Thematic Relations Based QA Generator for Sinhala

T Dissanayake[1] and B Hettige[#2]

[1]Information Technology Unit, Faculty of Humanities and Social Sciences University of Sri Jayewardenepura,
[2]Department of Computer Engineering, General Sir John Kotelawala Defence University

[#] budditha@kdu.ac.lk

Abstract: Question-Answer Generation is one of the research areas in the field of Natural Language Processing. This paper presents how the Sinhala question-answer generator operates through the thematic relation on the given input sentence. This Sinhala question generator operates on an input of a well-defined Sinhala sentence with the subject, object, and verb along with some specifically identified POS tags for the system like Nouns, Pronouns, Adjectives, Verbs and Adverbs which can make noun phrases and verb phrases from all the possible combinations. The system uses a Sinhala tokenizer, Sinhala POS tagger, Chunker, and the semantic relationship identifier with the support of the NLTK libraries. Through the syntax analysis, it can identify thematic relations for the Sinhala input sentence. Through these semantics relations, it has the ability to generate Sinhala questions and related answers through the rule-based approach. The present system has been tested with 56 sample sentences. According to the evaluation, the system shows a 93% correctness rate.

Key Words: Question Generation, Sinhala, Thematic Relations, Natural Language Processing

Introduction

Natural Language Processing is the technology used to assist computers to understand the human's natural language. It is a branch of artificial intelligence that represents the interaction between computers and humans using the natural language. Natural language processing can be used to read, decipher, understand, and make sense of the human languages in a valuable manner. In the context of education, NLP can assist teachers, students, authors, and educators for assisting in writing, analysis, and assessment procedures. Researches have been conducted on this discipline to generate smart question generation systems like ArikIturri (Intelligent question generator based on corpora and NLP techniques) (Itziar Aldabe, 2006), Intelligent factual question generation from text to generate factual WH questions (Heilman, 2011), Intelligent Question Generation using Relative pronouns and adverbs and Question generation using NLP to solve the inverse task (PayalKhullar, 2018). Natural language processing can be integrated with a large number of educational contexts. In the present scenario, teachers and any other academics would spend a lot of time manually preparing question papers, quizzes, and other related assessment materials. Similarly, students have no automated process for self-calibration. But by automating this calibre can save a lot of time on lecturers as well as students. Particularly this research can help in calibrating yourself and remove any dependencies on mentors. Thematic relations based QA generator for the Sinhala language can be used to assess students in theory-based Sinhala language subjects which can ask questions based on sentences and provide answers using sentences. The

Sinhala Language is one of the Indo Aryan family languages and it is the spoken and written language of the majority of Sri Lankans. Most of the already developed research-based methodologies are based upon English but generating a question generator from a different language is challenging because new languages would likely present new challenges for Question Generation as different languages have different syntactic constraints that form different realizations than in English. (Heilman, 2011) This paper presents thematic relations based QA generator for Sinhala which is developed to input a sentence in Sinhala and then the system extracts the thematic relations in the texts independently, analyses, and generates questions. The rest of this paper is organized as follows. Section 2 describes related works with an overview of some existing question generation systems. Then section 3 reports the Thematic relations based Sinhala Question Generator's design and implementation. Section 4 reports the results and discussion. Finally, Section 5 concludes the paper with a note on the conclusion and further works.

Related Works

Question generation has been a task which has got slightly less attention in the research domain specially for languages like Sinhala which is the written and spoken language of Sri Lanka. Despite the language type, similar set of steps can be incorporated to generate question from an inputted text as per the Fig. 1 A tokenizer breaks a stream of text into tokens, usually by looking for whitespace (tabs, spaces, newlines). Tokenization is the process of demarcating and possibly classifying sections of a string of input characters (Anon., n.d.). The resulting tokens are then transferred on to some other form of processing. The process can be considered a sub-task of parsing input (Anon., n.d.). POS tagging is the intelligent assignment of

descriptions to the tokens. Descriptions can be called tags, which could be one of the parts of speech tags (Anon., n.d.). So POS tagging is the assignment of the appropriate part of speech tags to the tokens of a sentence.

