An Automatic Gap Filling Questions Generation using NLP

Miss.Pranita Pradip Jadhav
M.Tech Student, Computer Department
Dr. Babasaheb Ambedkar Technological University
Lonere-India
pranitaj16@gmail.com

Mrs.Manjushree D. Laddha.
Assistant Professor, Computer Department
Dr. Babasaheb Ambedkar Technological University
Lonere-India
mdladdha@dbatu.ac.in

***Abstract*— An Automatic Blank space-fill Multiple Choices Question Generation method is one of the research fields which is aim to support increasing demand for specialized educational system and active learning. An automatic blank space-fill generation method can proposed to form blank space-fill question (BFQ) with multiple choices (one correct answer and three choices).The crafting of such type of questions is time-consuming for teachers because making the BFQ from the external source material like textbooks and other electronic texts are a very tedious task. It can be generated in three parts, selecting the Descriptive sentence to ask the question, choose blank space of the resulting selected sentence and Search the choices which distract the learner from the correct answer of the question. Natural language Processing (NLP) techniques like Tokenization, Part-of-Speech tagging, Name Entity Recognition are applied on each of these sentences. The advantage of this automatic generation method (AGM) is to provide the services that make it easy for teachers to generate the BFQ and many other competitive exams in which the evaluation can be done through conducting multiple choice questions test (Quiz test).**

Keywords- Automatic Generation method (AGM); Blank fill Questions (BFQ); Named Entity Recognition (NER); Natural Language Processing (NLP); Natural Language Toolkit (NLTK); Part-of-Speech (POS)

I. INTRODUCTION

Assessment is an essential facet of education. However, developing an assessment test is a grueling task. It engrosses teacher's time and efforts which could be spent on teaching performance. There is a necessity for developing educational materials for language learning. An automatic blank space-fill question generator helps to diminish teacher's load and generates questions of consistent caliber which provides an objective assessment. Assessment evaluation plays a deciding role in education and increases its importance in a changing demand teaching environment.

In this paper, we propose the system for the automatic blank space-fill multiple choices questions from the text file and paragraph using Natural Language Toolkit (NLTK) which is a ruling platform for building python programs to work with a human language data processing [1].To initiate the BFQ from the text file, separate the sentences using the symbols like full-stop, exclamatory mark, and question mark.

After that sentence is divided into tokens known as tokenization and then using Part-of-Speech (POS) tag, we acquire the separate word called as token and its type like the word is noun, pronoun, verb, adjective etc. Selection of descriptive sentence for BFQ is based on the number of noun, pronoun and superlative degree present in the sentence.

For blank space selection, put priority to the noun, pronoun and superlative degree present in the sentence and removes the appropriate word from the sentence and makes the blank space. Formerly the question is generated; the annoying task is to find choices for the blank which to be selected. For this cause, used Named entity Recognition [2] and used Wikipedia is a help to find the applicable choices for blank.

II. RELATED WORK

Generating automatic BFQ is relatively new and very emerging research topic which is useful in education technology. Here we first discuss the few models or systems for automatic blank space-fill question generation.

Sheetal Rakangor and Dr. Yogesh Ghodasara (2013) can proposed the system finds fill in the blanks, blanking key generates from the selected statement. Syntactic and lexical features are used in this process. NLP parser is used. POS taggers are applied on each of these sentences to encode necessary information [3].

Sheetal Rakangor and Dr. Yogesh Ghodasara (2014) both are proposed the distractor selection on the basis blank space selected name, organization and place using NER [4].

Manish Agarwal and Prashanth Mannem (2011) they used to selects most informative sentences of the chapters and generates blank space-fill questions on them using syntactic features like height of tree i.e Heuristic function [5].

Brown et al (2005) have conducted the task of Automatic Question Generation with a linguistic motivation. A multiple-choice cloze question is generated in the way that the correct answer of the question is the target word. They have restricted the distractor selection from their targeted set of words. He used WordNet for finding definition, synonym, antonym, hyponym and hyponym in order to generate the questions as well as the distractors [6].

Mitkov et al. (2006) used NLP techniques like shallow parsing, term extraction, sentence transformation and computation of semantic distance in their works for generating MCQ semi-automatically from an electronic text. They did term extraction from the text using frequency count, generated stems using a set of linguistic rules, and selected distractors by finding semantically close concepts using WordNet [7].

Aldabe and Maritxalar (2010) developed systems to generate MCQ in the Basque language. They have divided the task into six phases: selection of text (based on learners and length of texts), marking blanks (manually), generation of distractors, selection of distractors, evaluation with learners and item analysis [8].

