Automatic Fill in the blanks with Distractor Generation from given Corpus

Sheetal Rakangor
Research Scholar
R.K. University
Rajkot, India

Y. R. Ghodasara, Ph.D.
Associate Professor
College of Agricultural Informa & Technology,
Anand, India

ABSTRACT

In this paper, Researcher present an automatic fill in the blanks with distractors generation from the given corpus. Distractor means multiple choices provided, i.e. four options are provided out of that three are the distractor and one is correct answer. System is developed in java using JDBC and mysql for storing the question, both are open source.

Standford NLP parser is used for parsing the sentences and generated informative questions. POS tagger and NER functionality of parser used to encode the sentence. NER functionality is also used to identify whether key selected is Name, Place or Organization.

General Terms

MCQ Question, Standford Parser

Keyword

Distractor, NER Feature, POS tagger, Name Place, Organization

1. INTRODUCTION

Now a days, in any examination system Fill in the blanks with distractor and Multiple choice question is widely used to judge the student knowledge and Evaluation of this type of question is computerized, but construction of such question is still manual process, which is very time consuming and labour task. Researcher had studied and constructed system through which automatic fill in the blanks question will generate from given corpus and distractors (wrong answers) are generated.

In this paper research has concentrated on fill in the blanks with distractor, where in MCQ WH styles question generates, WH style questions contain sentences with blanks form a question. Researcher has first concentrated on Fill in the blanks with distractor where sentence with blanks is generated.

The shape of the earth is ____________and not a perfect sphere ; it is flattened at the poles

a). ball

b). 3-D

*c). **spherical***

d). square

In above example fill in the blank is generated from paragraph and four alternatives are also generated through system out of that three options are distractor and one is right answer. option c) is correct answer spherical for above blank generated.

The aim is to go through the paragraph's and extract the informative sentence from the given paragraph and generate fill in the blanks question. System takes paragraph's as input and produce list of fill in the blanks with distractor questions as output.

In this model, fill in the blanks with distractors system first find maximum number of noun or superlative degree available in the sentence and informative sentence is selected . And on the basis of Informative sentence, key is selected and blank part is generated by extracting key from the selected sentence. Distractors are generated automatically, first it will check whether key selected is Name, Place or Organization for extracting distractors from relevant database (NER feature of standford parser is used). if other than Name, Place or Organization distractors will be generated from the paragraph's entered by user, as well as from database.

2. DATA USED

Different paragraphs downloaded from the internet and textbook paragraphs are also used to generate fill in the blanks with distractors.

3. APPROACH

For generating Fill in the blanks with distractors four stages are used: *Data Processing, sentence selection, key selection and distractor selection*. Sentence selection involves identifying important sentences in the paragraph which can be used to generate a fill in the blanks with distractors question. These sentences are then processed in the key selection stage to identify the key on which to ask the question. In the final stage, the distractors for the selected key are identified from the given paragraph, and outside database are used like thesaurus, homonyms, Organization, Person name, City, State and Country databases are used for extracting distractors.

