Textual Analysis by using Knowledge-based Word Sense Disambiguation Approach

Lilyana Jelai, Edwin Mit, Sarah Flora Samson Juan and Wai Shiang Cheah
Universiti Malaysia Sarawak, Sarawak, Malaysia.
16020121@siswa.unimas.my

***Abstract*—Textual analysis had been widely used in the software engineering area. Even though some approaches had been suggested over the time, these approaches encounter number of challenges, especially dealing with information extracted from the text requirement. Most studies had chosen to analyse the text manually in order to overcome this challenge. However, the long and complex text would consume more time. This paper will discuss a framework based on the knowledge-based word sense disambiguation approach, an attempt to improve the knowledge representation. In this approach, WordNet 2.1 would be used as the knowledge source used to identify concepts represented by each word in a text.**

***Index Terms*—Knowledge-based; Textual Analysis; WordNet; Word Sense Disambiguation.**

I. INTRODUCTION

Software requirement specification (SRS) understanding is required to establish a system because the requirement specification portrays complete system behavior. Several approaches had been introduced to aid in this matter such as textual analysis. Textual analysis had been proposed in various area to aid the software engineering tasks. One of the reason textual analysis had shown its effectiveness is because it involves tokenization and several approaches of lexical analysis [1]. Generally, textual analysis in software engineering consists of several steps. These steps, as mentioned in [1], including extraction of text documents from the corpus, indexing the corpus and compute the similarity between documents.

However, recent studies founded that textual analysis faced some challenges. For example, constructing the textual analysis techniques [1]. Even though information retrieval method had been used widely, the approaches are based on particular purpose, such as configuring the solutions, components, and configuration. Moreno et al. [2] overcome the text retrieval configuration problem by proposing a new approach which considers both the query and software corpus. The aim of this approach is to find out the best text retrieval configuration to be used by individual query based on software engineering task context [2].

This paper is to present the propose framework for textual analysis by using knowledge-based word sense disambiguation approach in textual analysis. One of the motivations of proposing this approach is to build knowledge representation from text requirement.

Knowledge-based word sense disambiguation exploits knowledge resources like dictionaries, ontology and thesauri to determine the sense of words in a context [3]. As this approach uses huge amount of structured knowledge, it may have the advantage of producing a better knowledge representation. This research will deploy the combination knowledge-based word sense disambiguation approach and the use of WordNet 2.1 to produce knowledge from text. One of the benefits is knowledge representation can assist to identify the possible meaning of a particular sentence.

The remainder of this paper is organised as follows: Section II reviews the knowledge resource. Section III discuses about related works. The proposed framework is covered in Section IV. Section V will discuss evaluation methods to validate the proposed framework. Last but not least, Section VI is the conclusion of this paper.

II. KNOWLEDGE RESOURCE

WordNet [4], [5], [6], is one of the most preferable lexical resources available today [7]. Unlike Longman's Dictionary (LDOCE) [8], the information organised in WordNet is completely different. For instance, noun senses are organised in a hierarchy and it covers more than 117 000 synonyms set (synsets) [5]. The database also contains features like part of link and antonym links [8].

Table 1 shows the number of words, synsets and senses in Wordnet 2.1 database statistics. This statistic is available online and can be retrieved from the WordNet web page.

Table 1
Number of words, synsets, and senses in WordNet 2.1 database statistics

Part-of-Speech	Unique Strings	Synsets	Total word-sense pairs
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Total	155327	117597	207016

Wordnet is adopted as knowledge resources in this study because it consists super-subordinate relation [9]. It links more general synsets to increasingly specific ones. As mentioned by Fellbaum [9], all noun hierarchies ultimately go up the root node.

The majority of the WordNet's relations connects words from the same part of speech (POS). Hence, WordNet consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers [6], [9]. Cross-POS relations include the morphosemantic links that hold among semantically similar words sharing a stem with the same meaning. In other words, Fellbaum [9] use the word observe (verb), observant (adjective) observation, observatory (nouns) as an example to show links between semantically similar words. These features would enrich the construction of knowledge later discussed in Section IV.

