# Information Retrieval In Assamese Using Wordnet & Assamese Wikipedia

M P Bhuyan, R Purkayastha, S K Sarma, S Sarmah, P Sarma, V Deka

**Abstract**— Importance of Information Retrieval is increasing day by day. Non English languages are also getting importance in information retrieval, because people are more comfortable in their native language than English. India is a country of different culture and different language. Assamese is one of the recognized languages of India, which is the official language of Assam. In this research work, a Assamese Information Retrieval (IR) system is designed. The IR system has three phases 1) Structuring the query, 2) Use of Assamese WordNet in query expansion, 3) Accessing information from Assamese Wikipedia. The IR system is reliable and performance of the system is 60.08 % in terms of relevant document retrieval.

**Index Terms**— Assamese WordNet, Assamese Wikipedia, Information Retrieval (IR), Query Expansion.

## 1 INTRODUCTION

Assamese is an Indo-European language spoken in the north-east region of India [1]. There are around 15 million native speakers yet at the same time; very little progress can be seen when compared with other non-English languages. For digital computing, it is necessary to develop the computing technology related to the Assamese. This language is recognized as a regional language in eight schedules of the Indian constitution. In other states of India like Arunachal Pradesh and Nagaland Assamese language is widely spoken. Assamese is likewise spoken in a couple of nations like Bhutan.

Natural Language Processing (NLP) is a subfield of computer science, information engineering, computational linguistics, and artificial intelligence which focus on the enhancement of the interactions between computers and human languages. NLP is computationally very complex and challenging for researchers across the world. The main challenge in NLP is how to program computers to process and analyze large amounts of natural language data.

Assamese WordNet is a large lexical database which is designed by referring Hindi WordNet. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each communicating a particular idea. Synsets are interlinked by the method of conceptual-semantic and lexical relations. Assamese WordNet works as a resource in this research work.

The WordNet contains ID, Category, Concept, Example and the Synset. The Synset field is very crucial for the proposed IR system. Synset words which are relevant to user query are used to expand the query. So that the system can easily understand what the user is looking for.

The purpose of an IR system is to provide the answer or result of a query given by a user into the system. The IR system starts working when a user fires a query to the system. The IR system compares the user's query with the resources in its collection and returns the information which is highly related to the user's query. The IR system maps various parameters of the user's query with its collection and this is a partial to exact matching process, which leads to a result of the user's query. In the proposed IR system the query is matched with the WordNet database(specifically keywords present in synsets) if the words(tokens) of the query matches with any word of the synset then most frequent synset words are added to the query for retrieval of maximum documents which will me more relevant to the user query.

In this proposed-model to retrieve data from Assamese Wikipedia, we used Wikipedia python API (Application Programming Interface). Wikipedia API can be used to search Wikipedia page titles, article summaries, links and images from a Wikipedia page. After query expansion phase each word is passed to Wikipedia API and relevant documents have been retrieved and presented to the user.

Assamese Wikipedia is used as the knowledge base for the proposed IR system. Assamese Wikipedia is the Assamese language edition of Wikipedia. Currently, Assamese Wikipedia has 5,602 articles with 22,840 registered users and the community of contributors is still growing in this platform.

## 2 RELATED WORKS

Assamese IR system for retrieving the users' information by comparing the users' query and the documents at the lexical level. Shorter queries cause some problems like a lexical mismatch and to improve the query additional information like phrases related to the query, synonyms are added, the Assamese


- *M P Bhuyan is currently pursuing PhD degree in Information Technology in Gauhati University, India, E-mail: mpratim250@gmail.com*
- *R Purkayastha has completed (2019) BTech degree in CSE in Gauhati University, India.*
- *S K Sarma is currently working as a Professor in Information Technology department in Gauhati University, India.*
- *S Sarmah is currently working as an Assistant Professor in Information Technology department in Gauhati University, India.*
- *P Sarma is currently working as an Assistant Professor in Information Technology department in Gauhati University, India.*
- *V Deka is currently working as an Assistant Professor in Information Technology department in Gauhati University, India.*

IR is based on the vector space model, the original query is called a query vector and this query vector is extended by using the synset of the Assamese WordNet. Assamese IR system can satisfy the user by providing the most relevant information related to the user's query [1].

