Sri Lanka Institute of Information Technology

QAG - Question and Answer Generation System from Course Materials

Project Proposal Report

Project ID: 17-062

Submitted by:

1. IT14058424 – A.S.M Nibras
2. IT14033506 – M.F.F Mohamed
3. IT14121852 – I.S.M Arham
4. IT13001162 – A.M.M Mafaris

BSc. Special (Honors) Degree in Information Technology

Submitted on 17-03-2017

DECLARATION

We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Student Name	Registration No	Signature
A.S.M Nibras	IT14058424	
M.F.F Mohamed	IT14033506	
I.S.M Arham	IT14121852	
A.M.M Mafaris	IT13001162	

Supervisor

……………………………..

Ms. Anjalie Gamage

ABSTRACT

This project is intended to create an automation of generating questions and answers based on school text books of grade 6 History subject in Sri Lanka. Currently there are some existing systems which are providing questions and answers based on some text materials from the internet. There are no any systems so far developed targeting school teachers in Sri Lanka to prepare question paper and answer scheme. This research is all about full filling above mentioned gaps in the existing systems.

This system contains four main areas namely, extract the information (Lecture Contents) from the Portable Document Format(PDF) or word document and Categorizing the text book content according to the topic and store in the database. Then Filtering and identifying the content where the questions can be generated to test the knowledge. From each paragraph of each chapter, important sentences have to be extracted. Unnecessary contents will be filtered out from the paragraph. Then Identifying key phrases and words that should be used to generate appropriate questions. Finally Forming proper questions and suitable answers using the identified key phrases from the context. Generated questions will be stored in the question bank. Teacher has their preference to select question for exam paper.

Throughout from this system, teacher’s will get many benefits. They can easily prepare both questions paper and answer scheme. So far teacher’s preparing question paper and answer scheme manually. Along with that, teacher can customize the paper by their selves. They can allocate marks based on the weight of the question. The whole research idea is to provide better, more efficient and useful system for preparing question paper and answer scheme automatically.

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1 Background 1
1.2 Literature Review 1
1.2.1 Revup: Automatically generating questions from educational texts 1
1.2.2 Question Generation as a Competitive Undergraduate Course Project 2
1.2.3 Automatic Question Generator in Tamil 3
1.3 Research Gap and Research Problems 5
1.3.1 Research Problems 5
1.3.2 Research Gap 6
2. OBJECTIVES 7
2.1 Main Objectives 7
2.2 Specific Objectives 7
3. RESEARCH METHODOLOGY 9
3.1 Categorizing the text book content according to the topic/ chapter and storing in the database 10
3.2 Filtering and identifying the content where the questions can be generated........ 12
3.3 Identifying key phrases and words that should be used to generate appropriate questions. 13
3.4 Forming proper questions using the identified key phrases and words from the contexts. 15
4. DESCRIPTION OF PERSONAL AND FACILITIES 18
5. REFERENCES 20

1. INTRODUCTION

This system is mainly targeting school teachers who going to prepare question paper and answer scheme automatically. The reason why we chose to develop this kind of system is, nowadays school teachers have their busy schedule. Most of the teachers are taught to many classes. When the period of school examination, preparing question paper is tedious task for teachers. Some teachers preparing question paper for each chapter after it is completed at classroom. Also, teacher has to allocate marks for questions manually according to the weight of the question. This system considered all these problems to build better solution. For this research, we consider the factoid type questions only. A factoid question can be any of these types: “What...”, “Where...”, “When...”, “Who...”, and “How many / How much...”. The questions are straight forward from the text materials.

1.1 Background

1.2 Literature Review

There are some successful Question Generation projects done by many researches. The main flow of all projects is analyzing the text materials and generating different type of questions such as essay type questions, Multiple Choice Question(MCQ), gap-fill multiple choice questions etc.

1.2.1 Revup: Automatically generating questions from educational texts

RevUp is Automatically generates gap-fill multiple choice questions from online text. The system analyzes online text and finds most important sentences, then select the main gap-phrases from the selected sentences and choose distractors that are semantically and syntactically similar to the gap-phrase and have contextual fit to the gap-fill question [2].

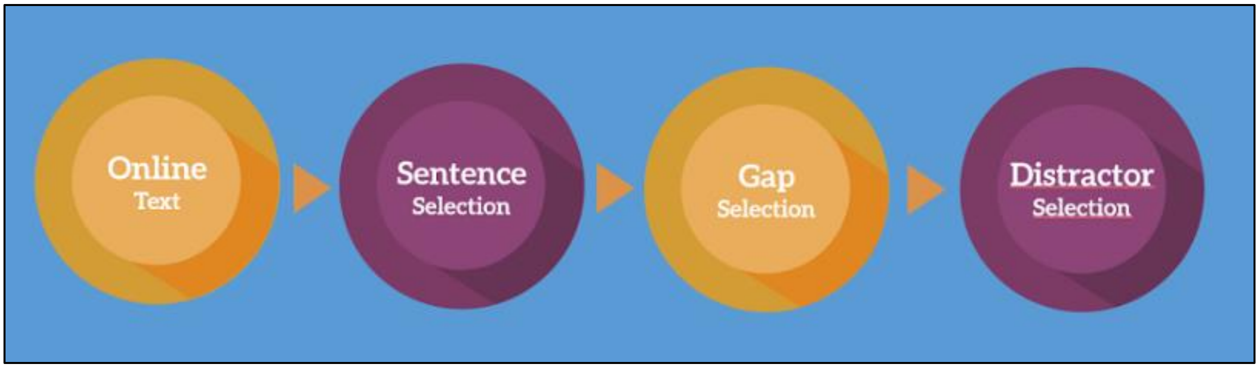


Figure 01: Core Methodology

- **Sentence Selection**

  They select the topically important sentences by ranking them based on topic distributions obtained from a topic model. And they used the Deep Autoencoder Topic Model (DATM) methodology, which discovers topics that are more coherent than the widely-used Latent Dirichlet Allocation and Latent Semantic Analysis.