III. Related Works

This section would focus on architecture, text understanding and type of knowledge representation from related work. Currently, there is no authenticated architecture that gives precise results. The related work is detail expressed in the remaining part of the section.

Lami et al. [10] introduced an automatic tool for the analysis of natural language requirements which consists of Syntax Parser, Lexical Parser, Indicators Detector, View Derivator, Input and Output that enable to find defect in requirement text. Besides that, type and frequency of false positive can be obtained from the output. To understand certain word, this tool use Dictionaries inside the architecture. However, this tool does not apply knowledge representation. The analysis also depends on the completeness and the accuracy of the dictionary used.

Simov et al. [11] believed by adding context information improves the accuracy of knowledge-based word sense disambiguation. Therefore, an architecture which includes conversion of WordNet 2.0 to Wordnet 3.0 is produced to aid the enrichment of word sense disambiguation knowledge bases with context information [11]. The combination of various approaches of relations from WordNet had shown the improvement knowledge-based word sense disambiguation. For example, the addition of syntactic-based relations improves the results of knowledge-based word sense disambiguation. The knowledge representation was expressed in knowledge graph. However, the accuracy of knowledge produce depends on the integrity of the domain.

Ta and Thi [12] combined statistical method and natural language processing to extract semantic relation from text documents. The purpose is to identify the semantic relation might be found in text documents of the ACM Digital Library automatically. This approach can be divided into two main parts which are Computing Domain Ontology (CDO) and identifying the semantic relations among the instances of CDO using WordNet and other resources. The document used focusing on the computing domain [12]. The ontology produced able to display various semantic relations among the instances. However, not all information was fully extracted.

Hassan et al. [13] used semantic technology to annotate the text requirements expressed in a natural language. The purpose of this research is to determine the meaning of particular sentences [13]. In order to do so, Data Cleaning, Graph Construction, Sparse Matrix and Ontology Construction were included in the architecture. However, this research does not include any lexical tool to identify the meaning of each word.

Gaeta et al. [14] merely focuses on the construction of knowledge in form of ontology from heterogeneous text. This research introduces an architecture consists of Pre-processing, First ontology creation, Concept and relationship creation, Harmonization Refinement and Validation. Since the ontology produced consist of concept and relationship between words, therefore contextual understanding of certain text can be obtained.

IV. Proposed Framework

This section will discuss about the proposed framework for textual analysis by using knowledge-based word sense disambiguation. The proposed framework will adopt the WordNet 2.1 as the knowledge source.

Various type of text file such as PDF, TXT and DOC had been considered to be use in the proposed network. However, other formats used in this proposed framework in the near future. For this research purpose, text file used is software requirement specification documentation for Hospital Management System [15] and Lane Management System 2 [16]. Both files are in DOC format. Both files can be obtained from the web.

The structure of software requirement specification documentation for Hospital Management System [15] and Lane Management System 2 [16] used in this framework also should be defined. According to Gaeta et al. [14], assumptions on initial knowledge representation is possible if the document is structured.

The construction of knowledge in this module is adopting from [12] which is acquiring knowledge from documents. The differences between this study and [12] is this study used software text requirement while [12] used text documents of ACM Digital Library.

Figure 1 shows the proposed framework of textual analysis by using knowledge-based approach. Based on Figure 1, the proposed framework contains several modules. Each module exploits results from previous modules.

The expected output of the proposed framework as shown as the following: 1) knowledge representation from a text requirement and 2) the knowledge content conceptual relation between terms. The discussions of every module are in following section.

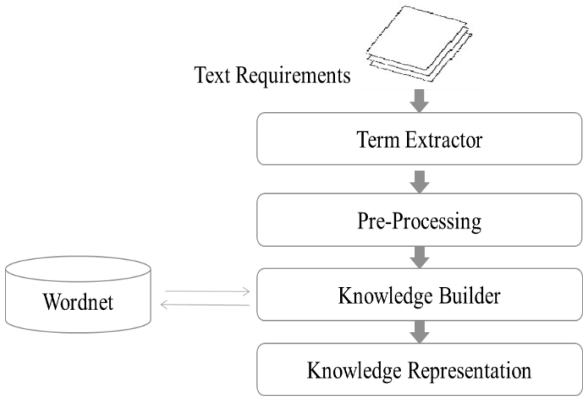

Figure 1: Proposed framework

A. *Term Extractor module.*

This module is to perform selection task of relevant term from the text. This module will consist set-of-term filter [14] adapting the method mentioned in [14]. According to Gaeta et al. [14], this step is important in order to avoid losing of document information structure. However, in order to identify the most relevant terms, appropriate algorithm should be considered. Therefore, this module is based on algorithms in [12].

