**Pragmatics:** The pragmatic component explains how the statements or sentences relate to the world. To recognize these languages, a proxy should contemplate more than the sentence or statements.

The main four levels that are used in Natural Language Processing are:

**Morphological:** The main drive of using this level is to breakdown the strings of language input into set of tokens equivalent to separate words, sub-words and punctuation forms i.e., carry discrete meaning.

**Syntactic:** The purpose of syntax analysis is two-fold: to check that a string of words (a sentence) is well-rounded and to break it up into a structure that shows the syntactic relationships between the different words.

**Semantic:** The purpose of using this semantic analysis is the lexicon that must be stretched to include semantic definitions for each word it contains and the grammar must be extended to specify how the semantics of any phrase are formed from the semantics of its component parts i.e., derive absolute meaning.

**Pragmatic:** The purpose of using this pragmatic analysis is to deduce the results of semantic analysis from the viewpoint of a specific context i.e., derive knowledge.

Some of the important researched tasks that are included in Natural Language Processing are:

**Automatic Summarization:** It produces a readable piece of text which often used to produce the reviews of various texts.

**Machine Translation:** It is used to translate text from one human language to another by performing mechanically i.e., requiring all of the different types of knowledge that humans in order to solve properly.

**Natural Language Generation:** It converts the information from computer database into the human readable language.

**Parsing:** It determines the parse tree (grammatical analysis) of a given sentence. The grammar for natural language is ambiguous and typical sentences have multiple possible analyses.

The relationship between the Goals can be extracted using Taxonomy creation which computed by the standard cosine vector as:

$$S_{cos} (u,v) = \frac{X_u X_v}{||X_u|| \cdot ||X_v||}$$  \hspace{1cm} (1)$$

Where $X_u$ is a binary-term vector and $X_v$ are the keywords

The lists of Goal-specific keywords are defined by isolating the goal-related sentences where the goals should be analyzed.

| Intentional | Objective, aim, purpose, achieve, maintain, avoid, ensure, want, wish, motivate |

**Table 1: List of Goal-Related Keywords**
Amelioration | Improve, increase, decrease, reduce, enhance, support, provide, make

B. ONTOLOGIES
The Semantic Web is the second generation which is mainly used to share and reuse the data across application, creativity and community (unrestricted) boundaries [17]. Important module of the Semantic Web is Ontology i.e., the study of existence [17]. The term Ontology is mainly used in the Semantic Web that refers the designed data for wide-ranging and transferable or lightweight machine understanding [1]. The Ontology learning significantly supports ontology engineers to construct the ontologies. The ontology-learning framework includes

Importing: It is a technique which is used to merge the existing structures or defining the mapping rules between the structures for ontologies.

Extraction: It is the major parts of the target ontology, which can support for the web documents.

Pruning: It is the process of the target result outline from import, reuse and extraction which is used to pruned for the proper fit for the prime purpose.

Refinement: It completes the ontology at a fine granularity that profits for the pruned process.

The ontology delivers the main function to machine-processable data on the semantic web. Ontology mainly functions as the metadata pattern for the system. The other techniques that can be used for the future implementation in ontologies are reverse engineering of ontologies from database schema or from learning XML documents [1].

The chief objective of the Semantic Web is to permit the web contents that should be sanction by both human and software agents. The web contents cannot be offered in a formal method. Ontology is a recognized representation which is mainly used for the representation for the concepts of domain and semantic relationship between concepts. The semantic web and ontology techniques are used in software engineering with various advantages. The software process ontology is defined with innumerable software activities, process phases and existing process. Each activity is associated with software artifacts. Building the ontology is not an easy task to be performed.

Ontology plays an important role in Semantic Web [7]. Ontology is represented in the form of Class Diagrams. By using the ontology approach, we can estimate the quality of requirements and describe the four quality characteristics as correctness, completeness, consistency and unambiguity [7].

IV. PROPOSED METHODOLOGY

A. PREDICTION-BASED NLP TECHNIQUE
Prediction is a data mining technique which can be applied for various domains such as Software Engineering, Artificial Intelligence and other domains. Prediction is a model which is used to track the continuous valued- function. Prediction is different from classification. As we all know that, classification is used to predict the categorical class label whereas prediction is used to predict the identity of one kind based purely on the description of another which is related to thing. Predictions are often tell about what is going to be happen in future and the predicted values are always continuous. Neural network is one of the important predictor used in prediction technique. Most of the prediction techniques are Simple Statistical Model, Non-Linear Statistics and Neural Networks.

Prediction-Based is a fundamental approach that is used to predict the random variable to another random variable. By using the supervised learning method, computer program captures all structural information or data which is also used to derive the conclusions [15]. In this paper, machine learning techniques mainly rely on the prediction model. In prediction model, the discretization is used significantly to improve the performance of the model. Normally, discretization refers to the method of renovating (converting) the numeric range into nominal classes [15].
B. BOYER-MOORE SEARCH ALGORITHM

Boyer-Moore string search algorithm is one of the efficient string searching algorithms which preprocess the string being searched for the pattern, but not the string being searched in the text [16]. It uses the information gathered during the preprocess resulting in the lower constant factor than the other string algorithm. The Boyer-Moore string search algorithm scans the character from right to left which widely uses the string in sequential form. This algorithm is fastest known search algorithm which has the advantage of the dependencies between the characters. It has the main advantage in search and replaces the operations in text editor.

Example:

```
<table>
<thead>
<tr>
<th>FAST SEARCH STRING METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>P</td>
</tr>
</tbody>
</table>
```

Consider ‘T’ as Text and ‘P’ as Pattern. Check whether P occurs in T at some position in the right-to-left scanning manner. If ‘P’ is present in ‘T’, then the position of the text is match in ‘T’ and the character will be displayed.

The Boyer-Moore algorithm uses two shifting functions are bad character rule and good suffix rule. The bad character rule keeps information about how pattern matches against shift of itself. This information is used to avoid useless shift of the pattern. The good suffix rule uses the prefix function, the pattern and the text as inputs to find occurrence of the pattern within the text and returns the number of shifts after the first occurrence.

The algorithm preprocesses the pattern and requires the time efficiency as $\theta(m+\sigma)$. The time complexity of the Boyer-Moore search algorithm is $O(m+n)$. This algorithm mostly tries to find a given position in the text which ends at the position and runs faster as the pattern length increases.

The steps involved in the Boyer-Moore string search algorithm as

*Step 1:* Start comparing from the right end of the pattern
*Step 2:* Identify the mismatch character in the text and its position
*Step 3:* Find the rightmost occurrence of this character in the pattern, i.e., to the left of the mismatch position
*Step 4:* Move the pattern rightwards to align these two characters
V. CONCLUSION

This paper has presented a unique approach of NLP-KAOS which is existing one mainly based on data mining techniques which aims to make more efficient of early stages of requirement elicitation in software engineering. The main aim of using this Natural Language Processing is to extract the useful requirements from the given problem statement. It plays an important role in Software Engineering, Artificial Intelligence and even though in Data Mining. It helps the requirement engineers to gather the information from abstract of any research publications by extracting requirements from text.

In the proposed system, Prediction-Based NLP Technique with Boyer-Moore algorithm is efficient and secure search algorithm for searching the string from the sentence. It is mainly implemented to reduce the ambiguity in meaning and improve the performance of the system by space and time complexity to frame the perfect goals from the collected requirements. The accuracy of the perfect goals can be improved from the collected requirements.
REFERENCES