- **Gap Selection**

  To select gap-phrases from each selected sentence, they collected human annotations, using the Amazon MTurk, on the relative relevance of candidate gaps. This data is used to train discriminative classifier to predict the educational relevance of gaps, they achieved an accuracy of 81.0%.

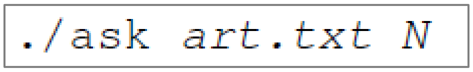
- **Distractor Selection**

  They proposed a novel method to choose distractors that are semantically and syntactically similar to the gap-phrase and have contextual fit to the gap-fill questions.

1.2.2 Question Generation as a Competitive Undergraduate Course Project

Some groups of students in Carnegie Mellon University(USA) created Question and Answer Generation projects for them under graduate Natural Language Processing course. Each group tried with different techniques of Natural Language Processing(NLP) such as Language Modeling, Part of Speech Tagging, Named Entity Recognition, Parsing etc. And they allowed for students to use any programming language and any existing NLP components to complete their projects [1]

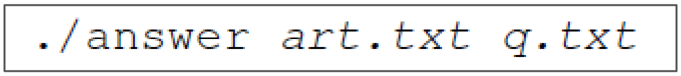
They created command line interface for question generation program. The sample format of command line input is,



```
./ask art.txt N
```

Here, ‘art.txt’ is the file which contains text document, and N is a positive integer to tell how many questions to be generate.

The answer generating program has similar interface, and the sample format of command line inputs is,



```
./answer art.txt q.txt
```

Here, ‘q.txt’ is the list of questions which generated in previous step.

Outcomes

Most of the students developed the systems had the following common characteristics [1]

- Multiple levels of preprocessing of the input articles using existing, freely available NLP tools for various tasks such as sentence boundary detection, tokenization, parsing, named entity recognition, and coreference resolution.
- Use of existing knowledge sources, in particular the WordNet lexical database to replace words with synonyms, antonyms, and other related words.
- Hand-written transformation rules for generating questions from source text. These rules operated on either list of words, dependency parse trees, or phrase-structure parse trees.

1.2.3 Automatic Question Generator in Tamil

One of students group of Indian Institute of Management Ahmedabad implemented an Automatic Question Generator System for Tamil language. They used the Natural Language Generation(NLG) concept of Natural Language Processing(NLP). This automatic Tamil Question

Generator is a form of NLG which produces optimal question set for any given Tamil sentence provided the input is according to the Tamil Grammar [4].