Language models based on the N-grams are used to design some predictive model to analyze the contextual information of the words and finally, keystrokes saving results are targeted in the research, the whole research work uses the basic building element N-gram as a context search. Unigram, bigram and trigram models are used and explored extensively to achieve the goal of the research work [2].

Designing of WordNets in various languages becoming a crucial task for the people in the field of linguistics and Computation. This is due to the digitization of information and the Internet. Regional languages are not developed very much in the computation field; Initiatives are taken from both at Government level and Academic level. Development of Assamese word has started a part of a Government project on the development of Indian languages. Assamese WordNet describes the complete details of a word by providing synset ID, concepts, characteristics, example, etc. Also, mapping has been done from Assamese to Hindi and English. A common interface is used for all the Indian languages to develop the WordNet of each language [4].

A quantitative comparison is done for every synset position of the Assamese WordNet by looking at the occurrence of these synsets (sets of synonymous words) in a corpus of around 1.5 million words, Sarma et al. (2014) research was an attempt to focus on the time-line of each synset of the WordNet-based on this 1.5million size corpus, the time-line is covered from 1900 to 2008, the first five entries of the synsets are found most frequently used this time period, timeline analysis is shown graphically for proper visualization and analysis [5].

Cross-Lingual Information Retrieval (CLIR) systems provide an interface to the user to trigger his query in one language and retrieve the result in another language, a CLIR system is designed for the farmer of Tamil Nadu, India, and the user of the CLIR system will send his query in Tamil and the result will be returned in English. The research work uses the morphological analyzer, machine translation from Tamil to English, bilingual dictionary, Name Entity Recognition, etc., the CLIR system is having some adaptive learning technique, like if a new word is encountered in the machine translation the bilingual dictionary is updated subsequently [6].

In [7], authors have used term proximity for query expansion. The term proximity is used to determine the similarity of the documents with the queries. They have divided the query expansion into the statistical approach and semantic-based approach. They have used Spectral-Based Information Retrieval Model (SBIRM) to rank the documents.

In [8], the authors have described a query expansion model for information retrieval, they have given an overview of information retrieval methods, and they have focused on the query expansion technique like Kullback-Leibler Divergence (KLD) and synonym search using the WordNet. For term proximity, spectral analysis technique was used.

In [9], they have mentioned the up-gradation of the information retrieval system due to advanced users' queries. They have analyzed the effect of query expansion on Urdu language using Kullback-Liebler (KL) model and Bose-Einstein1 (Bo1) Model & Bose-Einstein2 (Bo2) Model. KL model was found superior to the other two models which were able to enhance the Mean Average Precision (MAP) value by 22% to 24%.

The research work describes two-level information retrieval one at monolingual and the next is cross-lingual, cross-lingual information retrieval uses the query in Hindi or Bengali and the documents are returned in English. Hindi and Bengali queries are transliterated based on phoneme and the equivalent queries in English are formed and automated query generation, Machine Translation techniques are used in the research work. Query expansion, query refinement, Word Sense Disambiguation, Multi-word Expression such techniques are used to improve the performance of the IR system [10].

## 3 METHODOLOGY

The proposed IR system consist of following phases:

*1. Structuring the query:*

1.1. Tokenization
1.2. Removal of punctuation and numbers
1.3. Removal of stop words
1.4. Stemming

*2. Use of Assamese WordNet in query expansion.*

*3. Accessing information from Wikipedia:*

3.1. Set Language
3.2. Extract the Wikipedia page titles
3.3. Extract the entire page

These phases are explained below:

### *1. Structuring the query*

*1.1 Tokenization:*

The first step is to convert the query sentence into tokens. Tokenized form of the query will be searched in the WordNet. For example, given a piece of text: " মানুহজন আগতে নিজৰ ঘৰলৈ গ'ল।" , "maanujn aagote nijor gharlei gol." in English "At first the person went to his home" it outputs [ মানুহজন, আগতে, নিজৰ, ঘৰলৈ, গ'ল, ।],

*1.2. Removal of punctuation and numbers*

Query words matching in WordNet and Wikipedia are

independent of the presence of punctuation and numbers. In this step, we removed the punctuation and numbers. For example, the output of tokenization [ মানুহজন, আগতে, নিজৰ, ঘৰলৈ, গ'ল, ।] is stripped down to [ মানুহজন, আগতে, নিজৰ, ঘৰলৈ, গ'ল].