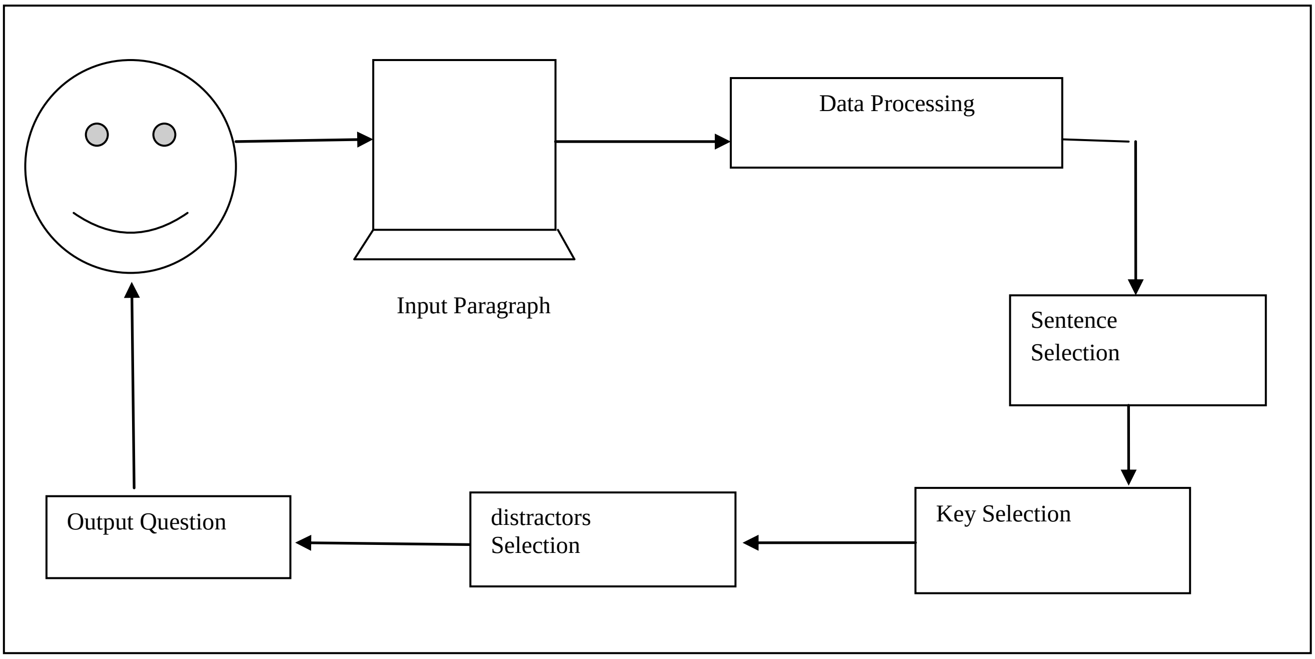


Figure 1 Architecture of Fill in the blanks with distractors

3.1 Data Processing

In Data Processing module goes through all sentences from given paragraph and use NLP Stanford parser which parser the sentences and divided into small fragments called *token*. And from that token POS tagger is used, which provides a representation of grammatical relations between words in a sentence.

3.2 Sentence Selection

In Sentence Selection module, (Agarwal and Mannem) [4] uses many features for sentence selection some of the features are used here to generate informative sentence from the corpus. For Informative sentence extractions a set of features use are,

Count number of Sentences: Paragraph's entered by user, count number of sentences from that paragraph entered.

Count number of words: Count number of words in the sentence. Short sentence generate unanswerable question because short content and very long content might have enough content to make the question generated.

Count number of nouns: Noun gives an idea about the sentences, if maximum number of noun in sentence means potential key can be generated from that sentence and that sentence having good content which can generate the key for fill in the blanks.

Superlative: Superlative degree defines exaggerated mode of expression or height of quality. Superlative are typically formed with suffix –est (healthiest) or the word most, good, best are used. Sentence which contains superlative degree can generate good fill in the blanks.

On the basis of these features important sentence will be extracted from the given paragraph's.

Algorithm for sentence selection

```
1.  Enter the paragraph P
2.  Read the statements from the paragraph S.
3.  Calculate number of sentences CtS.
4.  Calculate number of words from each sentence CtW
5.  For each CtNoun and CtSuper from S do
      Select the sentence which contains superlative degree and then calculate maximum number of nouns which contain
      superlative degree.
6.  IF CtSuper and CtNoun from S then
              SetenceSelected SS
      ElseIF Max (CtNoun) from S then
              SetenceSelected SS
      If there is no superlative degree then select that sentence which having maximum number of noun
      Else
              Without Noun and Superlative degree, blanks will not generated
      EndIF
      EndFor
7.  Display SentenceSelected SS
```

3.3 Key Selection

Key selection is most important stage, to identify the key from important sentence to ask the question on. Previous work in this area, [5] takes key as input and [6] select key on basis of regular expression on noun. Or first search the key and then basic sentence is selected. key selection approach is divided into two stages. Generate *Potential keys* from the statement and select *Best Key* form that key list [4]·

The Solar System consists of the Sun Moon and Planets.

The/DT Solar/NNP System/NNP consists/VBZ of/IN the/DT Sun/NNP Moon/NNP and/CC Planets/NNPS

The Solar System consists of the Sun Moon and Planets.

First Stage: Potential keys are generated from selected sentences, POS tagger is used to identify the words and there type. Suppose, need to generate key for noun then select total numbers of noun available in sentence and list them, and if any noun word is repeated in that list, it would be removed from list [4]. As shown above example the five nouns extract *Solar, System, Sun, Moon, Planets*. are pushed into the keylist.

Second Stage: Best keys are generated from the key-list, select the word from key-list and search that word in paragraph. Count how many times that word has been used in paragraph. Select best key which has found in selected sentence and noun repeated in paragraph maximum number of times.

As shown below the algorithm of key selection and Table 1: Describes how paragraph's entered and through system Key List is generated and from Keylist, BestKey is selected which is in red color.

