Anup Kumar Barman
Shikhar Kumar Sarma

JEDNOJĘZYCZNE POZYSKIWANIE INFORMACJI W LOKALNYM JĘZYKU: PRZYPADEK JĘZYKA ASSAMSKIEGO

[**słowa kluczowe:** pozyskiwanie informacji, język assamski, przetwarzanie języka naturalnego]

Streszczenie

Duża ilość informacji zawsze implikuje potrzebę dobrego systemu wyszukiwania. Badania nad pozyskiwaniem informacji (IR) stają się bardzo ważne, ze względu na ogromny wzrost cyfryzacji informacji. System wyszukiwania informacji dostarcza najbardziej istotne informacje z dużej kolekcji, na podstawie zapytania użytkownika. Ze względu na konieczność znalezienia odpowiednich informacji, badania nad pozyskiwaniem informacji rozpoczęto od 1950 r. Wdrożono kilka systemów IR w zależności od charakteru informacji i użytkowników. Znalezienie najistotniejszych informacji w oparciu o zadane zapytanie w lokalnym języku jest celem pobierania informacji jednojęzycznych. W wielojęzycznym kraju, takim jak Indie, gdzie istnieje 23 języków urzędowych, cyfryzacja treści w językach lokalnych ogromnie wzrasta. Aby zaspokoić zapotrzebowanie na istotne informacje każdej osoby, jednojęzyczny wskaźnik IR w lokalnym języku jest bardzo istotny. Poniżej analizujemy podstawowy wymóg opracowania jednojęzycznego IR. Omawiany tutaj system IR jest implementowany dla języka Assamskiego, który jest jednym z zaplanowanych języków Indii. Wydajność pobierania statystycznego systemu IR można zwiększyć dzięki informacjom językowym generowanym w różnych aplikacjach przetwarzania języka naturalnego.

* * *

Anup Kumar Barman, Shikhar Kumar Sarma

MONOLINGUAL INFORMATION RETRIEVAL IN LOCAL LANGUAGE: CASE STUDY IN ASSAMESE

[keywords: Information Retrieval, Assamese Language, Natural Language Processing]

Abstract

Large amount of information always implies the need of a good retrieval system. The research on Information retrieval (IR) is become very important due to the tremendous growth of digitalized information. Information retrieval system provide the most relevant information from a large collection based on the user query. For the necessity of finding relevant information the research on Information retrieval has been started from 1950. Several IR systems were implemented depending on the nature of information and users. Finding the most relevant information based on the fired query in their own language is the aim of monolingual information Retrieval. In multilingual country like India where 23 official languages exists digitalize local language contents are growing tremendously. To meet the need of each individual's relevant information the monolingual IR in own language is very essential. Here we analyze the basic requirement of developing the monolingual IR. The IR system discussed here is implemented for Assamese Language which is one of the scheduled language of India. The retrieval efficiency of a statistical IR system can be enhanced using linguistic information generated through various Natural Language Processing applications.

Introduction

Information Retrieval is a discipline that deal with the retrieval of relevant information from a collection of structured and unstructured documents depending on the user query. The use of an IR system in various application varies based on the user interest. The user's interested information need to be expressed through a query. Documents that can satisfy the user query in the judgment of the user are said to be relevant and others are non-relevant. An IR engine may use the query to classify the documents in a large collection, and then return a subset of documents to the user that satisfy some classification criterion. Natu-

rally, the higher the proportion of documents returned to the user that he/she judges as relevant, the better the classification criterion.

Alternatively, an IR engine may "rank" the documents in a given collection measuring the relevancy. IR systems can be categorized by the scale at which they operate. In web search the system has to produce the relevant document by searching over billions of documents stored on various servers. Indexing helps to make an efficient system at this enormous state. At the other end is personal information retrieval where the documents are less in number and the user are also limited. In between this two one another type of IR system is present which may be called as institutional charge where retrieval might be provided for collections such as office internal documents, a database of patient information etc. Many Natural Language Processing applications like stemmer, Parts Of Speech Tagger, Named Entity Recognizer, transliteration, Multiword Expression recognizer, clunker, Word sense disambiguation, snippet/ summary generation etc. have been used in Information Retrieval to enhance the performance.

