IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-4, Issue-11 E-ISSN: 2347-2693

A Natural Language Processing Based Approach Using Stochastic Petri Nets For Understanding Software Requirement Specifications

Rinku S.Ashtankar[1*] and Warsha M.Choudhari[2]

[1*] Assistant Professor, ITM College of Engineering, Kamptee
RTMNU, Nagpur ,India
[2] Assistant Professor, Datta Meghe Institute of Engineering Technology & Research,
RTMNU, Wardha, India

Available online at: www.ijcseonline.org

Received: Oct/22/2016 Revised: Oct/31/2016 Accepted: Nov/24/2016 Published: Nov/30/2016

***Abstract*— Language is a hallmark of intelligence, and endowing computers with the ability to analyze and generate language as a field of research is known as Natural Language Processing (NLP) - has been the dream of Artificial Intelligence. Software requirements are typically captured in natural languages (NL) such as English and then analyzed by software engineers to generate a formal software design/model. However, English is syntactically ambiguous and semantically inconsistent. Hence, English specifications of software requirements cannot only result in erroneous and absurd software designs and implementations but, the informal nature of English is also a main obstacle in machine processing of English complex specification of the software requirements. To tackle this key dispute, there is need to introduce a controlled NL representation for software requirements, to generate perfect and consistent software models. Proposed framework aims to model complex software requirements expressed in natural language and represent them with a new methodology that captures the natural language understanding(NLU) of events and models them using Stochastic Petri Nets (SPN) instead of only intermediate graph based structure using techniques of Natural Language Processing (NLP), this helps in removing ambiguity and corrects interpretation of requirements. To eliminate ambiguity, work combines all the different meanings (SPN graphs) of each ambiguous sentence into colored SPN graph. SPNs are state machines that help us to visualize better, the combined SPN graph. It can also represent knowledge about the requirement, which can be used to derive test case in early development phase. Hence aim of proposed work is twofold that overcomes the problem of ambiguity and knowledge representation. Stakeholder's document is input to framework, pre-processed by some pre-filter with certain functionality to improve the parsing. This parsed output gets converted into simple graph which in turn is converted into SPN graph with color representation to improve ambiguity. Pre-filter may be designed with self-learning capabilities to perk up output without human involvement.**

***Keywords*— NLP, SPN, SRS**

I. INTRODUCTION

NLP [1] is a relatively recent area of research and application, as compared to other information technology approaches, there has been sufficient successes to-date that suggests that NLP-based information access technologies will continue to be a major area of research and development in information systems now and far into the future.

The technologically driven world in which we live in has increased the necessity for human interaction with systems, particularly with computer-based systems that are used to accomplish a vast variety of tasks with the aim of helping the user in achieving its goal. This interaction is not always an easy one because of the unfortunate disconnect between how humans function and how technology responds. Especially important is the complex relationship between thought and language because a meaning or understanding has to be drawn from what is being said. That is why it is desired to integrate intelligence into these systems in order for them to become more efficient. For a system to be intelligent it should demonstrate some level of understanding, for instance a help system should understand what is being requested and respond appropriately. This kind of intelligent reaction or response between humans makes common sense, but it is more complex to implement it into machines that think. Natural Language processing support in such situations.

In general, software development projects are started with gathering stakeholders' requirements. In most cases, these requirements are specified using a natural language. Requirements specified [2] using a natural language, unlike those which are specified using formal models, are inherently ambiguous, inconsistent, and to some extent incomplete. Even though it has been widely acknowledged, that specifying software requirements using natural languages is problematic, these kinds of practices in real projects cannot easily resist. It is due to the fact that natural languages still become the easiest means of communication among project stakeholders. On the other hand, most practitioners believe that using formal models can provide better requirement specification in terms of low ambiguity and also high degree of consistency and completeness.

II. GAP ANALYSIS

In research paper [3] author worked on the issue of ambiguity and incompleteness of software requirement specification in natural language. He proposed a method to transform s/w requirement in NL to formal specification i.e. in object-oriented specification. However a lot of work is done, but that has certain limitations such as, it works for requirement specified in specific format k/as Concern-Aware Requirements Engineering format. For syntactic analysis of text it uses Reed-Kellogg [4] sentence diagramming system. It filters requirements in particular format such as Requirement= Subject + Verb + Target + [Way]

A requirement represents an action; an activity performed by an agent who affects/changes one state of an entity/object (an agent or a resource).Subject represents the agent who executes the behavior (the activity prescribed by the verb). Verb describes the activity taken by the agent (subject). Target can be physical or conceptual entity (object). The entity (object) has a number of properties (attributes), which will be affected by the activity. Meanwhile, way defines the way in which an action will be taken. Way can either define a manner or a utilized instrument (a means) to take the prescribed action. In this specification format, the subject, the verb, and the target are required whereas the way is optional.