*1.3. Removal of stop words*

Stop words are the words which itself does not provide any information, but helps to complete the sentence. After removing the stop words the tokens left are [ মানুহজন, নিজৰ, ঘৰলৈ, গ'ল].

*1.4. Stemming*

In this step stemming operation is performed on the remaining tokens to the root form (e.g. মানুহজন→মানুহ). For example, after stemming operation the list of tokens becomes this: [ মানুহ, নিজৰ, ঘৰলৈ, গ'ল]

Query (list of the token) is now ready to be searched in Synset of Assamese WordNet.

## *2. Use of Assamese WordNet in query expansion.*

In our proposed model of Information Retrieval (IR) system which is more focused on query expansion. In this proposed system we are using WordNet for query expansion. In this proposed system Assamese WordNet helps by automatically extending the user query by adding semantically related words which are originally not present in the query so that the most relevant documents can be retrieved plus more information can be gathered for a specific query.

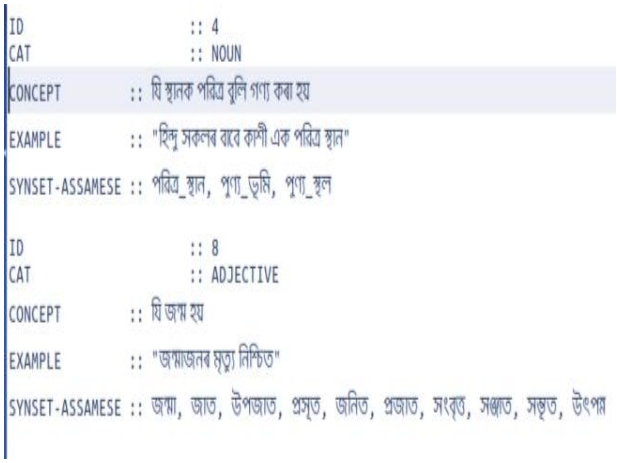


**Fig. 1.** A small section of WordNet

Assamese WordNet has five fields named as ID, CAT(category), CONCEPT, EXAMPLE, SYNSET-ASSAMESE. SYNSET-ASSAMESE contains a set of all the synonyms.

Query sentence after phase 1 is now changed to list of tokens. All the tokens are now searched in synset of Assamese WordNet simultaneously. If the tokens are found in the synset then we retrieve the maximum of three most frequent words present in the synset and add it to the original query.

If none of the tokens matches with the any of the words presents in synset of Assamese WordNet then we do not modify the query and directly move to phase 3 (forward the list of tokens to Wikipedia API to retrieve data which is relevant to the list of the tokens/ structured query).

## *3. Accessing information from Wikipedia*

After phase 2 the original query might be modified. In this proposed model we are using Assamese Wikipedia as our knowledge base from where we will retrieve the information and display it to the user. Each word of the query is now ready to be passed to the Wikipedia API.

*3.1. Set Language*

Before retrieving the data we have to first set the language on which we want to retrieve the information.

To do this below is the function (python code): wikipedia.set_lang("as")

"as" : For Assamese Language.

*3.2. Extract the Wikipedia page titles*

Each word of the query is searched in Wikipedia simultaneously and relevant Wikipedia page titles are retrieved. Here for a particular query word it is not guaranteed that it will find any relevant pages.

To retrieve the Wikipedia page titles which are relevant to the word we have use below function (python code):

syno_word = "ঈগল"

In English ("ঈগল" -> Eagle)

pg_titles = wikipedia.search(syno_word), pg_titles is a list of retrieved page titles

*3.3. Extract the entire page*

Now for a particular query word we have a list of all the relevant page titles. In this step, we will extract the whole content of each of those page titles.

Below is the code to perform this task (python code): wikipedia.page("ঈগল") #entire wiki page print(page.content)

After this phase, the user has displayed all the retrieved information from Wikipedia based on the user input query.