"As We May Think" the article published in 1945 by Vannevar Bush [1] initiated the idea of automatic access from large amount of digitalized data. Several new innovations and works were introduced in between 1945-1955 which lead the idea of searching text automatically to a mature stage. In 1960 the SMART system was developed by Gerard Salton and his student at Harvard and Cornell University [2] which allowed researchers to experiment with ideas to improve search quality. In 1970 to 1980 many developments built on the advances of the basic idea. Later in 1996 to 1998 various search algorithms were developed on IR field which can be employed in large scale data or in World Wide Web. The rapid growth of on line data has urged scientist to investigate new techniques and technologies to develop large scale search engine with high throughput.

The information may be structured, unstructured and semi-structured formats. If a record have a name component and organized according to some well--defined syntax then such type of information are called structured type. For example in a relational database there may be multiple record types but all records of a given type have the same syntax. On the other hand in a collection of unstructured natural language documents there is no well-defined syntax to organize the document. There is no simple well-defined way of telling where the required information occurs in a given document. Some time in a collection the documents may share a common structure but the data does not occupy well--defined columns in a well-defined table such type of documents are called semi--structured documents.

The representation of content of a document in a searchable format usually called the indexing process is the first step of a conventional IR system. This is done without concerning the end users. Indexing process transform a document in to a set of important terms which is directly understandable by the IR model. The query formulation is the second step of in IR system where the user's request to an IR system are processed. In third step the similarity is calculated between the represented documents and formulated query. Then produce the list of retrieved document according to the similarity score.

An IR model comprises the all three steps required to deal with the document representation, the query representation and the retrieval functionality. In semantic IR approach implementation is done based on some degree of syntactic and semantic analysis. In statistical IR approaches, the documents that are retrieved or that are highly ranked are those that match the query most closely in terms of some statistical measure. To fit the documents and queries properly in any one of IR model various preprocessing operations were performed. Stop word removal, stemming, Lemmatization, Normalization are some notable preprocessing tasks. There are various indexing and searching techniques are available to implement an IR system. The web crawling is a process through which we can collect the web pages to index. Crawler is a software to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them.

Methodology

This work concentrate on the discussion of structure and pipeline of Monolingual Information Retrieval System along with their important Natural Language Processing tasks considering the Assamese Monolingual Information Retrieval System as a case study. In order to implement the Assamese IR the open source software platform SOLR [3] – the search engine interface to the Apache Lucene [4] search library and NUTCH [5] – web crawler were used. [6] is a web crawler built in JAVA language that supports almost all features of Nutch and is suited for the purpose of Multimedia text retrieval.

Then the various language and informative resources are use as plugins in that system. Computational linguistic resources of this Indo-Aryan language is less compared to other and hence to incorporate the resources in IR we faced various challenges.

The architecture of our Assamese Monolingual Information Retrieval System is explained in figure 1.

