Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 12, December -2016

Automatic Question Generation from Paragraph

Dhaval Swali[1], Jay Palan[2], Ishita Shah[3]

[1,2,3] *Dept. of Computer Engineering, K.J. Somaiya College of Engineering, Vidhyavihar, Mumbai, Maharshtra, India.*

Abstract: *In this paper we have presented an approach to generate questions from a paragraph and the size of the paragraph is defined by its scope. A mix of syntax and semantic based approach to natural language processing is used to generate the questions from the paragraph. Important sentences from the paragraph are selected based upon the certain features and the questions are generated for these selected sentences. Our system implements generation of question from paragraph and also generating simple and complex types of questions. And the research till date works on either implementing question generation from single sentences or implementing generation of simple questions from paragraph or implementing question generation of complex questions from paragraph.*

***Keywords*:** *Discourse Connective, Feature Extraction , Named Entity Recognizer, POS Tagging, Stanford Parser.*

I. INTRODUCTION

Question generation can help a person to generate questions from the given text automatically. It is a process in which given an input text to the system it will create reasonable questions from the input as output. The potential benefits of using automated systems to generate questions helps reduce the dependency on humans to generate questions and other needs associated with systems interacting with natural languages. Question generation can be applied in many fields like intelligent tutoring systems, MCQ generation, FAQ generation etc.

The automatic question generation is an important research area which is potentially useful in intelligent tutoring systems, dialogue systems, educational technologies, instructionalgames etc. Since last few years Automatic QG from sentences and paragraphs has caught the attention of the NLP community through the question generation workshops and the shared task in 2010 (QGSTEC, 2010).

II. RELATED WORK

In earlier work on question generation, Sneider used templates, and Hielman and Smith used general-purpose rules to transform sentences into questions. But, here we present a system which uses the combination of both syntax and semantic based approach. The system will take a paragraph as input and generate important questions from the important sentences extracted from the paragraph. It will generate different wh-type of questions from those selected sentences. The questions will be generated from both simple and complex sentences.

III. PROBLEM DESCRIPTION

The system will take a paragraph as input and generate important questions from the important sentences extracted from the paragraph. It will generate different wh-type of questions from those selected sentences. The questions will be generated from both the complex and the simple sentences. It will consider only those complex sentences which consiststhe following discourse connective: because, and, since, when, as a result, for example, for instance.

IV. PROPOSED METHODOLOGY

Our system is divided into two main modules: *Sentence selection* and *Question generation*. In sentence selection we use the features of the sentences present in the paragraph which will then be selected for sentence selection. Thus only those sentences will be selected which are important in the paragraph on which questions can be generated. Thus ranking of the sentences is done rather than ranking the questions. In the second module of question generation all the sentences selected in the previous module of sentence selection are used for generating questions. Then depending on the sentence type, the framing of the sentence the appropriate questions are generated. The question generation module generates questions from simple sentences as well as complex sentences. Complex sentences are those sentences which contain discourse connective i.e. conjunctions. It will also generate summary type of questions.

The figure below is the overview of our entire system.