Descriptive details (mean, standard deviation, median, mode, minimum and maximum) of mother's age, mother's BMI, newborn's birth weight and head circumference are shown in table 1. In reference to the descriptive statistics, the sample consisted of three months postpartum mothers between ages 18 to 42 years.
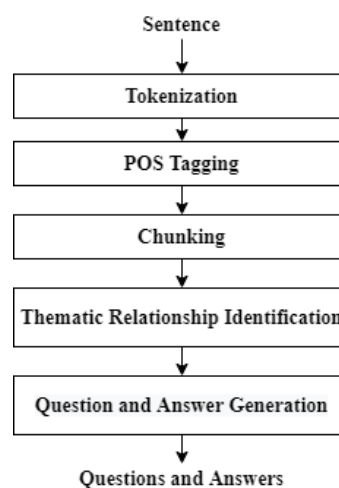

Figure 1. Question Generation process

Text chunking can also be defined as shallow parsing which follows POS tags and adds more structure to the sentence. The outcome of this is the grouping of the words in a sentence into "Chunks" (Anon., n.d.). Chunking is done using an identified chunk grammar. The thematic relations also called semantic roles of a language can be defined as various roles that a particular noun phrase play with the corresponding verb about the action or state it plays, frequently the main verb of the sentence (Anon., 2020) The concept of thematic roles has been presented to explain the syntax of a particular sentence as wells as the meaning it carries. It is presented with the assumption that the pivotal factor in the formation of a sentence's meaning is the role that sentence played by

each noun phrase with the verb (Anon., n.d.). There are several approaches and methods developed and proposed for the task of question generation. The literature of the thematic relations based question generator is reviewed based on the topics Transformation based Question generation which has sub arenas like templatebased which uses matching approach of Natural Language Processing (Andrea Andrenucci, 2005), syntax-based and semantics-based which uses a syntactic tree of the inputted or given sentence to convert it to an interrogative form and logical representation of the declarative sentence to make an agglomerative sentence respectively (Michael Heilman, 2009), Ontology-based Question Generation system called OntoQue, which is an engine for objective assessment item generation based on domain ontologies which are known as knowledge representation structures which provides a conceptual model of a domain (Al-Yahya, 2011), Question generation based on Discourse Cues that can intelligently generate questions from natural language sentences using discourse connectives. with discourse connectivity, the system can be able to generate questions that look at the problem beyond the sentence level which divides question Generation into a content selection and question formation. (Manish Agarwal, 2011), Question Generation based on Dependency parses information that provides a view of a sentence provided by a parser of a given type, such as phrase structure trees or predicate-argument structure. Which presents a solution using multiple views from different parsers to create a tree structure which represents items of interest for question generation. (Anon., 2015) Topic to Question Generation and Domain-Specific concepts for Question Generation paper addresses of generating all possible questions from a preferred topic and this system considers that each topic is associated with a body of texts containing

useful information about the topic. Questions are generated by exploiting the named entity information and predicate-argument structures of the sentences present in the body of texts (Yllias Chali, 2014). Though the usage of Sinhala language is minimal in the area of researches some Natural language processing and machine learning-based systems also have been developed to a favorable extent like Ananya - A Named-EntityRecognition (NER) System for the Sinhala Language discusses data-driven techniques to detect Named Entities in Sinhala texts, with the use of Conditional Random Fields (CRF) and Maximum Entropy (ME) statistical modeling methods (S.A.P.M. Manamini, 2016).A Parser for the Sinhala Language - First Step Towards English to Sinhala Machine Translation systems comprised of a Sinhala Morphological analyzer that connects three dictionaries called the base dictionary, rule dictionary, and concept dictionary. The morphological analyzer in this system implements implication and derivations rules of the Sinhala Grammar and gives grammatical information of the words which need to be considered by the parser (B. Hettigel, 2006) Hidden Markov Model-Based Part of Speech Tagger for Sinhala Language addressed the importance of a Post tagger as it is a vital component in a natural languagerelated task which includes the analysis of the construction, behavior, and dynamics of the language which can be utilized in both the analysis of computational linguistics and automation applications (A.J.P.M.P. Jayaweera, 2014) Mahoshadha", The Sinhala Tagged Corpus-based Question Answering System addressed a novel architecture for question answering which uses basic two steps. The first one is summarizing a tagged corpus and the second one is to use the summarization to generate an answer for a query (J.A.T.K. Jayakody, 2016). Though the approaches are introduced for several tasks in question generation, developing a model