Algorithm for key selection

```
1.  For Each Word from SS do
        From selected sentence extract the potential key now suppose need to generate key of noun then extract the nouns from the sentence. And add them into the keylist.
        KeyList =Select Noun from SS
        BestKey = No of Occurrence of that key in SS and Height of that key in the syntactic tree Structure.
    End For
2.  Remove that BestKey from sentence SS and generate Fill in the blanks.
```

Table 1 shows selected keys in red colored for sample and keylist which is generated from entered paragraph.

No.	Paragraph	Keylist
1	Khushbu is silent student in the class. **John** is the tallest in the class.	Khushbu, student, class, John
2	India is an agricultural country. Most of the people live in villages and are farmers. **farmers** grow cereals pulses vegetables and fruits.	India, country, villages, farmers, cereals, vegetables, fruits
3	**Delhi** is the capital of India. It is situated on the banks of the river Yamuna. It is a beautiful city. But it is becoming very crowded and polluted.	Delhi, India, banks, river, Yamuna, city

3.4 Distractor Selection

For Distractor selection, Named entity recognition feature of Stanford parser is used. first it identify the key, system will extract noun key so it can be i.e. Name of person, Organization name or Location of the world, or other than this.

3.4.1 Key is Name of the Person

If Key is Name of the person i.e. Sachin *is the good batman in India Cricket Team.* is the sentence system will generate "Sachin" as the key. so NER feature will identify sachin as PERSON. so for distractors, name start with "S" in the person_name database will fetch and set two distractors from database.

One distractor will extract from the paragraph. For fetching distractor from paragraph distractor key list is generated. distractors key list will contain *noun and person name* from the paragraph, distractor key list will not select key from selected sentence.

so if key is person name then two Distractors are generated from person_name database and one Distractor are generated from paragraph. Total three distractors are generated and one correct answer of the fill in the blank question.

As describes in Table 2: if Key is Name of the person then, Distractor key list will generate from paragraph as well as from database

Table 2 of Distractor list where key is name

No.	Paragraph	DistractorList from paragraph	Distractors from person_name database	Final Distractors
1	Sheetal is silent student in the class. **John** is the tallest in the class.	Sheetal, student, class	Joan, Jenny, Jenifer, Jenni	a) John b) Sheetal c) Joan d) Jenni