B. *Pre-Processing module.*

In this phase, the document will undergo several sub phases described in the following:

i. Stemming: This sub phase is to reduce a term of the analysed document. Combination of algorithm can be implemented so that term can be reduced to its stem or root.

Example: reading → read

ii. Part-of-speech tagging (POST): This sub phase is to distinguish the terms in the document to a particular part of speech, such as names, verbs, etc. This sub phase is assumed to rely on the context where the term is found and the use of dictionary. This method adopted from Hwang et al. [17].
iii. Stopword list: Removing the unused terms that do not bring useful information to the domain.

C. Knowledge Builder module

The creation of knowledge starts here. Assuming the connection between words that are semantically annotated can be determined by sequences of particular word corresponding to the other words in a sentence. The presentation of this knowledge adapting the ontology characteristics as this knowledge is expected to use HasPart (HP) relation. The function of HP relation is to see the overall relationship of every words for one another. In order to construct the first relation, it adopts the concept of synonymy graph construction [18]. Hence, once the knowledge representation is available, WordNet 2.1 would be used in order to identify concepts represented by each words as it connect word from the same part of speech. Other advantages of using WordNet is it able to provide useful information about the semantic correlation between concepts [9], [19].

D. Knowledge Representation module.

This module is to present the complete knowledge of a text requirement. Based of Figure 2 the extracted knowledge is represented by a graph. The graph consists of relevant nodes identified from previous modules. As illustrated in Figure 2, each node represents term, T and C, C_2, and C_3 is the plausible conceptualization represented by T. Element S as in Figure 2 represent the senses obtained from WordNet 2.1.

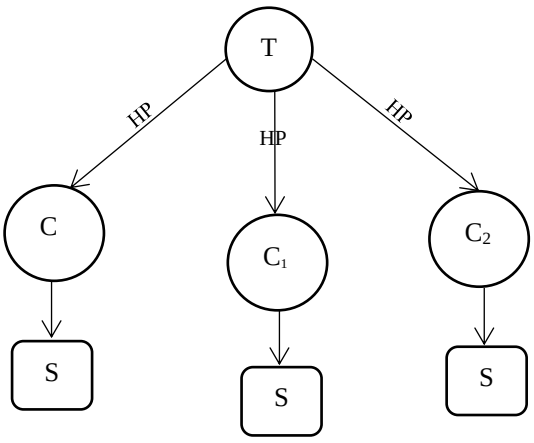

Figure 2: Simple HP relation for knowledge representation

However, Figure 2 shows knowledge representation of one term from the text requirement. The complete knowledge is expected to show wider knowledge with their respective concepts and senses. The main contribution of this research is it combine the textual analysis and word sense disambiguation to build knowledge from particular text requirement. Secondly, as the knowledge representation containing a conceptual relation between terms, it can assist in finding the meaning of a sentence, for example. Moreover, the knowledge representation in this approach is not domain dependent because the created knowledge is based on text requirement that had been extracted.

V. EVALUATION APPROACH

Since this research still in the early stage, the proposed evaluation to evaluate the proposed framework is by using precision and recall. Several researches had been using this method in order to validate the proposed framework. The evaluation is based on the proposed method in [14], [20]. Precision can be seen as measure of exactness and fidelity whereas recall is a measure of completeness.