that can generate questions from a Sinhala sentence based on thematic relations which was not addressed in any of the above approaches as per the best of my knowledge.

Design and Implementation

This chapter describes the design and implementation process of the thematic relations based on Sinhala Question Generator. The system is composed of basic modules likes POS taggers, tokenizers, chunkers, and semantic relationship identifiers. The overall design of the system is composed of the basic seven processes namely Tokenization, POS Tagging, Chunking, Thematic Relation Identification, Question Generation, and Answer Generation. To perform all these processes systems have Figure 1. Question Generation process required corpus, POS tag sets, and chunk grammar as extra mechanisms. The design and implementation of the intelligent Sinhala question generator have been developed based on the following processes and modules. A.Tokenization using a tokenizer The System uses a tokenizer to split the words punctuation marks and spaces of an inputted Sinhala sentence to a list of tokens to identify each word and space in the sentence. It is developed with the default NLTK library's Sinhala corpus. To carry out the tokenizing, the system uses "word_tokenize" from the NLTK Tokenize library. from nltk.tokenize import word_tokenize Tokenization has performed as in the following code EXAMPLE_TEXT = "Sample Sinhala Text." # Tokenize Defined Sentence Or Paragraph print(word_tokenize(EXAMPLE_TEXT))

Example: EXAMPLE_TEXT = "පා ෙගාඩි ළමයො ඇයජේ මවට ෙපපමන් කතොකරයි."

print(word_tokenize(EXAMPLE_TEXT))
Output: ['පා ෙගාඩි', 'ළමයො', 'ඇයජේ', "මවට", "ෙපපමන්", "කතොකරයි", "."]
B.POS Tagging

The Intelligent Sinhala Question generator uses a rulebased part of the speech tag system. A set of pre-defined parts of speech tags are used in the POS tag set. POS tag set is created with these POS tags specifically for this system by the author. Designed POS tag set includes Common nouns, Proper Nouns, Pronouns, Adjectives, Adverbs, and Verbs which are derived by a set of sample sentences initially used to test and develop the system. Common nouns and Proper nouns have collectively defined as Nouns in the designed POS tag set. The POS tag set is designed manually using a database table with MYSQL Libraries and SQL Queries by the author which has developed using a sample of 100 words

C. Chunking using Chunk Grammar This chunk grouping is done in the Sinhala question generation system by identifying a common chunk grammar after testing several set of chunk grammar rules to identify a common grammar which could work for all the sentences that can be made from the identified 8 patterns. Chunking is used to find phrases from a sentence like noun phrases, verb phrases, adverbial Phrases, and adjective phrases. This Intelligent Sinhala Question generation system uses Noun phrases and Verb phrases using chunk grammar to develop a parse tree. The following are the possible combinations of POS tags which can be identified as noun phrases and verb phrases according to the valid input Sinhala sentences Noun Phrase - [Noun], [Pronoun, Noun], [Pronoun], [Adjective, Noun] Verb Phrase - [Verb], [Adverb, Verb] As the sentence contains a Subject, Object, and Verb. Both Subject and Object can be considered as noun phrases. The chunking has been done using NLTK Text chunker with the use of identified chunk grammar. The system has done chunking for Noun phrasing and Verb Phrasing. The following grammar shows the finalized chunk grammar used for all

combinations For Noun phrases: NP: {||? |?} For Verb Phrases: VP: {|?} Overall Chunk Grammar: NP: {||? |?} VP: {|} The system uses 8 basic sentence patterns to generate questions as per table 1 as follows. These 8 sentence patterns have been obtained by taking logical possibilities of a sentence of three parts (Subject, object, verb) that is 2 3=8