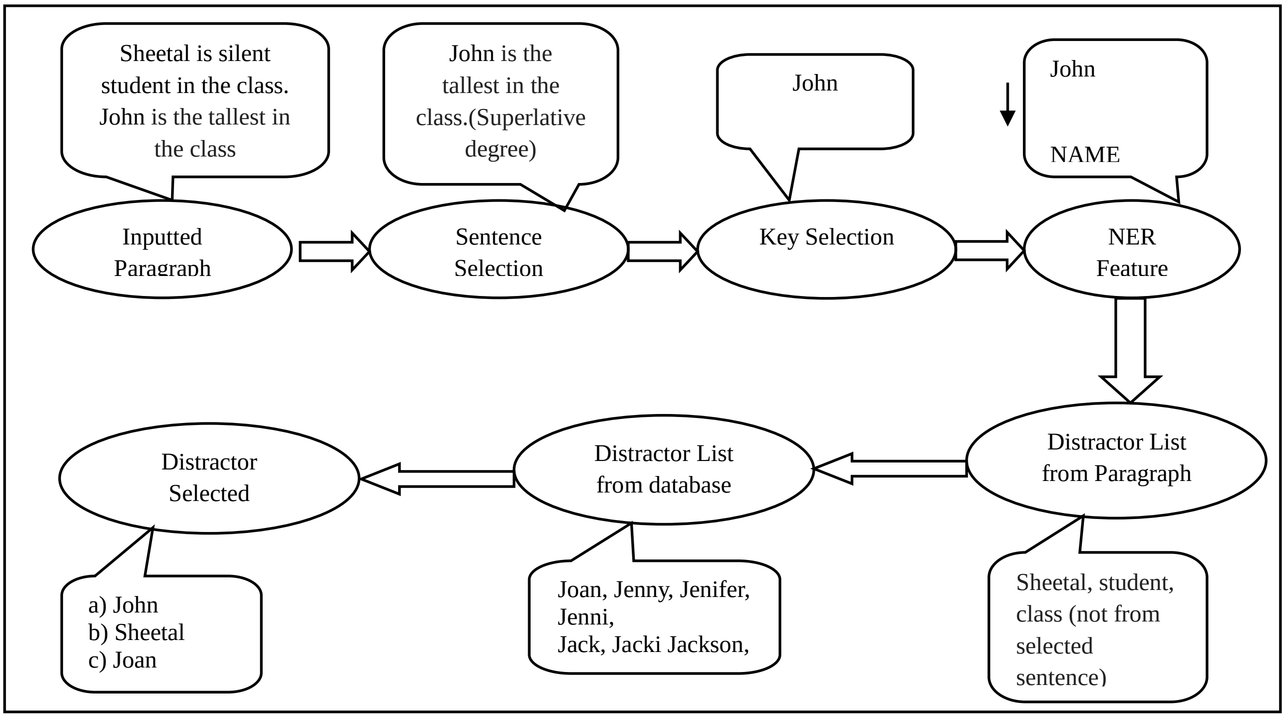

Sheetal is silent student in the class. John is the tallest in the class
John is the tallest in the class.(Superlative degree)
John
John
NAME
Inputted Paragraph
Sentence Selection
Key Selection
NER Feature
Distractor Selected
Distractor List from database
Distractor List from Paragraph
a) John
b) Sheetal
c) Joan
Joan, Jenny, Jenifer, Jenni,
Jack, Jacki Jackson,
Sheetal, student, class (not from selected sentence)

Figure 2 Process Flow of Fill in the blanks with distractors when keys is Name of person

3.4.2 Key is Name of the Organization

If Key is Organization i.e. *University of California is located in California.* is the sentence system will generate " *University of California* " as the key. so NER feature will identify *University of California* as ORGANIZATION so for distractors name start with "U" in the organisation_name database will fetch and set two distractors. one distractor will extract from the paragraph. For fetching distractor from paragraph distractor key list is generated. distractors key list will contain *noun that should be name of the organization* from the paragraph other than key is fetched from paragraph. so if key is organization name then two Distractors are generated from organisation_name and one Distractor are generated from paragraph. total three distractors are generated and one correct answer of the fill in the blank question.

As describes in Table 3: if Key is Organization then, Distractor key list will generate from paragraph as well as from database of organization.

Table 3 Distractor list where key is Organization

No.	Paragraph	DistractorList from paragraph	Distractors from organization_name database	Final Distractors
1	**Bharatiya Kisan Sangh** organization is not good as compare to Centre for Ecology & Rural Development organization. Centre for Ecology & Rural Development is also nice organization having good name in the market.	Centre for Ecology & Rural Development	Bharatiya Gorkha Parisangh, Bharatiya Vichara Kendra	a).Bharatiya Gorkha Parisangh, b). Bharatiya Kisan Sangh c). Centre for Ecology & Rural Development d). Bharatiya Vichara Kendra