The Architecture of Automatic Question Generator System they developed,

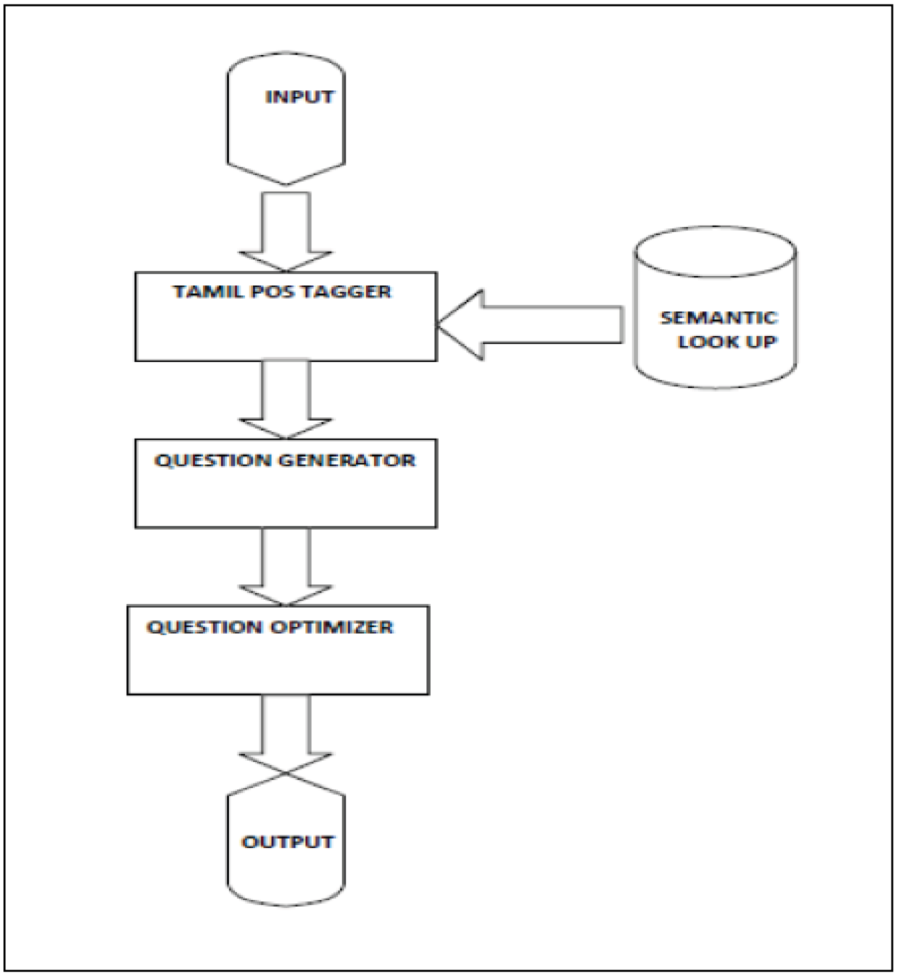


Figure 02

They have used the Part-Of-Speech(POS) Tagger to detect the verb, noun, adjective from the text. There are some taggers available under the POS tagger such as noun marker tag, verb marker tag, time marker tag, noun descriptor tag, verb descriptor tag. They have used these tags to extract the sentences from text and analyze each word of the sentence and identify them as noun or verb or adverb etc., and analyze the verb to identify the tense of the sentence.

1.3 Research Gap and Research Problems

1.3.1 Research Problems

Humans are almost subject to ambiguous, busy, or inconsistent mind in certain situations, question generation problem was raised to be solved and researched by many researchers over the last decade years. Therefore, the potential benefits from an automated **Question and Answer**

Generation (QAG) could assist humans in meeting their useful inquiry needs such as education, knowledge based, daily activities, and much more application(s).

Human process of understanding a problem and deriving solution applying the knowledge learnt, is a process which is hard to automate in the domain of Natural Language Processing. Although **QAG** are designed already, there are limitations in efficiency, accuracy and domain specificity. The problem addressed through this research is the difficulty in simulating the human process of natural language understanding, knowledge applying and retrieval and answer interpretation in a Dynamic machine environment.

QAG is focus to reduce the work load of reading the lecture contents by chapter and automate the manual question preparation.

1.3.2 Research Gap

- **Knowledge retrieval and applying**

For knowledge retrieval component, global and dynamic ontology will be used to process the questions. An efficient way of combining both ontologies for the knowledge applying, will be researched. Multi agent based concept will be applied in order to overcome the overloading problem and retrieve only relevant information, which would be favorable to both memory and efficiency. The logical reasoning agent proposed would consist of algorithms to match both the knowledge base and dynamic ontology which would be useful to generate more accurate questions and would require less time to process and derive the answer.

- **Information Extraction**

All the Information Extraction techniques perform the same tasks but with comparable effectiveness. This effectiveness depends on the algorithms and rules we create. In Information Extraction precision, which shows the number of correctly identified sentences as a proportion of the total number of sentences identified and recall which shows the number of correctly identified sentences as a proportion of the total number of correct sentences available, are the two most frequently used metrics for performance measurement. From the literature review it was found that effectiveness of the Information Extraction should be improved. It can be done by improving precision and recall by creating new Information Extraction techniques while integrating existing ones.

- **Natural language processing component**

According to the existing question generation system, that the current prevailing natural language processing techniques' lack of ability to cater for incomplete unstructured sentences with incorrect grammar and cater for incomplete data. This feature will be further addressed and will try to generate an approach to cater the above needs in a more efficient and a more accurate way.

- **Natural Language Interpretation**

Natural language interpretation is all about representing information to the user in a human readable format. In the proposed system, a linguistic realization component is required to verify English grammar and also to compose the sentence including kinematics terms and units. Currently many algorithms are available for linguistic realization. But they are inefficient and not manageable when building a sentence with English and kinematics terms. Our focus is to develop an algorithm to address the above-mentioned issues and which is efficient and manageable too.

In overall the 'Question Generation System' will be developed exploring more efficient ways of simulating the process of natural language understanding, knowledge applying and retrieval and knowledge interpretation to improve accuracy and efficiency.

The research gaps, which is proposed above will full fill the following.

- Produce appropriate question based on the student current knowledge.
- Produce questions that is semantically correct and the questions' representations is understood by students.
- Produce question that reflect what is the most important for learner to learn.
- Provide facilities for prepare the question paper from chapters.
- Generate Marking scheme according to the question paper.

2. OBJECTIVES

2.1 Main Objectives

Develop a system to school teachers, which will take lecture content as input and, generating suitable questions and appropriate answer scheme for it.

2.2 Specific Objectives

- Categorizing the text book content according to the topic and store in the database.

- Filtering and identifying the content where the questions can be generated to test the knowledge.
- Identifying key phrases and words that should be used to generate appropriate questions.
- Forming proper questions using the identified key phrases from the context.
- Allocate marks for each question.

3. RESEARCH METHODOLOGY

Our system will be going to automate the manual questions preparing process as you see in Figure 1. The lecture contents will be uploaded to system. Initially our system will gather information from lecture content and categorize as chapter and store in ontology driven database. Filtering process will be done from gathered information to identify the contents to form the questions. Complex sentences will be brake into elementary sentences for further process. The system will use elementary sentences to identify the key phrases and words that should be used to generate appropriate questions. The proper questions will be generated from identified key phrases from the context.