In information retrieval, precision is a formula to measure relevant results. In general, precision is defined as the number of true positives over the number of true positives plus the number of false positives. The formula is as the following.

$$\text{Precision} = \frac{A}{A+C} * 100\% \quad (1)$$

Meanwhile, recall is a formula to measure of how many truly relevant results are returned. Recall is defined as the number of true positives over the number of true positives plus the number of false negatives. The formula is as the following.

$$\text{Recall} = \frac{A}{A+B} * 100\% \quad (2)$$

where: A = The number of relevant words retrieved
B = The number of relevant words not retrieved
C = The number of irrelevant words retrieved

Combining both *precision* and *recall* are also one of method to evaluate proposed framework. F-measure [14], [20], which is defined as the harmonic mean of precision and recall. Below is the formula for F-measure:

$$\text{F-measure} = \frac{2 * Recall * Precision}{Recall + Precision} \quad (3)$$

To set the benchmark of this study, Stanford CoreNLP [21] is used to extract instances and relations among instances from two case studies as mentioned in Table 2. According to Manning et al. [21] Stanford CoreNLP supports the API functions to develop the applications related to natural processing language.

Table 2 shows the experimental results for knowledge extraction of Hospital Management System [15] and Lane Management System 2 [16] by using StandfordCoreNLP.

Both case study shows impressive results in term of precision when using StandfordCoreNLP. However, in terms of recall, Lane Management System 2 [16] is slightly higher than Hospital Management System [15]. In other words, knowledge extraction of Lane Management System 2 [16] is more complete than Hospital Management System [15].

The F-measure assumes values in the interval is from 0 to 1. Therefore, 0.5 is chosen to underline the importance of precision over recall [20]. Based on F-measure formula, both case studies show encouraging results.

Lastly, scores obtained in Table 2 would be used as a benchmark for this study. Future studies would focus on comparing the results obtained from the proposed framework and Table 2 in term of precision, recall and f-measure

Table 2
Comparative Evaluation Method

Case Study	Precision (%)	Recall (%)	F-measure
Hospital Management System [15]	69.46	62.13	0.65
Lane Management System 2 [16]	62.37	73.75	0.67

VI. CONCLUSION

The proposed framework exploits the use of knowledge source and text requirement to produce knowledge representation. The knowledge representation could be used as knowledge-based to assist in finding the meaning of a particular sentence from the text requirement.

To test the framework, two papers had been chosen. Once the results are revealed, it would be compared to scores reported in Table 2. Therefore, scores in Table 2 would be used as a benchmark for this paper.

However, there is a constraint of this current approach. As this approach using WordNet to identify concepts represented by each word, identifying senses on each word would be difficult. Four different senses in WordNet can be hard to differentiate not just for computers, but also for humans. Due to that reason, not all senses may be relevant to disambiguate a word.

ACKNOWLEDGMENT

The authors would like to thank Universiti Malaysia Sarawak for providing support/facilities to conduct this research and reviewers for their comments to improve this paper.

REFERENCES

[1] G. Bavota, A. De Lucia, R. Oliveto, F. Palomba, and A. Panichella, "Textual analysis and software quality: challenges and opportunities," unpublished.

[2] L. Moreno, G. Bavota, S. Haiduc, M. D. Penta, R. Oliveto, B. Russo, and A. Marcus, "Query-based configuration of text retrieval solutions for software engineering tasks," in *Proc. 10th Joint Meeting on Foundations of Software Engineering Conf.*, 2015, pp. 567-578.

[3] N. Roberto, "A quick tour of word sense disambiguation, induction and related approaches," in *Proc. SOFSEM 2012: Theory and Practice of Computer Science Conf*, 2012, pp. 115-129.

[4] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and Wordnet-based approaches," in *Proc. Annual Conference of the North American Chapter of the ACL Conf.*, 2009, pp. 19-27

[5] C. F. Baker, and C. Fellbaum, "WordNet and FrameNet as complementary resources for annotation," in *Proc. 3rd Linguistic Annotation Workshop Conf.*, 2009, pp. 125-129.

[6] F. Fabbrini, M. Fusani, S. Gnesi and G. Lami, "The linguistic approach to the natural language requirements quality: benefit of the use of an automatic tool," in *Proc. 26th Annual NASA Goddard Software Engineering Workshop*, 2001, pp. 97-105.