Paper explains this concept with real world industrial project that helps for tracing electors participation in a particular election using handheld devices called as the Voter Tracking System (VTS).

Research paper [5] presents framework to generate Semantics of Business Vocabulary and Rules (SBVR) standard based controlled representation of English Software Requirement Specification(SRS) to overcome ambiguity [6] and inconsistency. According to the author few scientists have proposed various approaches to identify and measure the typical ambiguities in a NL SRS. But drawback of the used approach is that, input should be in a constrained language, and this pitfall makes the approach impractical. This three phase framework is typically based on a set of SBVR business vocabulary and SBVR business rules to assist business people in creating clear and unambiguous business policies and rules in their native language. To demonstrate the potential of work, a small case study is discussed from the domain of office time management system with 68 sample elements classified into correct, incorrect and missing SBVR elements. Evaluation of case study showed that depending on recall and precision value F-value is calculated, encouraging future work.

In research paper [7] Annervaz K.M. & et al highlighted two issues of NL in business domain. Firstly, natural language is implicitly imprecise. The terms used by the Business Analyst (BA) in writing the requirement may not correctly capture the proper semantics of the actual domain terms. Such imprecise terms may give rise to ambiguity in the downstream activities (such as design). Secondly, since natural language [8], unlike a formal method, cannot force the notion of validation and completeness, the requirements often remain incompletely defined. Hence authors present a novel approach to perform requirement quality analysis with respect to a reusable domain model on the basis of ontology concept.

The methodology by Bourbakis-Manaris [9], based on SPNs [10] Modeling of the NL text sentences for Document Understanding. They describe four levels of processing: lexical to enforce case (subject verb) agreement, syntactic to combine words into sentences, semantic to assign meaning to words and sentences, and pragmatic to form context from relations to previous sentences, paragraphs, topics, and information from related data. This paper focuses on the more difficult syntactic and pragmatic process. Multiple modalities or external forms of information such as speech, images, text, video, gestures, facial expressions, hand signs, and handwriting are proposed to add to the context formed by the pragmatic process. The combination of ASGs(Augmented Semantic Grammars) and SPNs in this methodology provides significant capability in not only capturing semantic meaning from text but extracting contextual and other available information to resolve ambiguities. The methodology suggested in this paper shows how SPNs, used with ASGs, can model a tremendous amount of interrelationships that exists in both text and images. It provides significant potential for extended areas such as knowledge abstraction and representation and adding to their capabilities. The methodology presented in this paper also illustrates the potential for SPNs to model technologies in ways that significantly enhance their modeling capabilities compared to conventional approaches in using SPNs. The computational complexity however is high.

III. EXISTING METHODOLOGY

A. *Latent Semantic Analysis (LSA)*

Yeh et al (2008) presented the two methodologies, text relationship map and latent semantic analysis, that they used together for text summarization. In particular, the first methodology uses feature weights to create similar links between sentences forming a text relationship map. Sentence position (within a paragraph or document), keywords (that can add or negate), centrality, and resemblance to the title, together determine feature weights that contribute to sentence importance within the document. The authors also used

Latent semantic analysis (LSA) to extract and infer relations of words to their expected context. A sentence vs. word matrix analyzes use of words within context. Corpus-based information and scoring functions, use feature weights to trigger the creation of similar links between sentences, that are represented in a text relationship map (TRM), or graph. [11].This methodology captures various features that help in calculating the similarity of sentences throughout one or more documents.

Latent semantic analysis (LSA) is used for extracting and inferring relations of words with their expected context. The authors used it to derive latent structures from a document. They elaborate an LSA method that derives semantic representation and propose a method for generating a summary from a semantic representation. Four phases include:

(1) Pre-processes partitioned sentences using given punctuation and segment sentences into keywords using a toolkit called Auto Tag.

(2) Semantic model analysis uses a word-by-sentence matrix and produces a semantic matrix using singular value decomposition (SVD) and dimension reduction.

(3) Text relationship map is produced by the semantic matrix.

(4) Sentence selection uses the global bushy path from the text relationship map to select the important sentences that provides the summary.

In short the LSA approach uses a Word-Sentence matrix that can get very large due to the number of words in a document or in multi-documents.