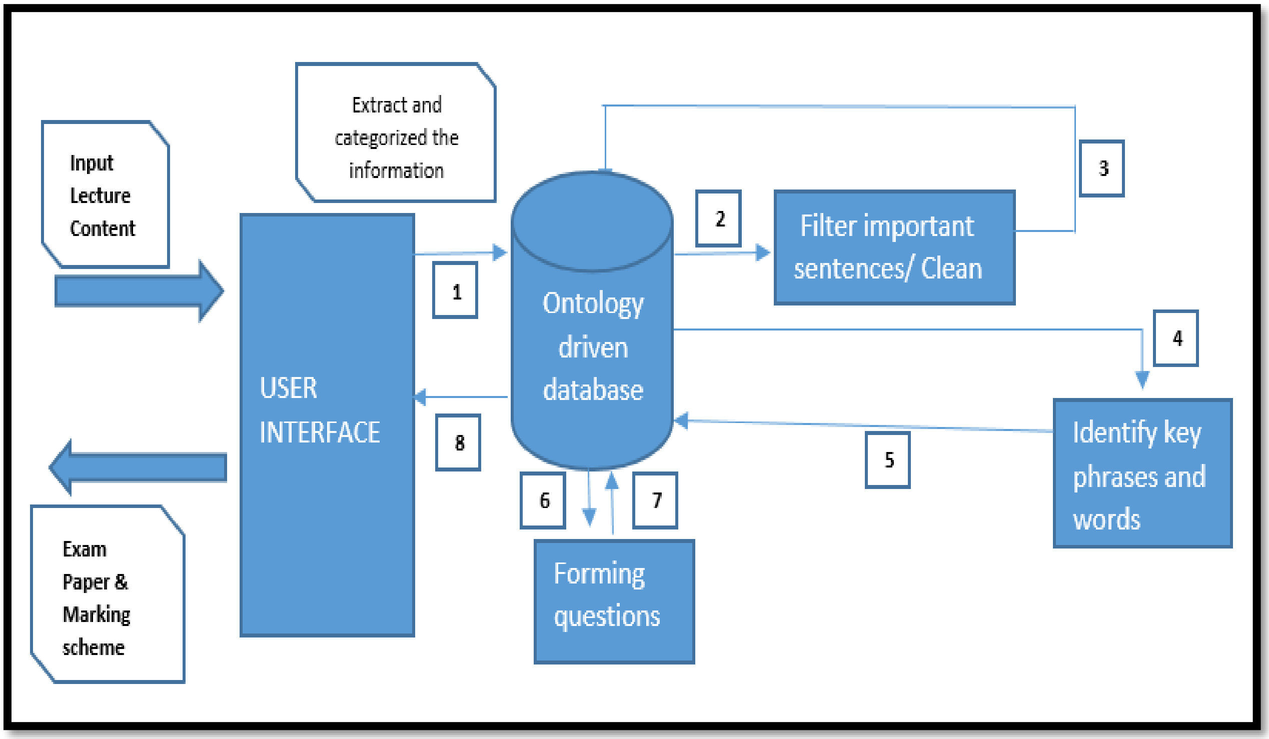

Input Lecture Content
Extract and categorized the information
USER INTERFACE
Ontology driven database
Filter important sentences/ Clean
Identify key phrases and words
Forming questions
Exam Paper & Marking scheme
1
2
3
4
5
6
7
8

Figure 03: Over view architecture

There are four main sub components in our research project.

1. Categorizing the text book content according to the topic/ chapter and storing in the database
2. Filtering and identifying the content where the questions can be generated to test the knowledge.
3. Identifying key phrases and words that should be used to generate appropriate questions.
4. Forming proper questions using the identified key phrases and words from the contexts.

3.1 Categorizing the text book content according to the topic/ chapter and storing in the database

In order to generate the questions by chapter/topic, the lecture contents have to be categorized according to the chapter and its related contents. Since the lecture content more structured, it is hard to store into the normal database to manage and difficult to identify the particular topic areas. In order to make this thing easier we are using the ontology driven database. Mean time it is most suitable to generate questions from the lecture content.

Selecting the Subject (History / Health Science)

The teachers have limitation to choose the subject from our system, because they have to consider only History or Health science. The selected subject soft copy (pdf/word) will be uploaded to the system. This uploaded document will be analyzed by the system. It will identify the contents by chapter and topic.

Identifying the domain of ontology

To identify the suitable domain of the ontology there are two rules we need to follow when we focus on the domain of the Ontology.

- Explicit mention rule - The name of the class itself being available in the text