Lee Becker, Sumit Basu and Lucy Vanderwende (2012) they presented to generate blank space-fill questions using the Wikipedia i.e Electronic Text. They proposed that the sentences of question can be divided by using NLP heuristic and then fills the blank space [9].

III. TECHNIQUE USED FOR QUESTION GENERATION

Blank space-fill multiple choice question generators can be creates BFQ in three distinct levels to lighten the load of teachers for creating quiz test paper.

The process of generating and analyzing the BFQ with multiple choices consists of the following steps:

In this Automatic generation method,

(1) **Descriptive sentence selection:** Pick out analytical and meaningful sentences from the text file to inquire the question.

(2) **Blank space selection:** From sentence determine the blank space.

(3) **Alternative choices selection:** Draft three choices which have the same context of the blank space and will trouble the learner to select the precise answer.

A. Descriptive Sentence Selection

Allow the text file as an input for selecting a consistent and logical sentence from the input text to form the question. Blank fill question can be asked on selected descriptive sentences and is done by using NLTK. Divide each and every sentence using a full stop(.), Question mark(?) and Explanatory mark(!).

Get the separate sentences. Apply the POS tagging and get the words with its type. In the case of the word are noun, pronoun, adverb, adjective, determiner, superlative degree present in the sentence. Perform the pattern matching for getting the noun, pronoun and superlative degree. If there is no noun, pronoun and superlative degree present in the sentence then discard the sentence.

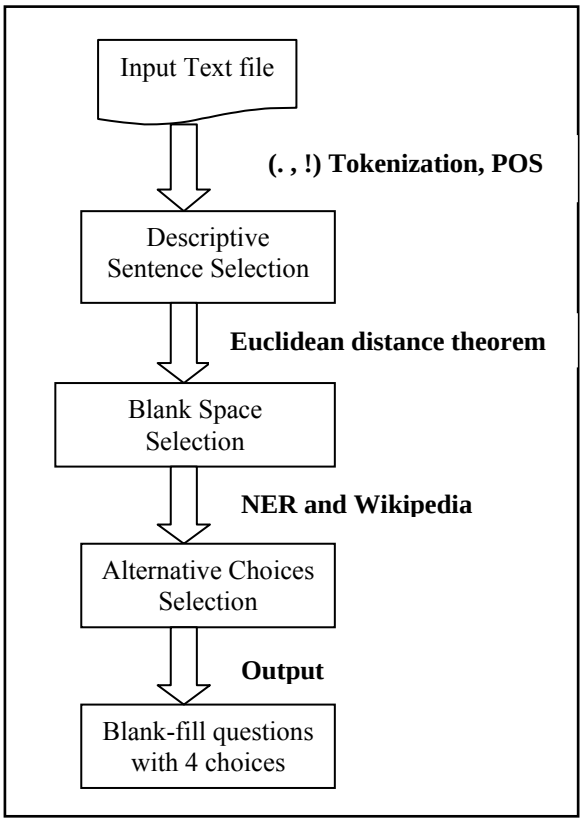

Fig 1: Technique for the blank space-fill question generation method

1) Extracting features by POS taggers are:

(i) Sentences calculated in the text file.

(ii) Shows the nouns (NN), pronoun (NNP), adverbs (RB), adjectives (JJ), determiner (DT) etc. present in the sentence.

(iii) Shows the adjective superlative (JJS) degree of sentence.

-Superlatives are suffix with –est. like biggest.

2) Algorithm for Descriptive Sentence Selection

i. Extract the text file.
ii. Read the sentences from the text file.
iii. Calculate the number of sentences.
iv. For all sentences, calculate the nouns, pronouns, adjectives, adverbs etc.
v. Find the named entity in each statement (Name, Location, city, country etc) Store the named entity into the database where all the previous name entities are stored.
vi. If the sentence which contains a noun, pronoun and superlative degree then the sentence is selected.
vii. Else if the sentence contains max [noun] then it is selected.
viii. If superlative degree is not found then catch the sentence which having a number of noun and pronoun present.
ix. Else if sentence will be disposed, if no noun, pronoun and superlative degree present.
x. End if.
xi. End for.
xii. Display the selected sentence.

B. Blank Space Selection

Once the sentence is selected for the blank space-fill question, there will be the task to select a blank which is very important level. POS tagger can administer a linguistic image between the words in the sentence.

The task of descriptive sentence selection is done from the text file and then pushes all the nouns, pronouns, superlatives present in sentences into the potential key list. From this key list, the generator will select one best key as a blank. If any noun is restated in the list it will be detached and formulate a blank space.

