Information Mining System in Bengali Language

G M Sohanur Rahman
18101590
Dept. of CSE, BRAC University
g.m.sohanur.rahman@g.bracu.ac.bd

Sayed Md. Rahmat Ulla
18101620
Dept. of CSE, BRAC University
sayed.md.rahmat.ulla@g.bracu.ac.bd

Md. Motahar Mahtab
18301023
Dept. of CSE, BRAC University
md.motahar.mahtab@g.bracu.ac.bd

Md Shahriyar Hossain
21141017
Dept. of CSE, BRAC University
md.shahriyar.hossain@g.bracu.ac.bd

Mohammed Julfikar Ali Mahbub
18301264
Dept. of CSE, BRAC University
mohammed.julfikar.ali.mahbub@g.bracu.ac.bd

Abstract

The Information Retrieval System is an effective tool that uses Natural Language Processing to help a user find important information (NLP). They proposed an algorithmic Information Retrieval Scheme(BIRS) based on information in this research paper, and the system is mathematically and statistically significant. Two algorithms for finding the lemmatization of Bengali terms, Trie and Dictionary Based Search by Removing Affix (DBSRA), are demonstrated in this paper and contrasted with Edit Distance for the exact lemmatization. They demonstrated the Bengali Anaphora resolution scheme, employing the Hobbs algorithm to obtain the correct knowledge language. The TF-IDF and Cosine Similarity are established as the behavior of question-answering algorithms to find the correct answer from the documents. They implemented a Bengali Language Toolkit (BLTK) and Bengali Language Expression (BRE) in this report, making our mission easier to implement. They also created a corpus of Bengali root words, synonym words, stop words, and a corpus of 672 articles from the famous Bengali newspaper 'The Daily Prothom Alo,' which we added. They generated 19335 questions from the presented data to validate the method and received a 97.22 percent accurate response.

1 Introduction

The term "information retrieval" refers to the process of retrieving data from a variety of databases using a related query. It is a science that allows you to scan for information inside a paper and search for text, pictures, and sounds. Full-text or content-based searching is the most popular method of searching. For historical purposes, IR stands for "information retrieval" (Singhal, 2001). The Recommended System, which is closely linked to the Information Retrieval System, functions without a query. It is a software environment that allows users to read documents such as books, newspapers, and magazines. This framework can be used to store and handle documents. Today, the most visible IR application is online search engines. Newspapers, social networking networks, and other blogs contain a massive amount of material every day. One of the first methods proposed was to use text words to quantify the likelihood of importance. Other approaches analyze text metrics in papers and surveys before constructing the term weighting scheme. The BM25 ranking algorithm performs well in a variety of tasks (Robertson and Zaragoza, 2010). As a result, they are primarily interested in using TF-IDF and cosine similarity to solve problems. The value of each word in a sentence was calculated using TF. IDF determined the fundamental importance of the terms in a text. The relationship between questions and sentences was then determined using cosine similarity. Their main goal is to retrieve relevant information with high precision in a limited amount of time.

2 Related Work

The study of information retrieval is so satisfactory in English language such as Web Information retrieval (Agichtein et al., 2006), the retrieval of the picture and video (Sivic and Zisserman, 2003)and text retrieval method. These are the commonly used studies in the field of Information retrieval. In comparison with English, Bangla language is much legging behind. In recent years, we have seen a few researches have been performed on summarization, Bangla Sentence Extraction (Sarkar, 2012), Sentiment analysis and customer feedback portal (Hassan et al., 2016) etc. Apart from these works, using mathematics and statistics, we present an information recovery system on Bangla language.

3 Proposed Work

In this paper, they introduced a Bengali Information Retrieval System(BIRS), based on Bengali Natural Language Processing. they were following a 3 steps process to complete the task. At the beginning, they collected five types of corpus similar to Bengali root words,stop words and questions then preprocessed their collected data and lastly, to find out the relationship between questions and answers, they used Cosine Similarity. To deal with cosine similarity, they converted the datasets into vectors using TF-IDF model.

3.1 Corpus

They have used five types of corpus.The first has more than 28 thousands root words used to lemmatize Bengali words. The second corpus contains 382 Bengali stop words, it was used to eliminate all the unnecessary information from the documents using 672 informative articles from various sources. The third corpus based on technology from the popular news portals such 'The Daily Prothom Alo'. Furthermore, nearly 19500 information collected for fourth corpus. And lastly, to avoid synonymous words, more than 18000 similar words were gathered. Besides, few other corpus were used for unknown word processing, verb processing, removing punctuation and others.