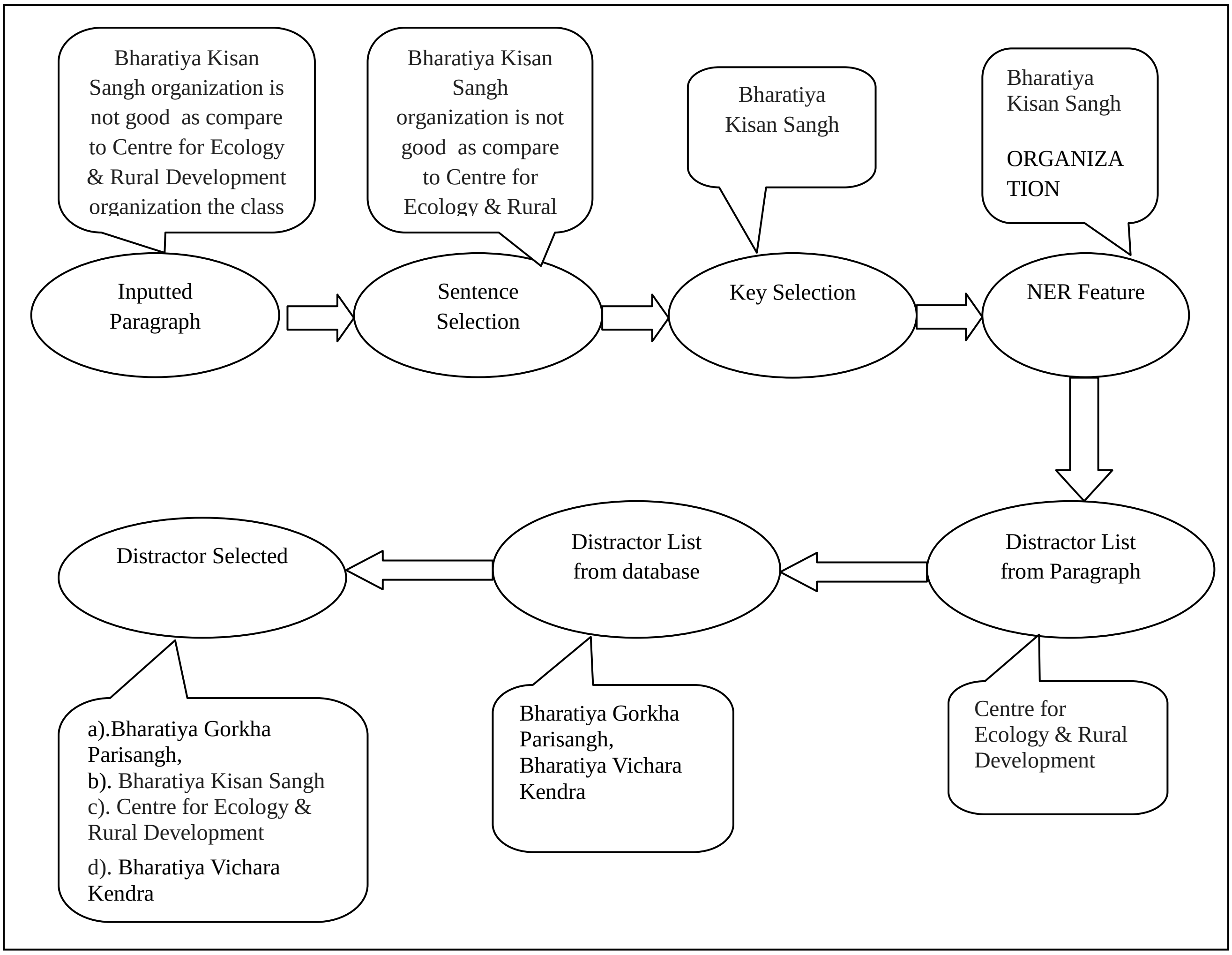

Bharatiya Kisan Sangh organization is not good as compare to Centre for Ecology & Rural Development organization the class
Bharatiya Kisan Sangh organization is not good as compare to Centre for Ecology & Rural
Bharatiya Kisan Sangh
Bharatiya Kisan Sangh
ORGANIZATION
Inputted Paragraph
Sentence Selection
Key Selection
NER Feature
Distractor Selected
Distractor List from database
Distractor List from Paragraph
a).Bharatiya Gorkha Parisangh,
b). Bharatiya Kisan Sangh
c). Centre for Ecology & Rural Development
d). Bharatiya Vichara Kendra
Bharatiya Gorkha Parisangh, Bharatiya Vichara Kendra
Centre for Ecology & Rural Development

Figure 3 Process Flow of Fill in the blanks with distractors when keys is Organization

3.4.3 Key is Location (City State or Country)

If Key is Location i.e. "*Gujarat is part of India* is the sentence, system will generate " *Gujarat* " as the key. so NER feature will identify *Gujarat* as LOCATION so for distractors name start with "G" in the state_name database will fetch and set two distractors. one distractor will extract from the paragraph. For fetching distractor from paragraph distractor key list is generated. distractors key list will contain *noun and location* from the paragraph other than key is fetched from paragraph. so if key is location name then 2 Distractors are generated from state_name and 1 Distractor are generated from paragraph. total 3 distractors are generated and 1 correct answer of the fill in the blank question

As describes in Table 4: if Key is City, State or country then, Distractor key list will generate from paragraph as well as from database of City, state and country.

Table 4. Distractor list where key is City, State or Country.

No.	Paragraph	DistractorList from paragraph	Distractors from location_name database	Final Distractors
1	**India** is populated Country. Pakistan and Nepal are neighbors of India.	Pakistan Nepal	Iceland Indonesia Iran	a). Nepal b). India c). Iceland d). Iran