Table 1. 8 basic sentence patterns

No	Subject		Object		Verb	
1	Adjective	Noun	Pronoun	Noun	Adverb	Verb
2	Adjective	Noun	Pronoun	Noun	-	Verb
3	Pronoun	Noun	-	Noun	Adverb	Verb
4	Pronoun	-	Pronoun	Noun	Adverb	Verb
5	Pronoun	Noun	-	Noun	-	Verb
6	-	Noun	Adjective	Noun	-	Verb
7	-	Noun	-	Noun	Adverb	Verb
8	-	Noun	-	Noun	-	Verb

The following sentences show examples for the 8 combinations of patterns as per the table 1

1. ප ොඩි ළමයො ඇයෙජේ මවට පෙපමන් කතොකරයි.

 2. ආදර්ශවත් ගුරුතුමො ඔහුෙජේ මවට ෙළකයි

3. ඇයෙජේ පියො ප ොත පෙපමන් කියවයි

4. මම මෙජේ ප ොත පෙපමන් කියවයි

5. ඔහුෙජේ මව න්ෙලට යයි

6. ළමයො සුන්දර ගම්ෙමොනයට යයි

7. අම්ෙමො කතන්දරය මිහිරිව කියයි

8. අම්ෙමො කතන්දරය කියයි

No Subject Object Verb 1 Adjective Noun Pronoun Noun Adverb Verb 2 Adjective Noun Pronoun Noun - Verb 3 Pronoun Noun - Noun Adverb Verb 4 Pronoun - Pronoun Noun Adverb Verb 5 Pronoun Noun - Noun - Verb 6 - Noun Adjective Noun - Verb 7 - Noun - Noun Adverb Verb 8 - Noun - Noun - Verb Table 1. 8 basic sentence patterns The Following parse trees (Fig. 2 and Fig. 3) shows possible noun phrase and verb phrase combinations of chunks identified for some patterns as examples.
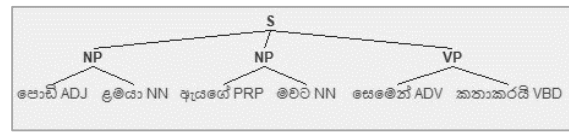

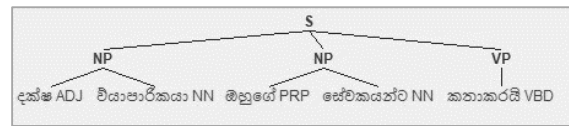*Figure 2. Chunking of the first pattern as per table 1*

Figure 3. Chunking of the second pattern as per table 1

D. Question Generation using thematic relations Thematic relations based Question Generator for Sinhala uses a Semantic relationship identifier model to generate questions that were designed as a novel approach by the author to identify thematic relationships between two words in noun phrases and verb phrases. As the system generates questions based upon inputted sentences it takes sentences that have basic three parts namely subject, object, and verb. These three parts produce noun phrases and verb phrases according to the POS tags.

Thematic relations based Question Generation System's semantic relationship identifier has identified basic three formations of Noun Phrases and Verb phrases and the system has analyzed and identified common relationships for each phrase. Relationships identified for noun phrases are, noun phrases with "Pronoun" and "Noun" as first and second words always

represent the quality "අයිතිය" which denotes the ownership of the noun to the pronoun. Noun phrases with "Adjective" and "Noun" as first and second words always represent the quality "ගුණය" which denotes the qualities of the noun based on the adjective. Verb phrases with "Adverb" and "Verb" as first and second words always represent the quality "ආකෝරය" which denotes the way a particular verb acts or reacts. These conclusions were derived by analyzing a sample set of noun phrases and verb phrases. The following tables (table 2 and 3 show the examples)