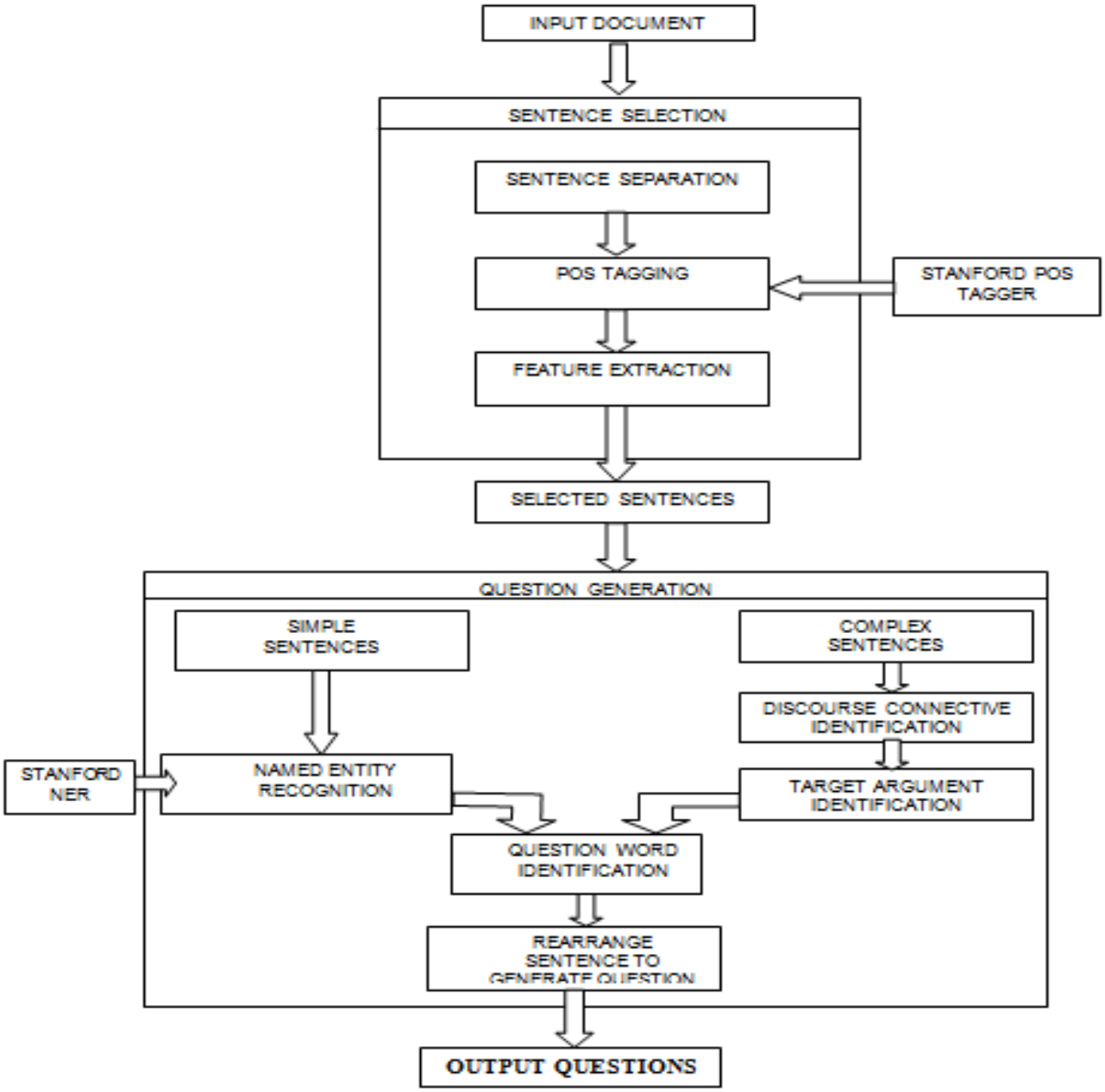

Figure 1.Flowchart of automatic question generation from paragraphs

V. IMPLEMENTATION

A. Sentence Selection

Content selection is crucial for any natural language generation system. In case Of our system it is sentences over which the question has to be asked. The sentence selection module is divided into two subtasks 1) paragraph processing 2) Feature extraction. At the end of this phase we get a set of candidate sentences which are then processed by the next module i.e. Question Generation.

1) Paragraph Processing

In this phase the entire input paragraph is scanned and split into individual sentences. This splitting is done based on full stop. Next each of these individual sentences is processed by a Parts Of Speech Tagger (POS Tagger). To generate the POS tagged sentence we use Stanford POS tagger [citation needed]. The output of this phase is the POS tagged sentences

from which we get information about the parts of speech of individual tokens in the sentences this information is further used by the Feature extraction and Question Generation phases.
E.g.: Input sentence: Robot is a machine

POS Tagged Sentence: Robot/NN is/VBZ a/DT machine/NN
Here NN: noun, VBZ: verb, DT: determiner

2) Feature Extraction

This phase goes through all the individual sentences and extracts a set of features from each of them. Depending on these features it selects all the important sentences on which questions can be generated. Following features are used to select the candidate sentences:
First sentence: This feature checks whether the sentence is the first sentence of the document or not. It is observed that the first sentence in the document usually provides a summary of the document. Hence, this feature has been used to make use of the summarized first sentence of the document.The first sentence of the paragraph is also used to generate the *general scope* question.