Lets see an example of a category information,

বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সালের ১৭ মার্চ টুঙ্গিপাড়া গ্রামে জন্মগ্রহন করেন। তার রাজনৈতিক জীবন শুরু হয়েছিল ১৯৩৯ সালে মিশনারি স্কুলে পড়ার সময় থেকেই। ভারত পাকিস্তান বিভক্ত হওয়ার পর, পূর্ব পাকিস্তানের উপর পশ্চিম পাকিস্তানের অন্যায় অবিচার বাড়তে থাকে। এজন্য তিনি ১৯৬৬ সালের ৫ ফেব্রুয়ারী লাহোরে বিরোধী দলসমুহের একটি জাতীয় সম্মেলনে ঐতিহাসিক ৬ দফা দাবি পেশ করেন যা ছিল পূর্ব পাকিস্তানের স্বায়ত্তশাসনের পরিপূর্ণ রূপরেখা। অবশেষে তিনি ১৯৭১ সালের ২৬ মার্চ বাংলাদেশের স্বাধীনতার ঘোষণা দেন

for this portion of corpus, the category questions will be,

Question-1:বঙ্গবন্ধু শেখ মুজিবুর রহমান কোথায় জন্মগ্রহন করেন?

Question-2:কত তারিখে তিনি ৬ দফা দাবি পেশ করেন ?

3.2 Pre-Processing

They have tokenized the informative documents to sentence tokens, removed the punctuations and stop words, found out the synonymous words for each word. They have used Hobb's algorithm for anaphora resolution and lemmetized the bengali words using DBSRA. This preprocessing has helped reduce their execution time and improve performance.

3.2.1 Anaphora

Anaphora is a process where a noun is replaced with other words (pronoun) for use in context. Anaphora resolution or coreference resolution aims to find out which pronoun is related with which noun. They have used Hobbs' algorithm for detecting the anaphora of the document taking into account the roles of subject, object and reflexive and possessive pronouns. Here is an example: বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সালের ১৭ মার্চ টুঙ্গিপারা গ্রামে জন্ম গ্রহণ করেন। তার রাজনৈতিক জীবন শুরু হলেঙ্গে ১৯৩৯ সালে মিশনারি স্কুলে পড়ার সময় থেকেই। Here, 'তাঁর' (his) is the pronoun of 'বঙ্গবন্ধু শেখ মুজিবুর রহমান' (noun).

3.2.2 Tokenization

Tokenization is the process of breaking a document into words, characters or sentence. They have used the BNLP toolkit to efficiently tokenize our document into sentence tokens and mapped each tokenized sentence to the root sentence.

3.2.3 Cleaning

They have removed punctuations(comma,question mark,semi-colon, colon, exclamation mark) from the sentences using Bangla Regular Expression (BRE) tool as most of the punctuations are unnecessary. They used punctuations corpora of BNLP toolkit.

3.2.4 Stop Words Removing

Stop words refer to the words that do not affect the overall meaning of the sentence. For instance, Bengali language stop words such as এবং (and), কোথায় (where),অথবা (or), তে (to), সাথে (with) are have no linguistic importance in natural language processing. They have used BNLP toolkit to efficiently remove the stop words from the documents using their pre-built stop word corpora.

3.2.5 Lemmatization for Bangla Language

Lemmatization is a process of transforming a word into its root word. In Bengali natural language processing, there are few verbs that cannot be lemmatized by any system because of the limitation of lemmatization algorithms.They have lemmetized Bengali words by combining three techniques e.g. Dictionary-Based Search by Removing Affix, Trie algorithm and Levenshtein distance. As each algorithm has some limitations, they have combined three of them for maximum performance with lowest execution time and space. Their propose plan is as follows:

At first, they have found out the lemma using dictionary based search and trie. If both lemmas match, they returned the lemma. If they do not match, they found the lemma that has the smallest Levenshtein distance to the target word. Then they calculated the probability $P(edit|word)$ (here word is the target word, not the root word). If the probability is greater than 50%, they take the target word as unknown word; otherwise they return the lemma. They have bult a corpora that contains words that included suffix of Bengali Language like তে (te), ছে(che), য়ের(yer), etc. They removed the longest matching suffix from the unknown word and returned the word.

3.2.6 Synonyms Words Processing

Moreover, there are possibilities when a question contain words which are not available in the system but there exist similar meaning question. Synonym means different words but same meaning. Therefore, to handle this unwanted situation, a synonyms word corpus were constructed containing total 13,189 words.