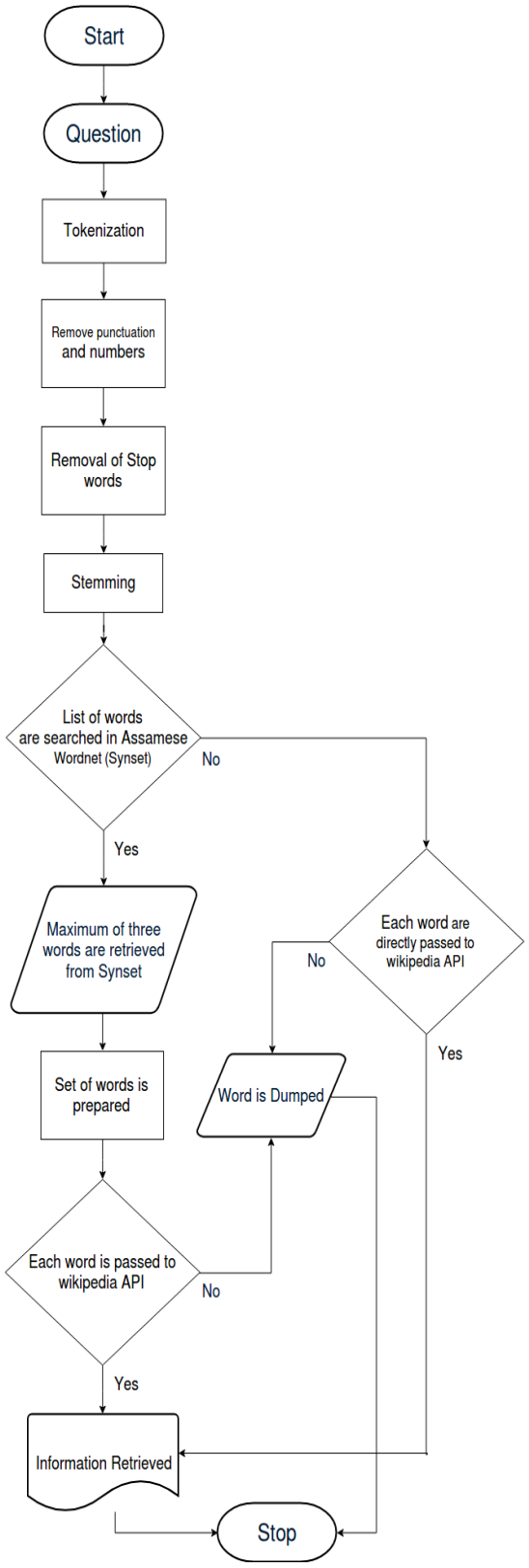


Fig. 2. Flow-chart of the proposed model

## 4 Result and Discussion

The proposed IR system was tested with lots of queries. A sample of a few queries has been shown in *Table 1*. Performance of the proposed IR system found satisfactory.

**Performance Measure:**

The following formulas are used to evaluate the performance of the proposed IR system:

$$\%\,of\ RelevantDocument = \frac{Relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \times 100$$

$$\%\,ofIrrelevantDocument = \frac{Irrelevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \times 100$$

TABLE 1
EXPERIMENTAL RESULT OF PROPOSED IR SYSTEM

| Query | Found in Word Net | Found in Wikipedia | Document Retrieved | Relevant Document | % of relevant document |
|---|---|---|---|---|---|
| Q1 | Yes | Yes | 56 | 35 | 62.5 |
| Q2 | Yes | Yes | 9 | 3 | 33.3 |
| Q3 | Yes | Yes | 19 | 10 | 52.6 |
| Q4 | No | Yes | 10 | 10 | 100 |
| Q5 | No | Yes | 10 | 9 | 90 |
| Q6 | Yes | Yes | 22 | 5 | 22.7 |
| Q7 | Yes | Yes | 7 | 3 | 42.8 |
| Q8 | Yes | Yes | 36 | 8 | 22.2 |
| Q9 | Yes | Yes | 9 | 8 | 88.8 |

| | | | | | |
|---|---|---|---|---|---|
| Q10 | Yes | Yes | 6 | 4 | 66.6 |
| Q11 | No | Yes | 10 | 9 | 90 |
| Q12 | No | Yes | 25 | 11 | 44 |
| Q13 | Yes | Yes | 1 | 1 | 100 |
| Q14 | Yes | Yes | 23 | 11 | 47.8 |
| Q15 | Yes | Yes | 21 | 16 | 76.1 |
| Q16 | Yes | Yes | 28 | 23 | 82.1 |
| Q17 | Yes | No | 0 | 0 | UNDEF |
| Average | | | | | 60.08 |

