IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Automatic Fill in the Blank Question with Distractor Generation Using NLP

Arpit Agrawal[1*], Pragya Shukla[2], Aishwarya Panicker[3], Trisha Dhawe[4]

[1,2,3,4]Dept. of Computer Engineering and Information Technology, Institute of Engineering and Technology, Devi Ahilya Vishwavidyalaya, Indore, India

Corresponding Author: aagrawal@ietdavv.edu.in, Tel.: 09424090249

DOI: https://doi.org/10.26438/ijcse/v7i6.892897 | Available online at: www.ijcseonline.org

Accepted: 11/Jun/2019, Published: 30/Jun/2019

Abstract— Today, in the advancement of Information Technology, machine learning has many applications especially in the field of Natural Language Processing. Thus with the help of NLP and algorithms of machine learning, we can automatically generate fill in the blank questions. Thus the task of manually constructing questions is no more a burden. An algorithmic approach has been deduced in this model to generate fill in the blank questions. A paragraph will be provided from which we have to select sentences with relevant content so that these sentences can be considered as options to generate fill in the blanks. The sentences with meaningful information are chosen for question generation. The questions will have relevant blanks which can be filled by one amongst the four choices provided just like multiple choice questions. Each question will have only one choice as the correct answer. The rest of the three choices will be wrong answer known as distractors.

Keywords— NLP-Natural Language Processing, NER-Named Entity Recognition, POS-Part-of-Speech, RNN-Recurrent Neural Network, NLTK-Natural Language Toolkit, GloVe-Global Vector

I. INTRODUCTION

In this paper, with the help machine learning and python libraries we formalize an algorithmic approach to generate fill in the blank questions. Previously the fill in the blank questions with multiple choices for answers were generally constructed manually. So, this approach used to consume a lot of time. Thus we try to reduce the burden of teachers by introducing this system of automatic fill in the blank questions with distractor generation using NLP. It is used for the assessment of knowledge of students in an examination by educational institutes. The questions can be constructed from any given paragraph from text books or any other source of information. This system will try to find out most significant sentences from any given corpus. These sentences will then be used to create blanks. The blanks are generally nouns and superlatives. These questions will be provided with multiple choices making them a lot easier to be evaluated. There will be four choices provided out of which one will be the key and the other three are going to be the distractors or wrong answers to these questions. They are named as distractors because these alternatives distract the students from the right choice. NER and POS tagger are used to extract the syntactical and lexical features from the sentences.

For example:

For the given paragraph, we use our model to generate fill in the blank questions as output-

Generative Adversial Networks or **GANs** are used for unsupervised machine learning. They are **neural networks** that are comprised of **two** computing models which are synchronized to function together. Its role is to capture and copy variations in dataset. It is used for image manipulation and generation and understanding risks. It majorly has two components i.e. the **generator network** and the **distributer network**. **Generator network** learns distribution of classes. **Distributer network** learns boundaries between classes.

_____ learns distribution of classes.

a) Neural network
b) Generative Adversial Network
c) Generator network

II. RELATED WORK

- **RevUP Automatic Gap-Fill Question Generation**

This paper describes RevUP which deals with automatically generating gap-fill questions. It uses a distribution based ranking method for sentence selection. For gap-selection, a discriminative binary classifier was trained on human annotations and a novel method was proposed for generating semantically-similar distractors with contextual fit.

- **Factual open cloze question generation for assessment of learner's knowledge**

In this paper, a system is presented which can generate fill-in-the-blank questions without alternatives. The system firstly extracts a set of informative sentences from the given input corpus and then secondly omits the answer-keys which are selected by identifying domain specific words in the sentences.

- **Automatic Question system of Multiple Choice Cloze Questions and its Evaluation : Takuya Goto**

In this paper, a statistical method of generating multiple-choice cloze questions automatically is proposed. Based on the machine learning and statistical patterns of existing questions, the system is able to select sentences which are appropriate to multiple-choice cloze questions from texts and generate various types of blank parts with distracters. Based on the experimental results, the system is proved to select correct sentences and blank parts.

- **Automatic Gap-fill Question Generation from Text Books : Manish Agarwal**

This GFQG system selects most informative sentences of the chapters and generates gap-fill questions on them. Syntactic features helped in quality of gap-fill questions. They evaluated the system on two chapters of a standard biology textbook and presented the results.