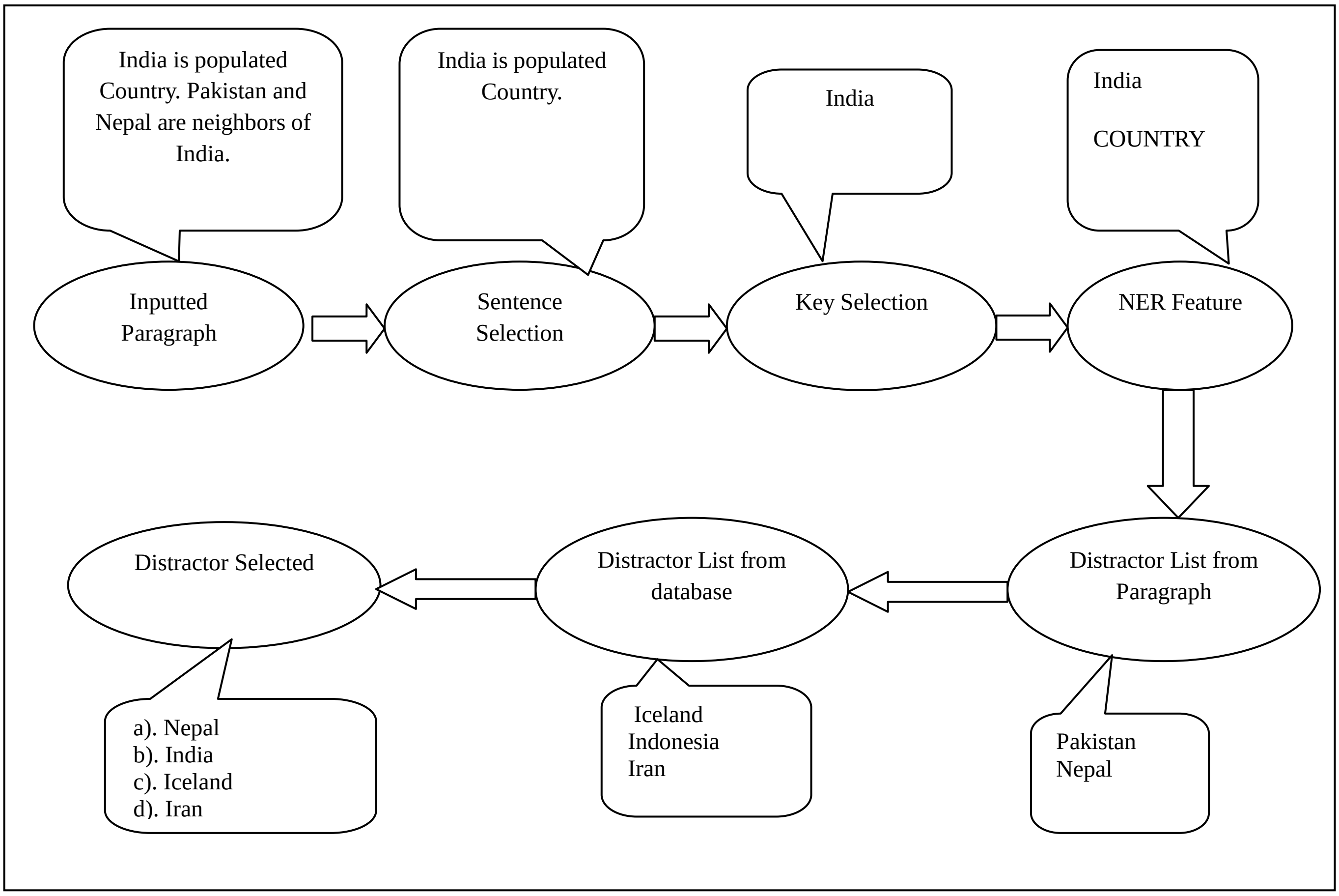


Figure 4 Process Flow of Fill in the blanks with distractors when keys is (Place)

3.4.4 Key is other (Other than Name, Location and Organization)

If key is Other than Name, Organization and Location i.e. " *Sun is huge ball of gases."* is the sentence, system will generate " *Sun* " as the key. so NER feature will identify *Sun* as OTHER so for distractors name start with "S" in the thesaurus database (which contain synonyms of the key) will fetch and set as a one distractors. one distractor will extract from the paragraph. For fetching distractor from paragraph distractor key list is generated. distractors key list will contain *noun* from the paragraph other than key is fetched from paragraph using NER feature. one is extract from table hyponyms (same pronunciation different word) i.e. Son and Sun both pronunciation same having different meaning . so if key is *other* then Distractors generated from thesaurus database, from paragraph and from hyponyms (if hyponyms are not available then fetch from paragraph distractor list) , total 3 distractors are generated and 1 correct answer of the fill in the blank question.