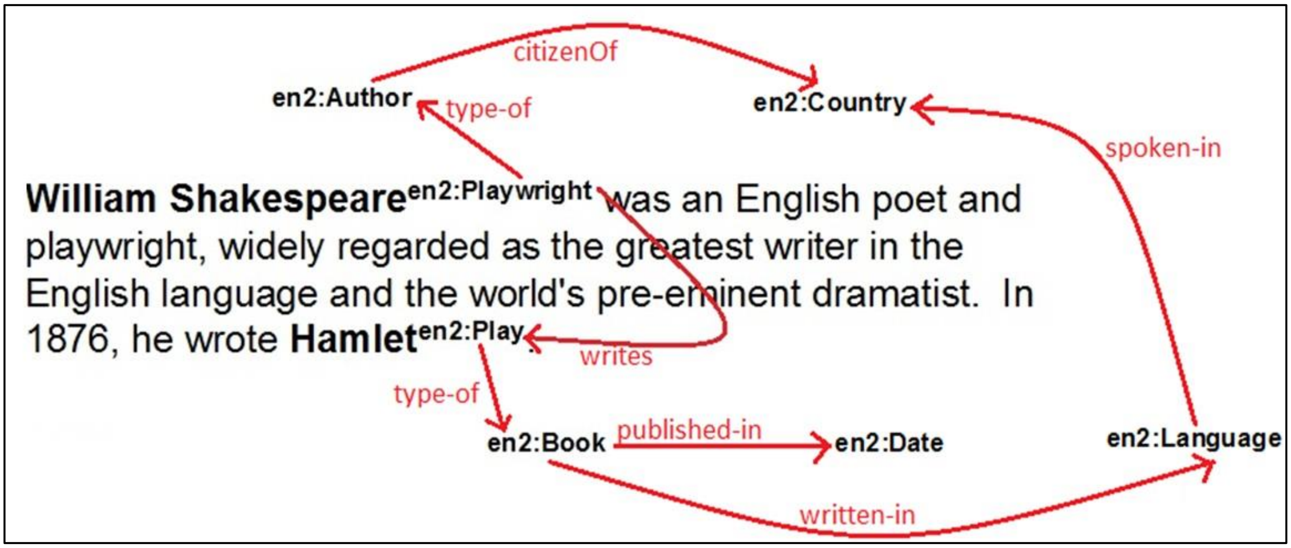
- Implicit lexicon match rule - A domain ontology lexicon which is derived from the ontology class name is used and a string patching is performed. Where we assume that there is a domain ontology existing that matches the derived ontology lexicon. Then we take the number of matches for each domain and we take an intersection of these sets to decide which domain ontology the input belongs to [8].

Storing into the Ontology Database

The analyzed lecture contents will be storing into the ontology database before that will be identify the domain according to the logic. When storing the lecture contents to the ontology we have to write ontologies to get the stored content accordingly. These ontologies can be static because the lecture contents are not going to be changed any more.

How the ontologies work with NLP (Natural language Processing)

Identifying entities in unstructured text is a picture only half complete. Ontology models complete the picture by showing how these entities relate to other entities, whether in the document or in the wider world.



William Shakespeareen2:Playwright was an English poet and playwright, widely regarded as the greatest writer in the English language and the world's pre-eminent dramatist. In 1876, he wrote **Hamlet**en2:Play.

Figure 04

This sentence is really marked up and there’s arrows and red text going all over the place. So, let’s examine this closely. We’ve only recognized (e.g. annotated) two words in this entire sentence: William Shakespeare as a Playwright and Hamlet as a Play. But look at the depth of the understanding that we have. There’s a model depicted on this image, and we want to examine this more carefully. You’ll notice first of all that there is a total of 6 annotations represented on the diagram with arrows flowing between them. These annotations are produced by the NLP parser, and modeled (here’s the key point), they are modeled in the Ontology. It’s in the Ontology that we specify how a Book is related to a Date, or to a Language, and a Language to a Country to an Author, to a work produced by that Author, and so on [7].

3.2 Filtering and identifying the content where the questions can be generated

Ontology Based Information Extraction

An Ontology Based Information Extraction system is a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies. [5] There are several ways that we can achieve the accuracy in Information Extraction. They are syntactic parsing of sentences and pattern matching using rules. When we compare the accuracy of both of these mentioned method, pattern matching gives most accurate results than syntactic parsing. [6] So in this system we will be focusing on pattern matching methodology for Information Extraction.

Sentence Simplification

Text material contents may be in a complex manner. We have to extract the elementary sentences from it. To achieve this, we syntactically parse each complex sentence using The BLLIP parser also known as the Charniak-Johnson parser or Brown Reranking Parser [9] to construct a syntactic tree representation from the bracketed representation of the parsed sentence. We use the depth first algorithm, to read the tree nodes and leaves, which help us construct the elementary sentences, were we maintain if the phrases to be joined are sequentially correct with the respect of the sentence syntactical structure.

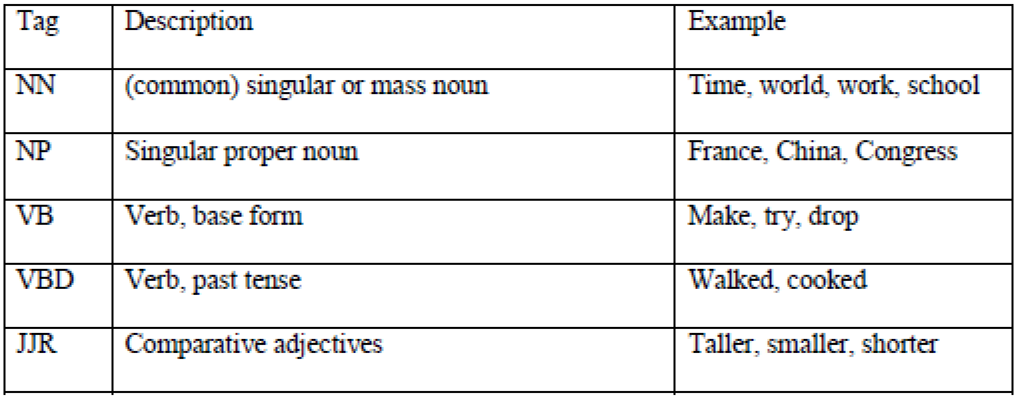
Sentence Classification

Based on the associated POS and NE tagged information; from each elementary sentence, we get the subject, object, preposition and verb which are used to classify the sentences. We use two simple classifiers in this module. Inspired by the idea proposed for question classification in [10], our first classifier classifies the sentence into fine classes (Fine Classifier) and the second classifies the sentences into coarse classes (Coarse Classifier) that are used to identify what type of questions should be generated from one simple sentence.