- **An Automatic Gap Filling Questions Generation using NLP : Miss. Pranita Pradip Jadhav**

This paper shows initial exploring experiments towards creating an automatic question. System selects the descriptive sentence from the paragraph and generates fill in the blanks with distractor from the paragraph and with the help of Wikipedia. Most of these methods use rule-based approaches which are heavily reliable on a manually crafted feature set. The questions are not fluent and grammatically incorrect. As some of the earlier approaches focused on frequency, they get a number of sentences with blank space and choosing the best blank space is a challenge.

III. METHODOLOGY

This model uses a basic approach for the generation of fill in the blank questions. It involves four steps which are mentioned below.

a) Processing Data
b) Sentence Selection
c) Key Selection
d) Distractor Selection

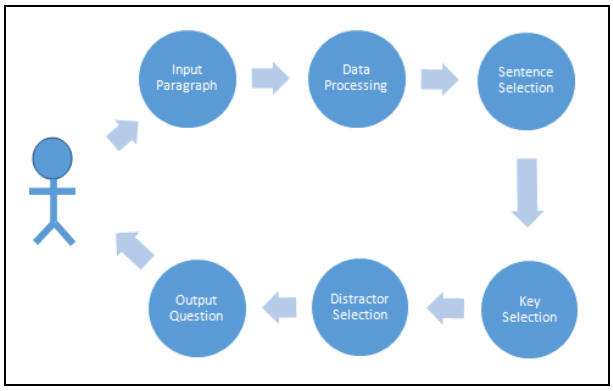

Figure 1. Flow Chart

III.I Data Processing:

In this step a paragraph is provided. From the given paragraph all the sentences are read with the help of the module. Tokens are created by breaking down these sentences into smaller parts. (give example of tokens here). Then parts of speech of these sentences are identified i.e., Nouns (NN), Pronouns (PN), Adverbs (AV), Adjectives (AJ), etc by using a POS tagger

III.II Sentence Selection:

The sentences with content and some features are important. Thus for better sentence selection, all the sentences from the paragraph are looked upon and informative sentences are identified. Thus to select appropriate key for generating relevant questions, we can use the selected sentences.

- **Summarization:** In this step, a single line summary of the text is generated. This provides us the gist of the passage in order to generate questions related to it. This sentence is denoted by SUMM.
  For summarization we may use word2vec or GloVe. Each word from the sentence is listed and converted into vectors with the help of word2vec. Then Word embeddings are generated.

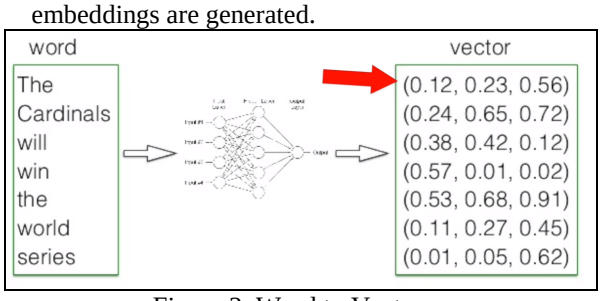

Figure 2. Word to Vector

Sequence2sequence model is used. The Encoder is given input sequence as an input. On passing the vectors through Decoder, output sequence is generated.

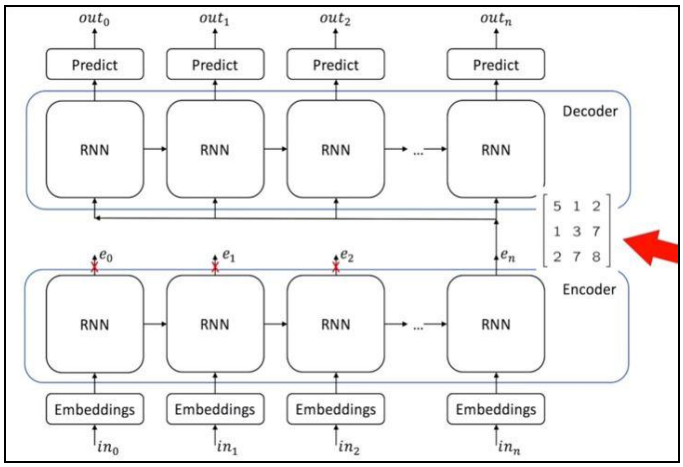

Figure 3. Recurrent Neural Network

- **Common Tokens:** Each sentence of the text is compared with the Summary sentence, SUMM. The number of words common to both the sentences is counted. A sentence with words from the SUMM in it is important and is a good choice to ask a question using the common words as the key. Based on the number of common words, the 'Feature' parameter is calculated.

$$\text{Feature} = \frac{\text{common words between sentence and summary}}{\text{Total words in selected sentence}}$$

Greater the value of 'Feature', more relevant the sentence is. The sentences are selected for question generation based on their 'Feature' value. This step is performed using NLTK and Word2vec.