Q1 = "অসম কি?"
Q2 = "গুৱাহাটী বিশ্ববিদ্যালয় ক'ত?"
Q3 = "বাঘ"
Q4 = "বিহু কি?"
Q5 = "ভাৰতবর্ষ ক'ত?"
Q6 = "কলম কি?"
Q7 = "পেপাৰ কি?"
Q8 = "ঈগল চৰাই কেনেকুৱা?"
Q9 = "গুৱাহাটী ক'ত?"
Q10 = "কম্পিউটাৰ কি?"
Q11 = "ভাই�ৰাছ কি?"
Q12 = "ভূপেন হাজৰিকা সেতু"
Q13 = "আস্তাবল কি?"
Q14 = "লোকসেৱা কি?"
Q15 = "লোকসভা কি?"
Q16 = "মুদ্রা কি?"
Q17 = "উপগ্রহ কি?"

## 5 Conclusion and Future Works

In this research work, an IR system for the Assamese language is designed and different level of queries are fired by expanding the queries with the help of WordNet and the relevant documents are retrieved from Assamese Wikipedia. The IR system is relatively stable and retrieving capacity is 60.08 %. However, in the case of a few queries, the system performance is outstanding but in few queries, performance has dropped abruptly. This is due to the Wikipedia corpus where the query is not found. In the future, a large corpus can be designed to reduce the failure rate and secondly multiple web links can be searched for large size queries. A better ranking system like the Vector Space Model can be used to display using the most relevant documents first.

## References

[1] Golok Chandra Goswami. 1982. "Structure of Assamese", Gauhati University, Guwahati, Assam.

[2] A. K. Barman, J. Sarmah and S. K. Sarma, "WordNet Based Information Retrieval System for Assamese", 2013 UKSim 15th International Conference on Computer Modelling and Simulation, Cambridge, 2013, pp. 480-484. doi: 10.1109/UKSim.2013.90

[3] M. P. Bhuyan & S. K. Sarma (2019), "An N-gram based model for predicting of word-formation in Assamese language", Journal of Information and Optimization Sciences, 40:2, 427-440, DOI: 10.1080/02522667.2019.1580883

[4] S. K. Sarma, M. Gogoi, R. Medhi, U. Saikia: "Foundation and structure of developing an assamese wordnet". In: Proceedings of 5th International Conference of the Global WordNet Association (GWC-2010), Department of Computer Science,Gauhati University (2010).

[5] Shikhar Sarma, et al. "A Quantitative Analysis of Synset of Assamese WordNet: Its Position and Timeline", Proceedings of the Seventh Global WordNet Conference 2014, pp: 246–249.

[6] D. Thenmozhi and C. Aravindan, "Tamil-English Cross Lingual Information Retrieval System for Agriculture Society", International Forum for Information Technology in Tamil Conference, October 2009

[7] S. Alnofaie, M. Dahab, M. Kamal, "A novel information retrieval approach using query expansion and spectral-based," Int. J. Adv. Comput. Sci. Appl. 7(9), 2016, pp. 364–373.

[8] M.Y Dahab, S. Alnofaie, M. Kamel, "A Tutorial on Information Retrieval Using Query Expansion," In: Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham, 2018. pp. 761-776

[9] I. Rasheed and H. Banka, "Query Expansion in Information Retrieval for Urdu Language", 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), Kota Kinabalu, 2018, pp. 1-6. doi: 10.1109/INFRKM.2018.8464762

[10] D. Mandal, S. Dandapat, M. Gupta, P. Banerjee, S. Sarkar: "Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources". In A. Nardi, C. Peters, eds.: Working Notes for the CLEF 2007 Workshop.