3.3 Identifying key phrases and words that should be used to generate appropriate questions.

Identify key phrases and their classification

In this module, we analyze each word of the sentence and divide them as noun, proper noun, verb, adjective etc., using Part-of-Speech(POS) concept. The part of speech tagging is used to identify and differentiate each word from other words in the sentence. For that there are different tag set, for example,



Tag	Description	Example
NN	(common) singular or mass noun	Time, world, work, school
NP	Singular proper noun	France, China, Congress
VB	Verb, base form	Make, try, drop
VBD	Verb, past tense	Walked, cooked
JJR	Comparative adjectives	Taller, smaller, shorter

table: 01

First, we identify the noun phrase and verb phrase of the sentence, and identify the noun, verb, adjective etc from the phrases. The work flow of divide the sentence is showing in figure 05

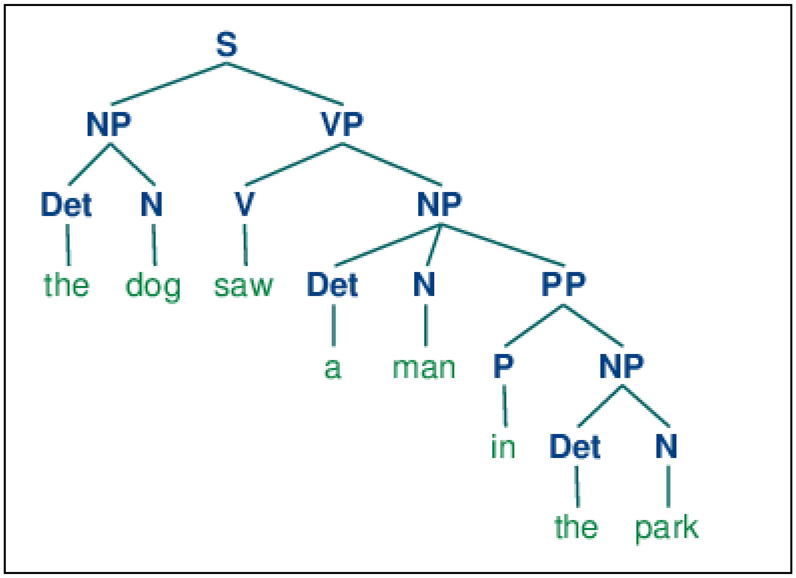


Figure 05

And, Syntactic parser to parse the elementary sentence, and based on the associated POS and Name Entity tagged information, we get from each elementary sentence the subject, object, preposition and verb. This information is used to classify the sentences. This module has two simple classifiers. The first classifies the sentence into fine classes (Fine Classifier) and the second classifies the sentences into coarse classes (Coarse Classifier).

We define the five major coarse classifications as:

1. Human: This will have any subject that is the name of a person.

2. Entity: This includes animals, plant, mountains and any object.

3. Location: This will be the words that represent locations, such as country, city, school, etc.

4. Time: This will be any time, date or period such as year, Monday, 9 am, last week, etc.

5. Count: This class will hold all the counted elements, such as 9 men, 7 workers, measurements like weight and size, etc.

Organizations which include companies, institutes, government, market, etc are all a type of category Entity in our classification. Once the sentence words have been classified to coarse classes, we consider the relationship between the words in the sentence. As an example, if the sentence has the structure “Human Verb Human”, it will be classified as “whom and who” question types. If it is followed by a preposition that represents date, then we add the "When" question type to its classification