Table 2. Identified relationships for noun phrases

Noun	Quality	Attribute	Noun Phrase
අම්මො	අයිතිය	මෑජේ	මෑජේ අම්මො(Pronoun,Noun)
න්ෙල	ගුණය	ජ්රඩොන	ජ්රඩොන න්ෙල (Adjective,Noun)

Table 3. Identified relationships for verb phrases

Verb	Quality	Attribute	Verb Phrase
යයි	ආකෝරය	පෙපමන්	පෙපමන් යයි (Adverb, Verb)

Questions have been generated using a semantic relationship identifier which was developed using MySQL and SQL Queries. Here, the noun phrases obtained from the parse tree are sent to three separate tables called NounConverter, VerbConverter, and SentencePattern. Then system writes identified Noun phrases to the NounConverter table, then noun phrase is assigned the appropriate semantic relations specified according to the rules. Identified verb phrases are also written to the Verb Converter table with the specified Semantic relationship by the system. Then the overall sentence is identified using the pre-defined pattern and written to the Sentence pattern table. Then through the identified sentence

pattern system generate a set of questions. Several questions can vary according to the number of words, various POS tags, and phrases. Answers are derived according to the created questions using the sentence patterns. The following examples show how questions are generated for the pattern one. Pattern 01 Example: ප ෝඩ් ළමයො ඇයජේ මවට පෙපමන් කතොකරයි ([Adjective,Noun],[Pronoun,Noun],[Adverb, Verb])

Figure 4. NounConverter Table for Pattern 01

id	noun	quality	attribute
1	ළමියා	ගුණය	ෝපාඩ්
2	මවට	අයිතිය	ඇයෝගේ

Figure 5. VerbConverter Table for Pattern 01

The following table (Table 04) represents set of questions and answers which generates based on the sentence pattern 01 as in the above example (ප ෝඩ් ළමයො ඇයෝජේ මවට පෙපමන් කතොකරයි) and figure 6 shows how questions generate in the real system.
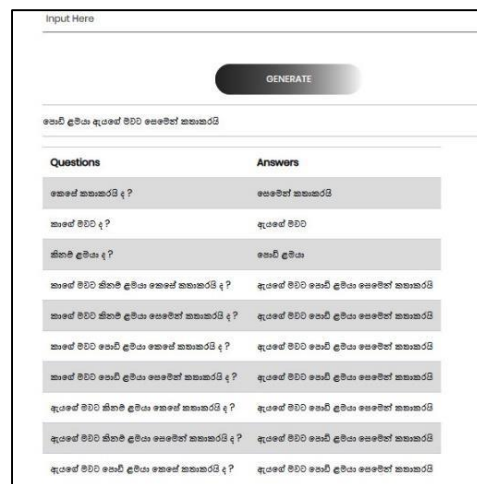

Figure 6. Screenshot of the system

Table 4. Generated Questions and Answers

Questions	Answers
පකෝජේ කියයි ද?	මිහිරිව කියයි
කොපේ ගුරුතුමෝ ද?	අපජ් ගුරුතුමෝ
කිනම් කතන්දරය ද?	ආදර්ශවත් කතන්දරය

කොපේ ගුරුතුමෝ කිනම් කතන්දරය පකපේ කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
කොපේ ගුරුතුමෝ කිනම් කතන්දරය මිහිරිව කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
කොපේ ගුරුතුමෝ ආදර්ශවත් කතන්දරය පකපේ කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
කොපේ ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
අපප් ගුරුතුමෝ කිනම් කතන්දරය පකපේ කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
අපප් ගුරුතුමෝ කිනම් කතන්දරය මිහිරිව කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි
අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය පකපේ කියයි ද ?	අපප් ගුරුතුමෝ ආදර්ශවත් කතන්දරය මිහිරිව කියයි