After pre-processing the sentences and question, they found the following,

Sentence 1: বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯২০ সালে ১৭ মার্চ টুঙ্গিপড়া গ্রাম জন্ম গ্রহণ করা

Sentence 2: বঙ্গবন্ধু শেখ মুজিবুর রহমান রাজনীতি জীবন শুরু হয় ১৯৩৯ সাল মিশনারি স্কুল পড়া সময় থাকা

Sentence 3: ভারত পাকিস্থান বিভক্ত হয় পূর্ব পাকিস্থান পশ্চিম পাকিস্থান অন্যায় অবিচার বাড়া থাকা

Sentence 4: বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯৬৬ সালে ৫ ফেব্রুয়ারী লাহোর বিরোধ দল জাতী সম্মেলন ঐতিহাসিক ছয় দফা দাবী পেশ করা থাকা কার্য পূর্ব পাকিস্থান স্থায়ত্তশাসন পূর্ণ রূপরেখা

Sentence 5: অবশেষে বঙ্গবন্ধু শেখ মুজিবুর রহমান ১৯৭১ সালে ২৬ মার্চ বাংলাদেশ স্বাধীন ঘোষণা দেওয়া।

Question 1:বঙ্গবন্ধু শেখ মুজিবুর রহমান জন্ম গ্রহণ করা

Question 2:তারিখ বঙ্গবন্ধু শেখ মুজিবুর রহমান ছয় দফা দাবী পেশ করা

4 Algorithms

4.1 TF-IDF

TF, Term Frequency, is the count of each word/ term in a corpus, it gives equal importance to all words. On the other hand IDF, Inverse Document Frequency, finds actual importance of term(s) in a sentence. The product of TF and IDF, TF-IDF, converts sentences into vectors and determines how important a word is to a given document in the corpus.

After normalization of the data, firstly, the term rule has been ensured to measure the value of TF in every pre-processed sentence. Secondly, IDF has been used to figure out a relevant sentences.

Finally, TF-IDF has been calculated by multiplying TF with IDF of only the words re-

lated to the input questions to reduce time and space complexity.

4.2 Cosine Similarity

Cosine similarity measures the cosine angle(judgement of the orientation, not magnitude) between two vectors. Cosine similarity of two non-zero vectors:

$$\frac{E_{i=1}^{n}(A_i * B_i)}{\sqrt{E_{i=1}^{n}{A_i}^2} * \sqrt{E_{i=1}^{n}{B_i}^2}}$$

Using TF-IDF,cosine similarity (the relation between) of questions were calculated. From the derived results, question 1's answer is 59.7% related to sentence-1 and 1.6% related to sentence-2. In this similar fashion it compares all possible combinations of questions and answers and finds the optimal relations. Thus, the answers of corresponding questions stays in the corresponding sentence.

5 Experiments, Tools & Final Results

The experiments were performed in Anaconda Distribution and Python 3.6 tools. A Bengali language processing tool and Bengali Language toolkit was also used to clear data and remove stop words. NLTK was also used in multiple pre-processing tasks.

Among the collected 672 articles, 19334 question were created for testing. The BIRS system correctly answered was 18797 and incorrect was 537, accuracy was 97.22% . The performance measured was noticeably good and flexible with lowest time complexity.

6 Conclusion and Future Work

The Bengali Information Retrieval System was established using numerous algorithms and methods like Anaphora resolution procedure, TF-IDF and Cosine Similarity. The whole study was processed in Bengali Language as a part of BNLP. The BIRS system was thoroughly tested, and the results were noted accordingly with expected optimal results.

Future plans include improvement of the system purposes like education, industry and business. Also implementation of deep learning algorithms for development of the system.

References

E. Agichtein, E. Brill, S. Dumais, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR 2006*, pages 19–26, New York, NY, USA. ACM.

A. Hassan, M. R. Amin, N. Mohammed, and A. K. A. Azad. 2016. Sentiment analysis on bangla and romanized bangla text (brbt) using deep recurrent models. volume abs/1610.00369.

S. Robertson and H. Zaragoza. 2010. The probabilistic relevance framework: Bm25 and beyond. volume 3, page 333–389.

K. Sarkar. 2012. Bengali text summarization by sentence extraction. volume abs/1201.2240.

A. Singhal. 2001. Modern information retrieval: A brief overview. volume 24, pages 35–43.

J. Sivic and A. Zisserman. 2003. Videogoogle: A text retrieval approach to object matching in videos. volume 2, page 1470–1477, Washington, DC, USA. IEEE Computer Society.