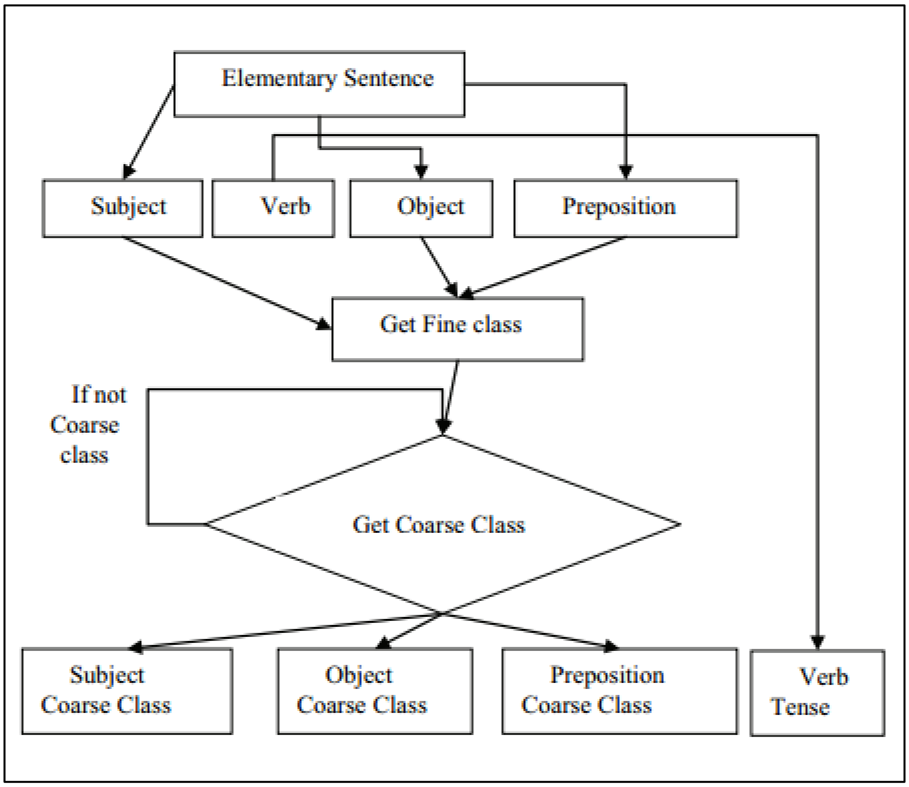


Figure 06: Coarse Classes and Fine Classes classification diagram

3.4 Forming proper questions using the identified key phrases and words from the contexts.

The final steps in syntax-based methods transform declarative sentences into questions by manipulating syntax trees and inserting question words. The system will use predefined interaction rules to define a priori the types of questions that will be generated from **subject-verb-object-preposition** patterns having specific named-entity types in each of those positions [11]. At some point, system needs to use the document-level key phrase identification [12], define separate heuristics for creating questions depending on whether key phrases are contained in subject noun phrases, object noun phrases, appositives, prepositional phrases, or adverbials.

A few examples will demonstrate the behavior and capabilities of syntax-based methods. Consider the sentence below.

Maithripala Sirisena is the president of Sri Lanka.

This sentence does not need any simplification. Its subject noun phrase is extracted and identified as an entity of type person. The question is formed by replacing the subject with the appropriate question word.

Who is the president of Sri Lanka?

Another example, in which sentence simplification would be performed.

Anuradhapura district, the biggest district in Sri Lanka, has many interesting archaeological sites.

This sentence contains an appositive, the biggest district in Sri Lanka. Extracting the appositive and transforming it into a question yields.

Which/Where is the biggest district in Sri Lanka?

A shorter sentence demonstrates the predictable behavior of syntax-based question generation.

Tom ate an orange at 7 pm.

From above sentence following questions can be derived.

Who ate an orange?

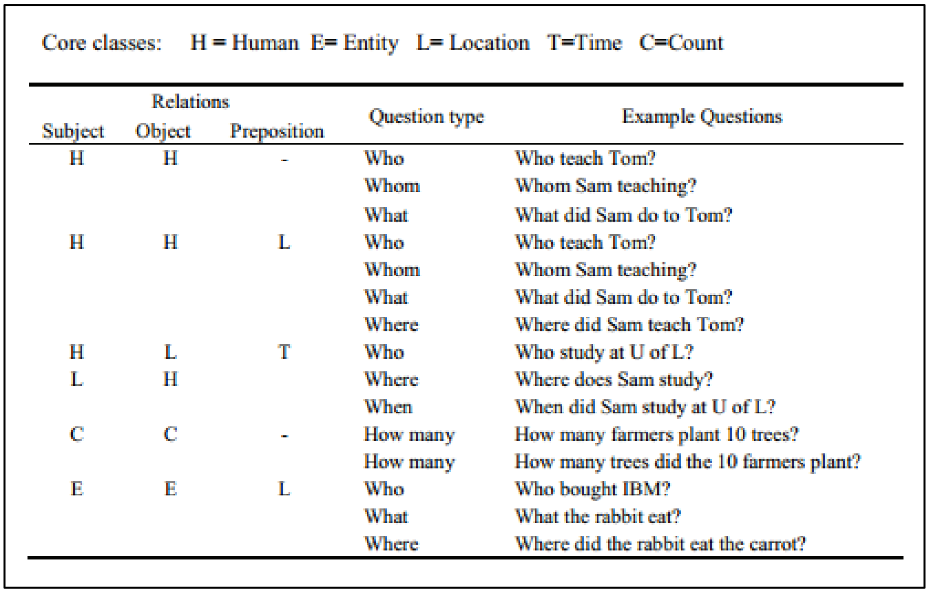
Who ate an orange at 7 pm?

What did Tom eat?

When did Tom eat an orange?

All of these examples serve to illustrate the fundamental behavior of syntax-based methods. QG from text is essentially reduced permuting syntactic elements of a sentence and replacing words or phrases with question words.

This module takes the elements of the sentences with their coarse classes, the verbs (with its stem) and the tense information. Based on a set of predefined interaction rules, we check the coarse classes according to the word to word interaction.



Core classes: H = Human E= Entity L= Location T=Time C=Count

Relations			Question type	Example Questions
Subject	Object	Preposition		
H	H	-	Who	Who teach Tom?
			Whom	Whom Sam teaching?
			What	What did Sam do to Tom?
H	H	L	Who	Who teach Tom?
			Whom	Whom Sam teaching?
			What	What did Sam do to Tom?
			Where	Where did Sam teach Tom?
H	L	T	Who	Who study at U of L?
L	H		Where	Where does Sam study?
			When	When did Sam study at U of L?
C	C	-	How many	How many farmers plant 10 trees?
			How many	How many trees did the 10 farmers plant?
E	E	L	Who	Who bought IBM?
			What	What the rabbit eat?
			Where	Where did the rabbit eat the carrot?