To find the occurrences of the key in the text file is calculated by the Euclidean Distance theorem [10].Applying theorem and assign the unique id to the name entities present in the key list and finds their occurrences in the text file and calculated by using the equation (1),

$$d^2_{xy} = (x_1 - y_1)^2 + (x_2 - y_2)^2 \quad \ldots\ldots\ldots \quad (1)$$

Where, *x* denotes the unique id of name entities present in potential keys list and the *y* denotes their occurrences in the paragraph and whole text file.

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad \ldots\ldots\ldots\ldots (2)$$

i.e. the distance itself is the square root value. Select the key which has minimum value is found.

1) Algorithm for blank space selection

i. For each sentence, extracts the words and its types (noun and pronoun) from the sentence and pushed it into the potential keys list of the sentence.

ii. From this list, selection of a blank space on the basis of their occurrences in selected sentence and text file.

iii. End for.

iv. Remove the best key from the sentence and generate blank-fill question.

C. Alternative Choice Selection

Once the best key is selected from the pool of potential keys and making the blank in selected sentence, we get one correct choice which is our answer to that question. To find out the choices for BFQ and AGM invoke the Named Entity Recognition (NER).The common NER task is mapping named entities to concepts in vocabulary and dataset.

Check whether the selected key is name, organization, number, time, place, country, disease etc and then retrieves the appropriate choices from the dataset. In which any text file is given as an input to the AGM, it will find the name entities present in each sentence (name, place, time organization, city, country etc.) and automatically stored into the dataset and dataset is dynamically updated every time. Second, fetch the alternative choices from the Wikipedia by giving the best key as an input.

To choose the choices from the dataset, used randomized theorem and gets the different choices in fraction of second like 75 entities are present in the dataset which are relevant to the answer then it will divide the 75 entities into three parts and then fetch the one entity from each section as choices. By doing this, it will not select the same choice frequently. It will always give the variant choices.

1) Select alternative choices on the basis of following properties:

i. **Semantic Checking**: The choices that are selected for the questions that should be in same meaning or context of the blank space.

ii. **Syntactic Checking**: Choices that is complementary and identical to the blank space of the sentence. For ex. T-phase probably as good distractor for G-phase.

iii. **Contextual Meaning**: Choices need to competent to the question.

2) Algorithm for alternative choice selection

i. For each sentences, diff (alternative choices, key) with comparable importance to the key.

ii. For each dataset, retrieve the arbitrary three name entities with equal importance perhaps close in their semantic meanings.

iii. Interchange their positions.

iv. Display the blank-fill question into text file with the four options.

v. End for.

Once all the three levels are achieved by an AGM. It will deliver the sorted blank-fill questions with suitable choices.

IV. PROCESS FLOW MODEL

Extract the text files and used as an input for AGM. It will separate all the sentences present in the file and calculate the number of sentences.

In Example, There are two sentences extracted from paragraph which are C++ is Object Oriented Programming Language and OOP follows Bottom-up Approach. .Selection of descriptive sentence can be done by applying the POS tagging, get number of noun, pronoun and adjective etc. present in sentence From the selected sentence.

Key list is generated in which all the nouns, pronouns, superlatives present in the sentence are pushed. Select one key as a blank space on the basis of their occurrences in paragraph and sentence and generate:

(1) ____ is Object Oriented Programming Language.

(2) OOP follows _____ approach.

Next level is to find the distractors for the blank space. For blank space selection put priority to noun, pronoun and superlative degree. It encounters in sentence then make a blank space. For C++ and bottom-up select the appropriate choices from the pool of name-entities dataset and Wikipedia i.e. for C++ and Bottom-up you get options like visual basic, python, C and top-down, parallel and serial respectively.

Interchange the position of choices and display the blank space-fill question with choices.