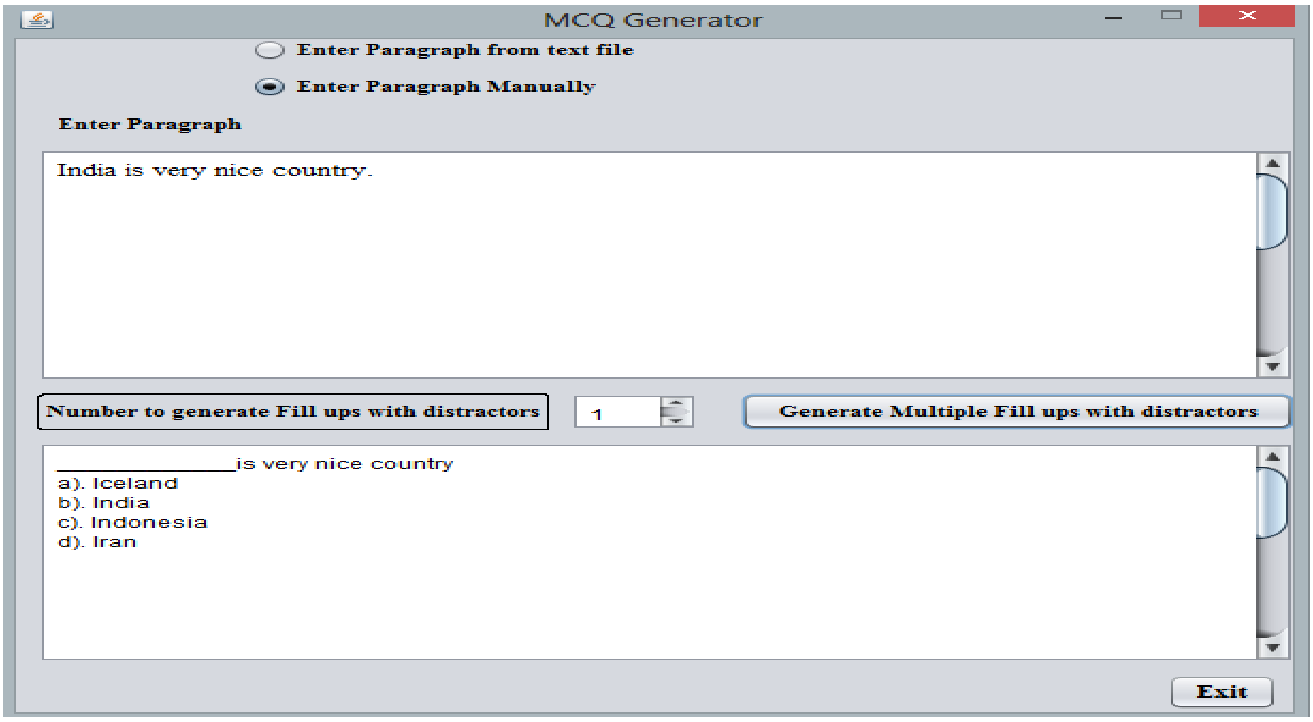
As describes in Table 5: if Key is Other then Name, Location, Organization Simple noun, Distractor key list will generate from paragraph as well as from database thesaurus and hyponyms.

Table 5 Distractor list where key is Other Noun (Not Name Location Organization)

No.	Paragraph	DistractorList from paragraph	Distractors from thesaurus and homonyms database	Final Distractors
1	The sun is a huge ball of gases. Sun is so huge that it can hold millions of planets inside it.	Millions	Amen-Ra Apollo Son (homonyms)	a). Amen-Ra b). sun c). millions d). son

As shown in Figure 5 screen shot of the system, User can enter the paragraph from the through open dialogue box and select file from the drive (doc or txt). or manually enter the paragraph.

Then user needs to enter the total number of blanks which you want to generate from the system. and Click on the button will display generated blanks with distractors on screen and store in database i.e. mysql as well as txt file will generate. user can make changes in that file if required.


MCQ Generator
Number to generate Fill ups with distractors
Generate Multiple Fill ups with distractors
Exit

○ Enter Paragraph from text file

● Enter Paragraph Manually

Enter Paragraph

India is very nice country.

Number to generate Fill ups with distractors: 1

______________is very nice country
a). Iceland
b). India
c). Indonesia
d). Iran

Figure 5: Sample screen shot of the system Fill in the blanks with distractors

4. OUTPUT OF THE TOOL

Output Questions will store in database as well as text file is generated. In text file, if user need to make some changes in the question becomes easy and In databases if user enter the same paragraph or generate the same question then, if question is already available in the database that will not store again. duplicate entry for question will not enter in database.

5. EVALUATION AND PERFORMANCE MEASUREMENT OF TOOL BY APPLING VARIOUS SAMPLE DATA

Sixty paragraph's from internet as well as from the book of essay paragraphs as been selected. And that sixty paragraphs has evaluated manually for generating objective question fill in the blanks with distractor. Same way questions are generated from that paragraph's through Tool.