Figure 07: Sample interaction rules

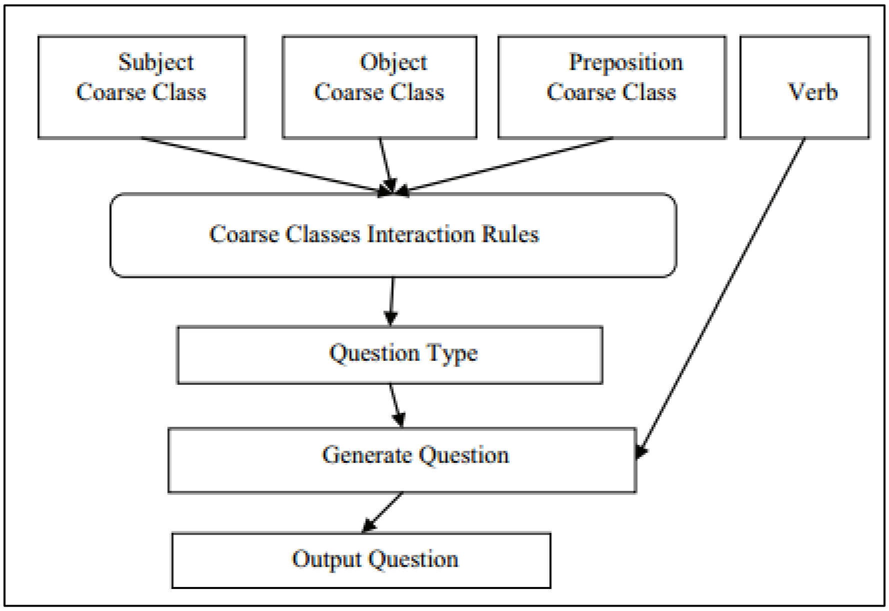


Figure 08: Question Generation process diagram

4. DESCRIPTION OF PERSONAL AND FACILITIES

This section describes the work load assigned to each four members. The research problem divided into four parts with equal proportion, so all members can work with equal effort and focus. Each member’s components related with other members, so everyone should focus on entire project.

Member assigned	Component / Task
IT14058424 - A.S.M Nibras IT14033506 - M.F.F Mohamed IT14121852 - I.S.M Arham IT13001162 - A.M.M Mafaris	Requirement Analysis & feasibility study, Analyse the project requirements and gather details.
IT13001162 - A.M.M Mafaris	Get the subject text materials as word/ pdf, and divide it into chapters and store into database.
IT14058424 - A.S.M Nibras	Get the text materials from database which were stored as chapters, and analyse each paragraph/ sentence and filter the main sentences which can be used to generate questions, then store them into database.
IT14121852 - I.S.M Arham	Get the sentences which were filtered out in previous stage, and identify the noun, verb, adjective, phrases (noun phrase, verb phrase), adverb etc. And store them into database.
IT14033506 - M.F.F Mohamed	Generating questions using phrases, verb, noun, which were identified in the previous stage such as
IT14058424 - A.S.M Nibras IT14033506 - M.F.F Mohamed IT14121852 - I.S.M Arham IT13001162 - A.M.M Mafaris	System Testing

IT14058424 - A.S.M Nibras IT14033506 - M.F.F Mohamed IT14121852 - I.S.M Arham IT13001162 - A.M.M Mafaris	System Deployment
IT14058424 - A.S.M Nibras IT14033506 - M.F.F Mohamed IT14121852 - I.S.M Arham IT13001162 - A.M.M Mafaris	Maintenance
IT14058424 - A.S.M Nibras IT14033506 - M.F.F Mohamed IT14121852 - I.S.M Arham IT13001162 - A.M.M Mafaris	Documentation

5. REFERENCES

[1] Noah A. Smith, Michael Heilman, Rebecca Hwa, "Question Generation as a Competitive Undergraduate Course Project" (2008)

[2] Revup: "Automatically generating questions from educational texts" [Online]

Available : https://www.googlesciencefair.com/projects/en/2015/7151ae4ff6b70198aafc08fbee39127ad0913cd407d98d8b596a85c14ed57ba9

[3] Wikipedia [Online] Available: https://en.wikipedia.org/wiki/Test_(assessment)

[4] N.Vignesh and S.Sowmya, "Automatic Question Generator in Tamil" (2013)

[5] Daya C.Wimalasuriya , "Ontology-Based Information Extraction" (2009) ,

[6] S.R.R Raghu A, "Ontology guided information extraction from unstructured text" (2013)

[7]https://www.ibm.com/developerworks/community/blogs/nlp/entry/ontology_driven_nlp?lang=en [Online]

[8] http://www.antlr2.org/doc/lexer.html [online]

[9] Eugene Charniak, Mark Johnson "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking" (2005)

[10] Xin Li and Dan Roth, "Learning Question Classifiers: The Role of Semantic Information" (2004)

[11]Xuchen Yao, Gosse Bouma, Yi Zhang "Semantics-based Question Generation and Implementation" (2012)

[12]M. Atif Qureshi, Colm O'Riordan, Gabriella Pasi "Exploiting Wikipedia to Identify Domain-Specific Key Terms/Phrases from a Short-Text Collection"