Results and Discussion

Black box testing has been conducted to test the system in the viewpoint of a user to check whether the system functions properly and provides expected outcomes concerning the execution. The finalized system of the Thematic relations based Sinhala question generation system can be performed by using a set of test cases and then these test cases have been tested by providing inputs to the system's input screen. Performance testing for the Thematic relations based Sinhala Question generator is carried out with the help of several end users with the use of a questionnaire and by giving instructions to the users to use the system in realtime to identify the quality of the questions it generates whether it has adhered to the expected requirement or not. So the final system is given to few users according to the selected sample space. The sample space for the evaluation was derived by analyzing the application of the system. Therefore, as the system can be focused a lot for the learning purposes sample space was obtained using a population of IT professionals, University Lecturers, and

University Students. So the sample space system has used 15 university students,5 university lecturers, and 10 IT professionals which derived a performance level of 53% excellent, 30% good, 17% neutral, and 0% of bad performance. Accuracy testing was conducted by obtaining a pile of valid sample questions and checked for accuracy following the generated outcome. Therefore, it has been done using 56 sample sentences. These 56 sample sentences are created following the defined validity of the system and the sentences are being fed to the system and obtained the outcome to check whether it can generate the expected patterns of questions and answers according to the defined rule-based question patterns. Here, the system has used 56 sample sentences and from that 52 samples Sinhala sentences have been generated correctly and 4 sample Sinhala sentences have been identified as incorrect which concluded the 93% correctness rate.

Table 5. Test case sample

Sentence	Valid Parse Tree	Total QNA	Valid CNA	Invalid QNA	Final Assumption
මල්ලි වඳුරන් එලවයි	T	0	0	0	T
මයිනෝ කුඩො පොජටන් නුවන් ඩැගනී	T	5	5	0	T
අම්මො රූ වොහිනිය ඔනැකමින් නරබයි	T	2	2	0	T
මෙනලගයො මල්ල පරොන් උරා පබොයි	T	5	5	0	T
නර්තන ශිල්ලපිනිය චිත්‍ර ටපේ රග ගොයි	T	2	2	0	T

කුරුල්ලන් නිද පේ පියඔයි	T	0	0	0	T
මිනිංංො අම්බලපම් ගිමන් රි	T	2	2	0	T
ගොයකයො ංංගීත ංංදර්ශන පේ ගීත ගයයි	T	5	5	0	T
ංෙද රොත්රි අ පේ දිපලයි	T	2	0	2	T
තොත්තො නිවපේ ව ලය ංංොංංෙදයි	T	2	2	0	T

Figure 7. Test Results (Accuracy Testing)

Figure 8. Test Results (Performance Testing)

Conclusion and Further Works

Thematic relations based question generator which is developed to generate questions based on input by processing it through several stages to generate questions and answers. To achieve this system has used a semantic relationship identifier which can generate questions based on 8 identified patterns for a Sinhala sentence with the basic subject, object, and verb with the additional POS tags called pronouns, adjectives, adverbs. Limitations of the research have identified currently developed mechanism cannot generate questions for a sentence with 3 words that is just the basic sentence with a subject, object, and a verb. With further enhancements, more varieties of sentences can be covered with the proposed method. This project has played a foundation for various projects in Sinhala Language Question generation. Several significant areas of further works are the developed Sinhala question generator only accepts sentences with a subject, object, and verb with different other supportive POS tags as described. This system can further enhance to generate questions for complex sentences. Furthermore, it generates questions from inputted paragraphs and documentations with extensions like .doc, .pdf, .txt which can be conceded as a further enhancement of the system.
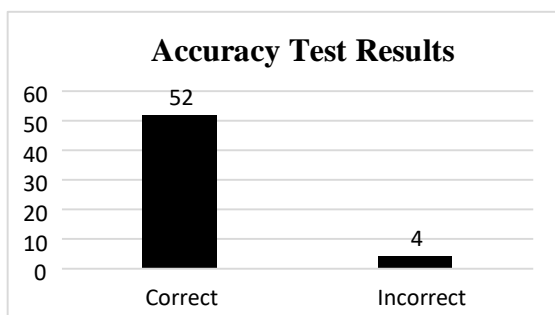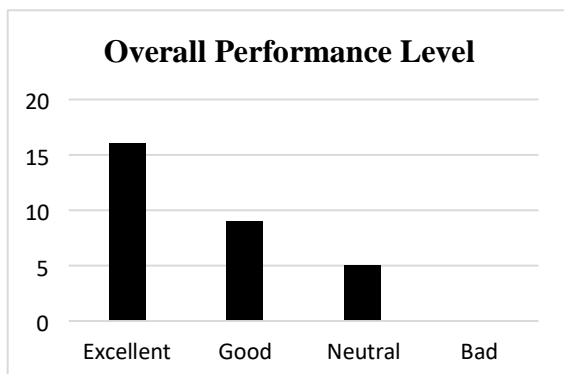
References