Example –

SUMM = India finally got **freedom** in the year **1947** and celebrated victory**.**

Sentence = **1947** has been the year of struggle for all the **freedom** fighters in **India**.

Feature = 3/14

The blank can be at India, freedom and 1947 as these are the common words.

- **Additional steps :**

1. Count of Nouns – The number of nouns in a sentence gives an idea about the degree of importance of sentence. Large number of nouns in a sentence means that potential key can be generated from that sentence and that sentence has good content which can generate an appropriate key the blanks.

2. Superlative - Superlative degree is related to exaggeration. Generally superlative degree words end with the suffix – est. (E.g. wealthiest). Sometimes the word most is used to indicate superlative degree (E.g. most beautiful).For generation of potential fill in the blank, sentences containing superlative degree words can be used.

Example: Indore is the **cleanest** city in India.

3. Count of Words – For generation of good fill in the blanks content is required. So very long sentences are preferred as they have enough content for question generation. Sentences which have short length are discarded because of lack of content in them. So we find length of the sentence by counting number of words in the sentence.

ALGORITHM:

Enter the paragraph P.
Create line summary as Summ.
Read the statements S from the paragraph P.
For each S
    Compare S and Summ
    Count the number of common words
    Compute the value of 'Feature'
End For
Compare the values of 'Feature' and give priority to the sentences accordingly.

III.III Key Selection:

In this step we identify the most appropriate key from the selected sentence to ask question upon. We can select the first cardinal as the key.

For example: There are **six** members in my family.

Key selection approach is divided into two stages-

- Generate Potential keys from the statement and create key list.
- Select Best Key from that key list.

Stage 1: With the help of sentence selection module we select appropriate sentences for generating potential keys. Now a POS tagger is used to identify and label the parts of speech in the sentences.

For example:

Here we select nouns from the provided sentence and generate keys. So we list all the nouns in the key list.

Delhi, **Kolkata** and **Bangalore** are the metropolitan cities of **India**.

Here, Delhi, Kolkata, Bangalore and India are listed in the key-list.

Stage 2: Select a word from the key-list and search that word in paragraph. For selecting the best key we use the following three features-

1. **Stage Frequency** - Count the frequency of occurrence of that word in the paragraph. Select best key with highest frequency.
2. **Summary** - If the word is present in the summary of the paragraph, then it is the best key for generating questions.

3. **Height of Syntactic Tree** - If the word has maximum height in the syntactic tree then it forms the best key.

ALGORITHM:

For Each Word from selected sentence
Do
Generate the potential key and
Add them into the key list KL.
KL = Select Noun from Selected sentence
BK = No of Occurrence of that key in paragraph
or
Key common to SUMM
or
Height of syntactic tree

End For
Remove that Best Key BK from selected sentence

III.IV Distractor Selection:

Finally, to find the grammatical structure of sentences so that distractors could be selected, we use the NER feature of the Stanford parser. Its NER feature further labels various word sequences present in the sentence into various categories of nouns such as the name of a person, place, animal, institute, etc.

There exists many pair of words that have the same spelling but mean different. Such words are known as homonyms. For example, let us take two sentences-

1). It was a bright and sunny morning.
2). Your child is very bright.

Here, the word bright means "filled with light" in the first sentence. On the other hand it means "very smart or intelligent" in the second one.

Therefore, the system must select distractors for such keys very wisely. Choosing relevant distractors belonging to the same context, as the key is necessary. This is where the NER feature comes in. The key selected in the key selection process is categorized by the Stanford parser into a particular category and then distractors are selected from the given corpus for the key.