Last sentence: This feature checks whether the sentence is the last sentence of the document or not. It is observed that the last sentence in the document usually provides a conclusion of the document. Hence, this feature has been used to make use of the concluding last sentence of the document.

Common tokens: This feature counts the words, only nouns and adjectives that the sentence and the title or the subtitle of the paragraph have in common. A sentence with words from the title in it is important and is a good candidate to ask a question.

Length: This is the number of tokens (words) in the sentence. This feature considers the fact that a very short sentence might not have enough contexts and therefore not a good candidate for generating question. In our case we have considered a sentence with less than 4 tokens as a poor candidate for generating questions and therefore such sentences will not be selected for further processing.

Number of Nouns: This is the count of the number of tokens that are tagged as noun (NN, NNS, NNP, NNPS) by the POS tagger. More number of nouns increases the informational context of the sentence and therefore a sentence with more number of nouns is a good candidate for generating question and is therefore selected for further processing.

Number of Pronouns: This is the count of the number of tokens that are tagged as pronoun (PRP or PRP$) by the POS tagger. More number of pronouns reduces the informational context of the sentence and therefore a sentence with more number of pronouns (in our case greater than 2) is a not good candidate for generating question and is therefore not considered for further processing.

Discourse Connective: This feature examines the presence of discourse connectives in the sentences. Discourse connectives make a vital role in making the text coherent and so wh-type of questions can be easily generated using them. The following table describes the discourse connective and the respective wh- Question Type generated.

Table 1.Question type for various discourse connectives

Discourse Connective	Wh-Question
Since	When
When	When
Because	Why
As a result	Why

Nowwe have a set of features of each individual sentence in the paragraph. Depending upon the combinations of these features a sentence is either selected or rejected for further processing. At the end of the first module (Sentence Selection) we have a list of the selected sentences which will now be processed by the next module i.e. Question Generation.

B. Question Generation

In this module the actual work of generating questions takes place. Here the questions are generated only on the sentences selected in the previous module. It involves two main tasks:

1) Generating questions on simple sentences.
2) Generating questions on complex sentences.

We begin with the dividing the selected sentences into simple and complex sentences each of which are processed separately. The sentences which contain discourse connectives are categorized as complex sentences.

1) Generating questions on simple sentences

Here in this phase we divide the simple sentences into subsections of a English sentence i.e. Subject, Verb, Object.
Then Named Entity Recognizer(NER) is processed over the Subject and Object of the sentence to identity the coarse class classification of it. The NER then specifies the tagged type of the words as Person/human, Location and Organization. The coarse class classification is as follows:
Human: This includes the name of a person.
Entity: This includes animals, plant, mountains and any object.
Time: This will be any time, date or period such as year, Monday, 9 am, last week, etc.
Location: This will be the words that represent locations, such as country, city, school, etc.
Count: This class will hold all the counted elements, such as 9 men, 7 workers, measurements like weight and size, etc.Organization: Organizations which include companies, institutes, government, market, etc.

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. So then based on the sentence structure and the sentences are classified based on the various rules.
Some sample rules are as specified below:

Some of the sample rules for generating questions based on the classes are as follows:

Table 1. Sample rules for generating questions based on the classes

Subject	Object	Preposition	Question type
H	H		who,whom, what?
H	H	L	who, whom, what, where?
L	H		where, when?
C	C		How many?

Here H=human/person, L=Location, O=Organization, C=Count, T=Time, E=Entity.