A.J.P.M.P. Jayaweera, N. D., 2014. HIDDEN MARKOV MODEL BASED PART OF SPEECH TAGGER FOR SINHALA LANGUAGE. International Journal on Natural Language Computing (IJNLC) , 3(3), p. 15.

Al-Yahya, M., 2011. OntoQue: A Question Generation Engine for Educational Assesment Based on Domain Ontologies. s.l., s.n.

Andrea Andrenucci, E. S., 2005. Automated Question Answering: Review of the Main Approaches. s.l., s.n.

Anon., 2015. Leveraging Multiple Views of Text for Automatic Question Generation. Springer International Publishing Switzerland, p. 10. Anon., 2020.

Wikipedia. [Online] Available at: https://en.wikipedia.org/wiki/Thematic_relatio n [Accessed 03 05 2020].

Anon., n.d. ello.us.de. [Online] Available at: http://www.ello.uos.de/field.php/Semantics/Se manticsThematicr elations [Accessed 03 05 2020].

Anon., n.d. Guru99. [Online] Available at: https://www.guru99.com/pos-tagging-chunkingnltk.html [Accessed 13 03 2020].

Anon., n.d. Stack Overflow. [Online] Available at: https://stackoverflow.com/questions/380455/looking-for-a-cleardefinition-of-what-a-tokenizer-parser-and-lexers-are [Accessed 12 03 2020].

Anon., n.d. Tutorials Point. [Online] Available at: https://www.tutorialspoint.com/natural_language_processing/nat ural_language_processing_part_of_speech_tagging.htm [Accessed 12 03 2020].

B. Hettige, A. S. Karunananda, 2006. A Parser for Sinhala Language - First Step Towards English to Sinhala Machine Translation. Colombo, Sri Lanka, s.n.

Heilman, M., 2011. Automatic Factual Question Generation from Text, s.l.: s.n.

Itziar Aldabe, M. L. d. L. M. M. E. M., 2006. ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. Taiwan, s.n.

J.A.T.K. Jayakody, T. G. W. K. P. A. N. Y. M., 2016. "Mahoshadha", The Sinhala Tagged Corpus based Question Answering System. p. 11

Manish Agarwal, R. S. M., 2011. Automatic Question Generation using Discourse Cues. Portland,Oregan, s.n.

Michael Heilman, N. A. S., 2009. Question Generation via Overgenerating Transformations and Ranking. Language Technologies Institute School of Computer Science Carnegie Mellon Universit, p. 16.

PayalKhullar, K. M. M., 2018. Automatic Question Generation using Relative Pronouns and Adverbs. Melbourne, s.n.

S.A.P.M. Manamini, A. A. R. R. G. R. S. J. G. D. S. R., 2016. Ananya - A Named-Entity-Recognition (NER) System for Sinhala Language. s.l., s.n.

Yllias Chali, S. A. H., 2014. Towards Topic-to-Question Generation. Association for Computational Linguistics, 41(1), p. 20