1) Key is Name of the Person

Suppose the key selected is categorized as the name of a person. In such a case, all the names in the database can be selected as the distractor. To select some out of all the possible choices, the system chooses names that particularly start with the same alphabet as that of the key. For example, the sentence "Shri Ram Nath Kovind is the 14th president of India." is selected from the president_database. Then the key will be "Ram Nath Kovind". So the names in the president_database like Pratibha Patil, A.P.J. Abdul Kalam etc. will be selected. The names like R. Venkataraman and Rajendra Prasad will definitely be selected as the distractors by the system since they start with the letter "R".

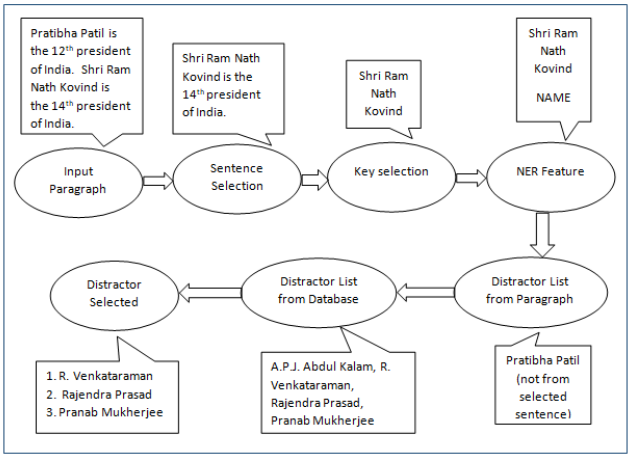

Figure 4. Process Flow of Fill in the blanks with distractors when key is Name of person

2) Key is Name of the Organization

Suppose the key selected is categorized as the name of an organization. In such a case, all the names of organizations in the database can be selected as the distractor. To select some out of all the possible choices, the system chooses names that particularly start with the same alphabet as that of the key. For example, the sentence "Bharatiya Kisan Sangh organization is not as good as compared to Centre for Ecology and Rural Development Organization" is selected from the organization_database."Bharatiya Kisan Sangh "will be the key. So the names of organizations in the organization_database will be selected. The names like Bharatiya Gorkha Parisangh will definitely be selected as the distractors by the system since they start with the letter "B".

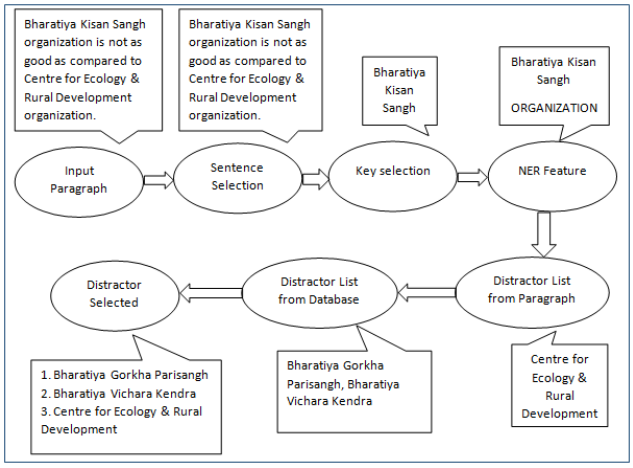

Figure 5. Process Flow of Fill in the blanks with distractors when key is Name of an organization

3) Key is Place (City State or Country)

Suppose the key selected is categorized as the name of a place. In such a case, all the places in the database can be selected as the distractor. To select some out of all the possible choices, the system chooses places that particularly start with the same alphabet as that of the key. For example, the sentence "New Delhi is the capital of India." is selected from the capital_database. Then the key will be "New Delhi". So the places in the capital_database like Canberra, Moscow etc. will be selected. The places like Nassau and Nagpur will definitely be selected as the distractors by the system since they start with the letter "N".

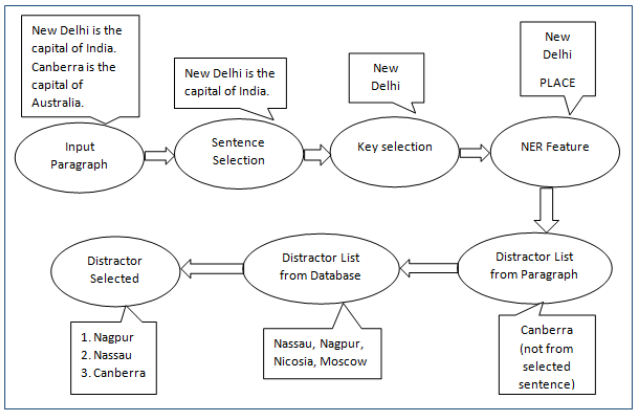

Figure 6. Process Flow of Fill in the blanks with distractors when keys is Name of a place