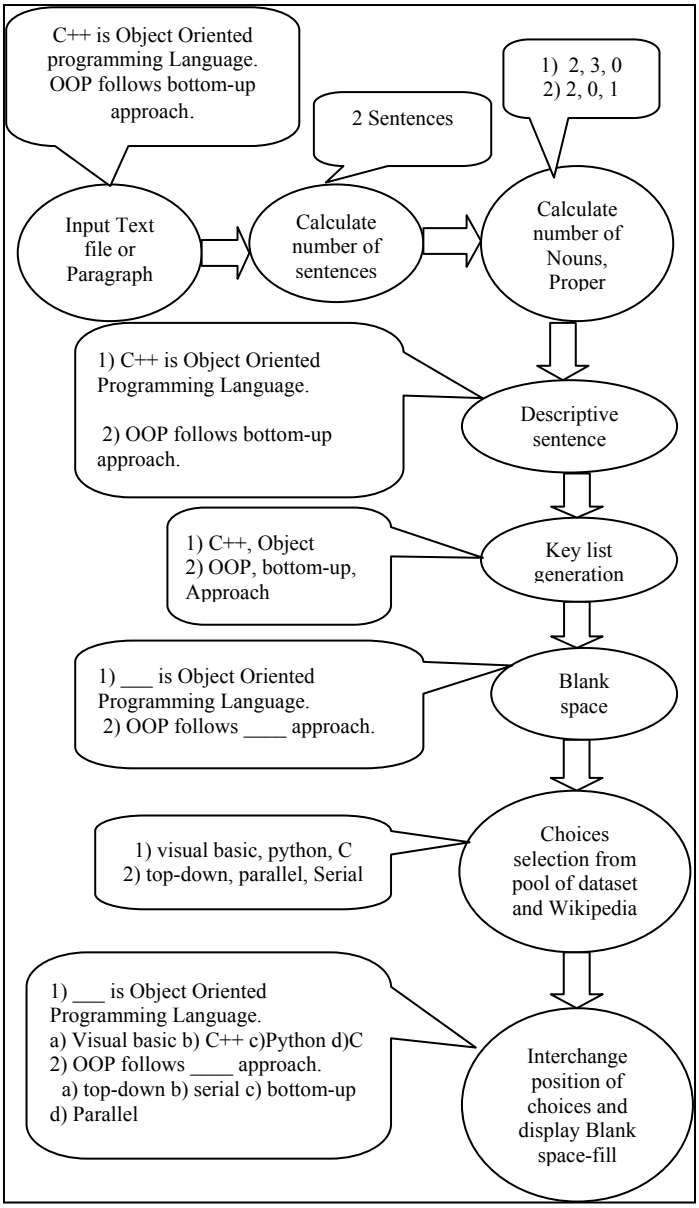

Fig 2: Process Flow of an automatic blank-fill multiple choice questions

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have shown our initial exploring experiments towards creating an automatic question. System will select the descriptive sentence from the paragraph and generate fill in the blanks with distractor from the paragraph and with the help of Wikipedia. For that used Natural Language Toolkit for selection of descriptive sentence and Selection of blank space from the paragraph.

To obtain the distractors we look for the synonyms, antonyms, and similar words for the distractors that are find from dataset of name entities, Wikipedia or the given paragraph.

It is very difficult for questions that are automatically generated to be as good as questions generated by human experts. Currently, our methodology focuses on improving the correctness of the answer. From paragraph, get number of sentences with blank space and choose coherent sentence and best blank space is first challenge.

Second is for quality improvement of distractors that fits in the sentence is contextually and semantically same. To obtain a better performance, we intend to develop an AGM to get the pattern based sentences and make no restriction on election of noun, pronoun and superlatives.

VI. REFERENCES

[1] Natural Language Processing with Python by Steven Bird, Ewan klein and Edward Loper, by O'Reilly Publication.

[2] Jia-Li You, Yi-Ning Chen(2008) Identifying Language Origin of Named Entity With Multiple Information Sources, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16, No. 6, August 2008.

[3] Sheetal Rakangor and Dr. Yogesh Ghodasara (2013) Computer aided environment for drawing (to set) fill in the blank from given paragraph. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 6 (Nov. - Dec. 2013), PP 54-58

[4] Sheetal Rakangor and Dr. Yogesh Ghodasara (2014)
Automatic Fill in the blanks with Distractor Generation from given Corpus, International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 9, November 2014.

[5] Manish Agarwal and Prashanth Mannem(2011) Automatic Blank space-fill Question generation from text books.

[6] Brown, J. C., Frishko_, G. A., and Eskenazi, M. (2005) Automatic question generation for vocabulary assessment. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 819-826.

[7] Ruslan Mitkov, Le An Ha and Nikiforos Karamanis (2006)
A computer-aided environment for generating multiple-choice test items, Natural Language Engineering 12(2): 177-194.

[8] Aldabe, I., Maritxalar, M., (2010). Automatic Distractor Generation for Domain Specific Texts. Proceedings of IceTAL, LNAI 6233. pp. 27-38.

[9] Lee Becker, Sumit Basu and Lucy Vanderwende (2012)Mind the Blank space: Learning to Choose Blank spaces for Question Generation. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 742–751, Montreal, Canada, June 3-8, 2012.

[10] Computational Optimization and Applications 12, 13–30 (1999) Kluwer Academic Publishers. Manufactured in The Netherlands. Solving Euclidean Distance Matrix Completion Problems.