[7] N. Kiyavitskaya, N. Zeni, L. Mich, and D. M. Berry, "Requirements for tools for ambiguity identification and measurement in natural language requirements specifications," *Requirements Engineering*, vol. 13, no. 3, pp. 207-239, Sept. 2008.

[8] K. Knight and S. K. Luk, "Building a large-scale knowledge base for machine translation," *AAAI*, vol. 94, pp. 773-778, Oct.1994.

[9] C. Fellbaum, *WordNet: The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd, Nov. 2012.

[10] G. Lami, S. Gnesi, F. Fabbrini, M. Fusani, and G. Trentanni, "An automatic tool for the analysis of natural language requirements," in *Informe técnico, CNR Information Science and Technology Institute*, Sept. 2009.

[11] K. Simov, P. Osenova and A. Popov, "Using context information for knowledge-based word sense disambiguation," in *International Conf. Artificial Intelligence: Methodology, Systems, and Applications*, Sept. 2016, pp. 130-139.

[12] C. D. Ta and T. P. Thi, "Automatic extraction of semantic relations from text documents," in *International Conf. Future Data and Security Engineering*, Nov. 2016, pp. 344-351.

[13] T. Hassan, S. Hassan, M.A. Yar and W. Younas, "Semantic analysis of natural language software requirement," in *6th International Conf. Innovative Computing Technology (INTECH)*, Aug. 2016, pp. 459-463.

[14] M. Gaeta, F. Orciuoli, S. Paolozzi, and S. Salerno, "Ontology extraction for knowledge reuse: The e-learning perspective," *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 44, no. 4, pp. 798-809, Jul. 2011.

[15] S. Prasad, "SRS documentation for Hospital Management System," Sept 2012. Available at http://www.freestudentprojects.com/ studentprojectreport/project-srs/srs-documentation-for-hospital-management-system/

[16] C. Dinsmoor, M. Mei, B. Wong, A. Agrawal, G. Levene, "Software Requirements Specification (SRS) Lane Management System 2," 2016, Available at https://pdfs.semanticscholar.org/ce98/ b1bada1186e7b09eeca764dff69c64dee0db.pdf

[17] M. Hwang, C. Choi and P. K. Kim, "Automatic enrichment of semantic relation network and its application to word sense disambiguation," *IEEE Trans. Knowledge and Data Engineering*, vol. 23, no. 6, pp. 845-58, Jun 2011.

[18] Y. Shin, Y. Ahn, H. Kim and S. G. Lee, "Exploiting synonymy to measure semantic similarity of sentences," in *Proc. 9th International Conference on Ubiquitous Information Management and Communication*, Jan 2015, pp. 40.

[19] C. Fellbaum, "Wordnet(s)" in *Encyclopedia of Language & Linguistics*, 2nd ed. vol. 13, *Keith Brown*. Oxford: Elsevier, 2006, pp. 665-670.

[20] J. Ma, W. Xu, Y. H. Sun, E. Turban, S. Wang and O. Liu, "An ontology-based text-mining method to cluster proposals for research project selection," *IEEE Trans. Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 3, pp. 784-90, May 2012.

[21] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations Conf.*, Jun 2014, pp. 55-60.

[22] A. Arellano, E. Carney and M. A. Austin, "Natural language processing of textual requirements," in *Proc.10th International Conference on Systems (ICONS 2015)*, Apr. 2015, pp. 93-97.

[23] G. Génova, J. M. Fuentes, J. Llorens, O. Hurtado and V. Moreno, "A framework to measure and improve the quality of textual requirements," *Requirements Engineering*, vol. 18, no. 1, pp. 25-41, Mar. 2013.

[24] W. Lu, Y. Cai, X. Che and Y. Lu, "Joint semantic similarity assessment with raw corpus and structured ontology for semantic-oriented service discovery," *Personal and Ubiquitous Computing*, vol. 20, no. 3, pp. 311-323, Jun. 2016.

[25] F. Sclano and F. Velardi,"Term extractor: A web application to learn the shared terminology of emergent web communities," in *Proc. 3rd International Conference I-ESA*, Mar. 2007, pp. 289-290.