For analysis of Fill in the blanks with distractor Questions, As shown in table 8: File analysis will give information about the paragraph of particular type which contain information like No of paragraph, No. of line, No. of Sentences and No. of Nouns. Measurement parameters like No. of Question generated through manual as well as through system is calculated. Time taken while generating questions manually as well as through tool. Informative and Not at all Informative Measurement is calculated manually for both manual question generation and tool through question generation.

From the sixty table below is example of table type City for Fill in the blank with distractor where data is feed about paragraph, manually number of blanks generated and time taken, Blanks generated through tool and through tool generates two types of blanks 1). Selected Blanks : In this algorithm is created through which selected blanks can be generated and chances of useful blanks is very high. It will not cover all the noun in the paragraph as a part of blanks. 2) All Noun Blanks : In this paragraph having noun. It cover all the noun in the paragraph as a part of Blanks. for example. In paragraphs 10 nouns are there than 10 blanks will be generated through this type.

Table 8 Fill in the blanks with distractor Analysis through tool as well as manually

Fill in the Blanks with distractor Analysis						
Type	**File analysis**		**Measurement parameters**	**Manual outcome**	**Output come from tool**	
					Selected	**All nouns**
City02	No of paragraphs	04	No of question	14	25	108
	No of lines	25	Time taken	20 min	1 min 41 sec	1 min 53 sec
	No of sentences	58	Informative	-	22	82
	No of nouns	108	Not at all Informative	-	03	26

Type	**File analysis**		**Measurement parameters**	**Manual outcome**	**Output come from tool**	
					Selected	**All nouns**
PHP 02	No of paragraphs	03	No of question	32	16	74
	No of lines	18	Time taken	40 min	55 sec	1 min 21 sec
	No of sentences	22	Informative	-	10	14
	No of nouns	74	Not at all Informative	-	06	57

Type	**File analysis**		**Measurement parameters**	**Manual outcome**	**Out come from tool**	
					Selected	**All nouns**
Software Engineering 01	No of paragraphs	02	No of question	27	24	123
	No of lines	23	Time taken	30 min	1 min 3 sec	3 min 17sec
	No of sentences	26	Informative	-	21	56
	No of nouns	123	Not at all Informative	-	03	67

6. CONCLUSION AND FUTURE WORK

System will select the informative sentence from the paragraph and generate fill in the blanks with distractor from the paragraph. Syntactic features from NLP parser helps to create the fill in the blanks with distractor questions from paragraphs. And For testing different paragraph's downloaded and tested through system as well as manual question were also generated from paragraph. Stil there is still much room for improvement. Firstly Comparison of Selected Blanks generation and All Noun Blank generated is remaining and Multiple Choice Question(MCQ) Question generation i.e. WH question's where question starts with Who, Where, Whom, What etc. Wh. generation is part of future work.

7. REFERENCES

[1] Pollock, M.J., Whittington, C.D., Doughty, G.F.: Evaluating the Costs and Benefits of Changing to CAA. Proceedings of the Fourth International Computer Assisted Conference CAA, http://www.caaconference.com/. (2000).

[2] Wolfe, J.H.: Automatic question generation from text - an aid to independent study. SIGCUE Outlook 10(SI) (1976)

[3] Kunichika, H., Katayama, T., Hirashima, T., Takeuchi, A.: Automated question generation methods for intelligent english learning systems and its evaluation, Proc. of ICCE01 (2001).

[4] Manish Agarwal and Prashanth Mannem : Automatic Gap-fill Question generation from text books

[5] Simon Smith, P.V.S Avinesh and Adam Kilgarriff. 2010.*Gap-fill Tests for Language Learners: Corpus-Driven Item Generation* .

[6] Nikiforos Karamanis, Le An Ha and Ruslan Mitkov: Generating Multiple-Choice Test Items from Medical Text:A Pilot Study

[7] Naveed Afzal and Viktor Pekar: Unsupervised Relation Extraction for Automatic Generation of Multiple-Choice Questions.

[8] RUSLAN MITKOV, LE AN HA and NIKIFOROS KARAMANIS: A computer-aided environment for generating multiple-choice test items(2005)

[9] Le An Ha :Multiple-choice test item generation: A demo Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev and Cristian Moldovan:The First Question Generation Shared Task Evaluation Challenge (2010)

[10] Husam Ali Yllias Chali Sadid A. Hasan: Automatic Question Generation from Sentences(2010)

[11] Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, Takeshi Yamada : Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation

[12] John Lee, *Stephanie* Seneff: Automatic Generation of Cloze Items for Prepositions (2007)