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

Suppose the key selected is categorized some other proper noun. In such a case, all the nouns of that category of noun in the database can be selected as the distractor. To select some out of all the possible choices, the system chooses nouns that particularly start with the same alphabet as that of the key. For example, the sentence "The Taj Mahal is the iconic symbol of love." is selected from the monument_database. Then suppose the key is "love". So the nouns in the monument_database like peace, knowledge etc. will be selected. The nouns starting with the letter "l" will definitely be selected as the distractors by the system.

Table 1. Distractor list where key is other noun.

No.	Paragraph	Distractor List from Pararaph	Distractors from Database	Final Distractors
1.	The Taj Mahal is the iconic symbol of love.	love	peace hate knowledge	a). peace b). love c). hate d). knowledge

IV. RESULTS AND DISCUSSION

We finally generate relevant fill in the blank questions by firstly selecting the sentences with good content from the paragraph and then generating potential keys. By using the syntactic features from NLP parser, distractors are also created for fill in the blanks. It is not easy for automatically generated questions to be as good as questions generated by human experts, so this approach is helpful in improving the quality of questions.

V. CONCLUSION AND FUTURE SCOPE

There are four features on the basis of which we generate questions which are very less. There is possibility of more features. Selection of first cardinal as a key is not always correct because the second cardinal in the sentence can be equally or more important. Since the number of features in this methodology is less, a hypernym or hyponym of the key could find way into the distractors list thereby providing a confusing list of distractors.

Here in this approach we have a paragraph. In future we can work on very large database. Comparison of Selected Blanks generation and All Noun Blank generation can be worked upon. The questions starting with Who, Where, Whom, What etc. are the WH questions which can be worked upon in future work.

ACKNOWLEDGMENT

Every endeavor, we understand takes an indomitable urge, perseverance and proper guidance especially when it is most needed. Internally motivated to take some appreciable work as our degree project, we came to undertake this project.

We would like to use this opportunity to thank with gratitude our respected project guide Mr. Arpit Agrawal, Lecturer Department of Information Technology for his motivation to take this endeavour and perform. We would also like to thank Dr. Vrinda Tokekar, Head of Department Information Technology for supporting us. Lastly we would like to extend our gratitude towards Dr. Sanjeev Tokekar, Director Institute of Engineering and Technology (DAVV), Indore for providing us a platform to make this project.

REFERENCES

[1] Wolfe, J.H.: Automatic question generation from text - an aid to independent study. SIGCUE Outlook 10(SI) (1976)
[2] Kunichika, H., Katayama, T., Hirashima, T., Takeuchi, A.: Automated question generation methods for intelligent English learning systems and its evaluation, Proc. of ICCE01 (2001).
[3] Simon Smith, P.V.S Avinesh and Adam Kilgarriff. 2010.Gap-fill Tests for Language Learners: Corpus-Driven Item Generation .

[4] Naveed Afzal and Viktor Pekar: Unsupervised Relation Extraction for Automatic Generation of Multiple-Choice Questions.
[5] RUSLAN MITKOV, LE AN HA and NIKIFOROS KARAMANIS: A computer-aided environment for generating multiple-choice test items(2005) .
[6] Husam Ali Yllias Chali Sadid A. Hasan: Automatic Question Generation from Sentences(2010)
[7] Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, Takeshi Yamada : Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation

AUTHORS PROFILE

Arpit Agrawal is a senior lecturer in the Department of Computer Engineering. His research areas are Natural Language Processing, Machine Learning, Data Mining and Algorithms.

Pragya Shukla is a senior lecturer in the Department of Computer Engineering. Her research areas are Artificial Intelligence and Analysis of Algorithm.

Aishwarya Panicker is pursuing Bachelor of Engineering from Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore and is expected to complete it in 2019. She is pursuing her engineering from Department of Information Technology. She holds a cross domain interest in Machine Learning, Algorithms and Computer Organization and Architecture.

Trisha Dhawe is pursuing Bachelor of Engineering from Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore and is expected to complete it in 2019. She is pursuing her engineering from Department of Information Technology. Her main fields of interest are Data Science, Machine Learning and Data Mining.