For example:
Sachin plays cricket at 5 am

Sachin is a subject of coarse class Human
Cricket is an object of type Entity

At 5 am is a preposition of type Time

Sample generated questions based on the rule "Human Entity Time" will be:

Who plays cricket?
Who plays cricket at 5 am?
What does sachin play?
When does sachin play cricket?

2) *Question Generation from Complex Sentences*

The sentences which contain *discourse connectives*i.e conjunctions like because, for example, for instances, since, when etc. are considered as complex sentences. There are 100 distinct types of discourse connectives listed in PDTB manual (PDTB, 2007). The most frequent connectives are and, or, but, when, because, since, also, although, for example, however and as a result. In this paper, we provide analysis for four subordinating conjunctions, *since*, *when*, *because* and *although*, and three adverbials, *for example*, *for instance* and *as a result*. Connectives such as *and*, *or* and *also* showing conjunction relation have not been found to be good candidates for generating wh-type questions.

- Question type identification

Based on the discourse connective the type of the question to be generated is selected. i.e a sentence containing "because" will generate why type of questions. Given below is the list which tells the question type for some discourse connective.

Table 3. Question type for the discourse connective

Discourse Connective	**Q-type**
Because	Why
Since	When, why
When	When
As a result	Why
For example	Give an example where
For instance	Give an instance where

- Argument Identification

Once we have decided the question type we have to then decide which part of the sentence is to be selected for generating question. The discourse connective separates the sentences into two parts ag1 and arg2.
E.g.*[Arg1 Organisms inherit the characteristics of their parents]* ***because*** *[Arg2 the cells of the offspring contain copies of the genes in their parents' cells.]*

So based on the type of discourse connective the argument suitable for question generation is selected. The table below gives the list of discourse connectives and its argument.

Table 4. The target argument to be selected for various discourse connectives

Discourse connective	**Target argument**
Because	Arg1
Since	Arg1
When	Arg1
Although	Arg1
as a result	Arg2
for example	Arg1
for instance	Arg1

- Auxiliary verb Identification:

Here in this part we extract the auxiliary verb present in the sentence which will then be used in framing the sentence. Auxiliary verb is helping verb present in the sentence. It is always present before the verb. It also helps in guessing the tense of the sentence which is very much important for generating questions.

- Question formation:

If the auxiliary is present in the sentence itself then it is moved to the beginning of the sentence; otherwise auxiliary is added at the beginning of the sentence. A question-mark(*?*) is added at the end to complete the question.
E.g.: Competitive badminton is played indoors because shuttlecock flight is affected by wind
Here
*[arg1: Competitive badminton is played indoors]***because***[arg2: shuttlecock flight is affected by wind]*

Arg1 is selected for question generation.

The question type why is selected for because.
The auxiliary is first moved at the start of the sentence to get *is competitive Badminton played indoors*. Then the question type *Why* is added just before the auxiliary *is*, and a question-mark is added at the end to get the final question,
Output: *Why is competitive badminton played indoors?*

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed an approach to automatically generate questions given a paragraph. We have used human effort to evaluate the system. We extract simple and complex sentences from the paragraph and generate question based on subject verb object and prepositions present in the sentence by mapping it to certain predefined rules. Our system does not support anaphora resolution i.e. pronoun resolution. Also our system has a human evaluation so steps can be taken for semantically providing proper sentences. Also many different types of questions like the yes/no question, summary question can be generated.

REFERENCES

[1] 2010 *Question generation shared task and evaluation challenge*, http://questiongeneration.org/QG2010

[2] Manish Agarwal and PrashanthMannem, *Automatic Gap-fill Question Generation from Text Books*, Language Technologies Research Center International Institute of Information Technology Hyderabad, AP, India – 500032

[3] Husam Ali,YlliasChali, and Sadid A. Hassn, *Automation Of Question Generation From Sentences*, In proceedings of third workshop of Question Generation, June 18, 2010.

[4] Manish Agarwal, Rakshit Shah and PrashanthMannem. *Automatic Question Generation using Discourse Cues*. Language Technologies Research Center International Institute of Information Technology Hyderabad, AP, India - 500032