B. Text summarization

Ko and Seo presented a hybrid sentence extraction methodology that uses some context information augmented with mainline statistical approaches to find important sentences in documents. Their model combines two consecutive sentences into a bi-gram pseudo sentence representation to overcome feature sparseness. By using traditional statistical methods, they calculate a score based on sentence similar to a query, location within a paragraph (first or last sentence, etc.), aggregation, and frequency of the same pseudo sentence. Each of these factors adds to the importance of the corresponding sentence by summing products of weights. A sliding window combines adjacent sentences to form a bigram. Once enough bi-gram representations are selected for a summary, each bi-gram is converted back to two sentences which are used in the resulting summary [12].Test results of the hybrid sentence extraction approach showed that it out performed other approaches listed by a small percentage. What the authors (of the hybrid approach) call context information is limited to two consecutive (i.e., adjacent) sentences with no global context capability implied. Normally, context would imply more extensive surrounding information than groups of two adjacent sentences.

C. Cluster based summarization

Methodology, Moens et al. (2005), extracts important sentences and detects redundant content across sentences. They used generic linguistic resources and statistical techniques to detect important content from topics and patterns of themes throughout text. From this, they build hierarchical topic trees from text. Then, they segmented topics and summarized at each level of topic detail. Their parser detects main grammatical constructs and finds semantic relations between content items. They use statistical techniques to cluster lexical and syntactic features of sentences, and then detect redundant content to generate summaries of multiple documents [13].

IV. IIMPLICATONS

This research will enable software developer to easily analyze the requirement of stakeholder by a single look at graph and understand what actually a user wish from developer. Also, work will emphasize on events in NLP by using firing concept of SPN to gain domain knowledge. Finally, color representation of sentences support to identify the ambiguity in complex sentences efficiently .Work will collect data from stakeholders for some specific domain and research will utilize it as a dictionary. Also, developer will create its own technical dictionary for this work

REFERENCES

[1] K.R. Chowdhary “Natural Language Processing” April 29, 2012

[2] Bures, T., Hnetynka et al. “Requirement Specifications Using Natural Languages” Technical Report D3S-TR-2012-05 December 2012.

[3] Agung Fatwanto “ Software Requirements Specification Analysis Using Natural language Processing Technique ” IEEE Quality in Research 2013

[4] A. Reed and B. Kellogg, Higher Lessons in English., 1877. (In Wikipedia: Sentence Diagram, en.wikipedia.org/wiki/Sentence_Diagram, last accessed: April 21st 2013).

[5] OMG. Semantics of Business voc bulary and Rules. (SBVR) Standard v1.0. Object Management

[6] Ashfa Umber & I. S. Bajwa “ Minimizing Ambiguity in Natural Language Software Requirements Specification” 2011 IEEE.

[7] Annervaz K.M., Vikrant Kaulgud & et. al “Natural Language Requirements Quality Analysis Based on Business Domain Models” ASE 2013, Palo Alto, USA New Ideas Track, 2013 IEEE

[8] M. Ilieva and O. Ormandjieva, “Automatic transition of natural language software requirements specification into formal presentation,” in Natural Language Processing and Information Systems. Springer, 2005, pp.392–397

[9] Bourbakis, N., Manaris, R “An SPN based Methodology for Document Understanding” IEEE nternational Conference on Tools for Artificial Intelligence, Tapei, Taiwan, 1998, pages 10-15.

[10] Haas, Peter J., Stochastic Petri Nets – Modelling, Stability,Simulation. Springer-Verlag New York, Inc 2002, ISBN 0-387-95445-7

[11] Yeh, J-Y., Ke, H-R., Y, W-P, Meng, I-H., Text summarization using a trainable summarizer and latent semantic analysis, Information Processing and Management, Vol. 41 (2005)pages 75-95.

[12] Ko, Y., Seo, J., An effective sentence-extraction technique using contextual information and statistical approaches for text summarization, Pattern Recognition Letters 29 (2008) pages1366-1371.

13] Moens, M.F, Angheluta, R., Dumortier J., Generic technologies for single- and multidocuments summarization, Information Processing and Management, Vol. 41 (2005) pages 569-586.

AUTHORS PROFILE

Mrs R.S. Ashtankar: She is received the B.E(Computer Technology) degree from RTM Nagpur, India, M.E(WCC) degree from G.H. Raisoni College of Engineering ,RTM Nagpur, University. Presently I am working as Assistant Professor, Department of CSE, ITM College of Engineering, Nagpur.

Research Area: Her interesting research field is Data Mining, Cloud Computing, and Natural Language Processessing. She is a author of one text book 'Data Warehousing and Mining" published for the student of Computer Science & Engineering.

Ms.W.M.Choudhari: She is received the B.E(Computer Technology) degree from RTM Nagpur, India, M.E(ESC) degree from G.H. Raisoni College of Engineering ,RTM Nagpur, University. Presently I am working as Assistant Professor, Department of IT, DMIETR , Wardha.

Research Area: Her interesting research field is Data Mining, Image Processing and Natural Language Processes sing