Question generation using NLP

Designed and built by:-
Jay Vaghani (ID-15CP057)
Jay Movaliya (ID-15CP066)
Baraiya Vivek (ID-15CP202)

4th Year, Computer Engineering

Prepared under the guidance of:-

Prof. Mosin I. Hasan

Prof. Hemant Vasava

Birla Vishvakarma Mahavidyalaya

Department of Computer Engineering

Vallabh vidyanagar-388120, Gujarat

Academic Year: 2018-2019

CERTIFICATE

This is to certify that the Project entitled "**Question generation using NLP**" has been carried out by Jay Vaghani(15CP057), Jay Movaliya(15CP066) and Baraiya Vivek(15CP202) under my guidance in partial fulfilment of the degree of Bachelor of Engineering in Computer Engineering (Fourth level) from Birla Vishvakarma Mahavidyalaya,Vallabh Vidyanagar during the academic year 2018-19.

Internal Guide Head of the Department

Prof. Mosin I. Hasan **Prof. Hemant Vasava**

ACKNOWLEDGEMENTS

We have taken sincere efforts in this 7th semester's Final Year project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

We are highly indebted to Birla Vishvakarma Mahavidyalaya and heart fully to our project guides for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

And finally, we would like to thank our colleagues who helped us in developing the project and people who have willingly helped us out with their abilities.

INDEX

1. Problem statement:

How would someone tell whether the **content** is **yours**? They might ask you to summarize it or describe how it relates to your own research.At first, they might just ask you questions to check that you remember the basic facts.

We aim to create a system for question generation (QG) that can take as input an article of text or reports that is in doc or pdf (e.g., a web page or encyclopedia article that a teacher might select to supplement the materials in a textbook), and create as output a ranked list of factual questions.

Or, as a first step, just to see if you bothered to get through it, they might ask you what features we used in our ranking model, or what corpora we tested on. That is, at first, they might just ask you questions to check that you remember the basic facts. Then, if it is clear that you read more than the title and abstract, they might move on to more challenging questions.

2. Problem summary:

As per our research, the following problems are generally faced today:-

- There’s no way to know that the factual information in the report or research paper by an individual is its his/her owns.
- There’s no any system that can generate question on based fn the factual information.

3. Literature Survey:

In the literature survey, we have read 1 thesis, papers in which, all paper has implemented an algorithm for detection in different ways. Brief information about all these research paper are shown below:-

1. Automatic Factual Question Generation From Text(2011)[1]. This thesis gives the overview and methods used for generating the questions from factual text.

2. A selection strategy to improve cloze question quality(2008)[2]. In this paper, different ways of imporving the cloze procedure were given.

3. Neural question generation from Text: A Preliminary Study[3]. This paper gives the idea of generating questiong by training a neural network based RNN model.

Also, we have gained so much information about Natural Language Processing from different resources. The resources are listed below:

1. Sentdex course on NLTK library from YouTube.
   - This course was an introduction to the NLTK library with its implementation in python

2. The Definitive Guide to Natural Language Processing
   - This article give some basic overview to what is NLP and some relating terms like Machine Translation, Automatic Summerization, Abstraction - Text generation with its own words, Extraction- Generally copy the text from the article,

4. IMPLEMENTATION

Methodologies:

All the methodologies are related to NLTK library:

1. **Text Extraction:** Generally the text can be in any form – e.g., an article, passage, report or a research paper.

2. **Text Preprocessing:**
   a) Tokenization
   b) Stop words removal

3. **Term Frequency:**
   a) **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

      TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

4. **Sentence Weight:** Frequency of each of the terms in the sentence and then the aggregate all the term frequencies is taken out.

5. TOOLS REQUIRED:

1.NLTK
2.Python
3. Tf-Idf

6. CONCLUSION

Thus, using NLTK and Tf-Idf we were able to find the term frequencies from each sentences of the given passage. And also from this we were able to find the weighted sentences.

7. FUTURE IMPROVEMENTS:

Using RNN to train a model to generate question with the above weighted sentences and term frequencies.

8. REFERENCES:

[1] Automatic Factual Question Generation From Text(2011) - Michael Heilman
[2] A selection strategy to improve cloze question quality(2008) - Pino, J., M. Heilman, M., and Eskenazi, M. (2008)**.**
[3] Neural question generation from Text: A Preliminary Study - Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, Ming Zhou.