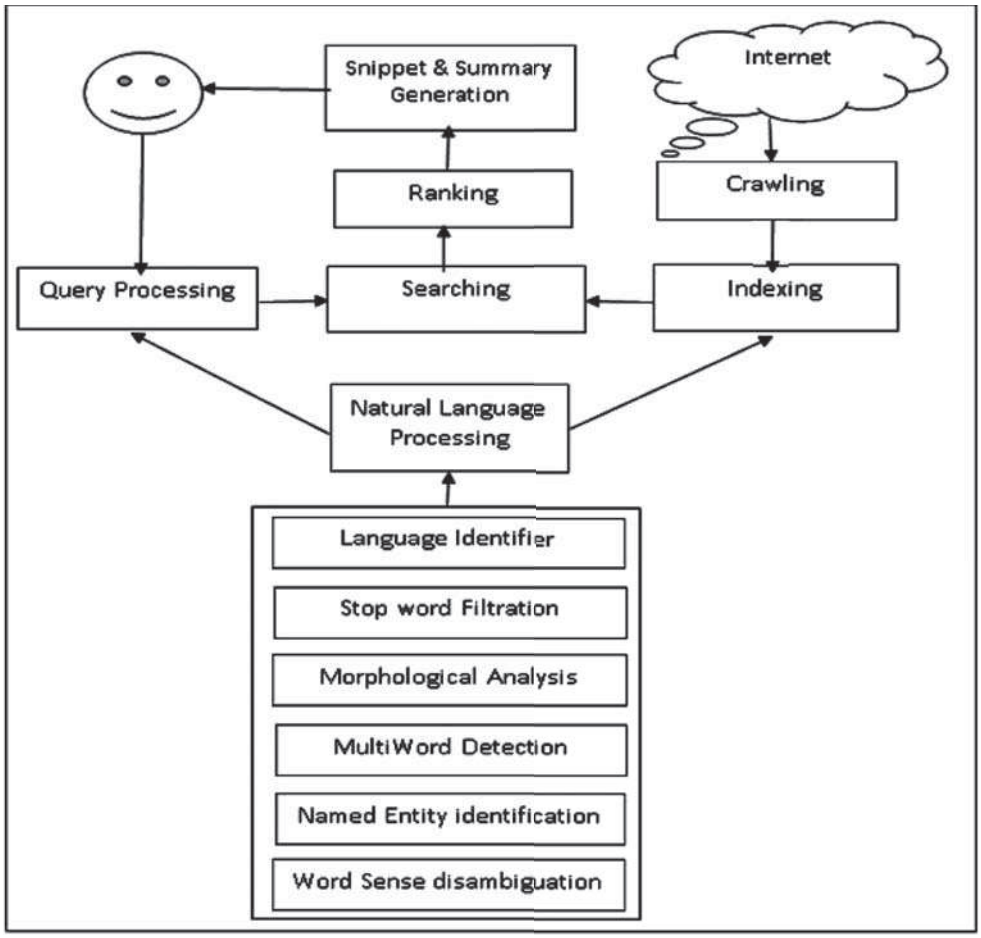

Figure 1. IR System Architecture

The various resources prepared to incorporate with the Information Retrieval module are discuss as follows:

Processing resources

Language Identifier: Sharing same script by different languages introduce the need of Language identifier (LI). We developed a LI system to exactly identify the Assamese language digitalised text.

Morphological Analyser: A rule based Morphological analyser was developed for Assamese language to obtain the lexicon. An accuracy of 85% was achieved by using 23 morphotactic stemming rules.

Multiword detection: To identify the collocations which indicate a single concept, the POS tagging is an important pre-processing task. [7] To prepare a statistical MWE detection system having Precision score 73, we first develop one statistical POS tagger and took help from Assamese WordNet. Parts of speech [8] information for words are use as basic information to detect the collocations. Assamese WordNet [9] the huge lexical resource helps in finding those MWEs.

Named Entity Recognizer [10]: Named Entities plays a vital role in information extraction. A statistical NE recognizer was developed to identify the Assamese Named Entity with an accuracy of 73%.

Language Resources

Corpus: Corpus [11] is a collection of various genre of texts. It is the basis of all NLP tasks. Assamese corpus consists of 1.5 million words of genre Arts, Science, Politics, History, Sports etc.

Spell Variation List: In conventional writing process some word's spelling may vary time to time. All spelling variation are acceptable as correct one. We prepare a list comprising of 5170 such entities.

Multiword expression list: Multiword Expressions (MWEs) are sequence of words separated by space or delimiter which determines a unique meaning instead of words' individual meanings A list comprising of 1627 Multi-word Expressions have been identified for Assamese language.

Stop-word list: The words like "The", "and" do not convey any important semantic information in context of IR and so those are filtered as stop-words. A list of 264 Assamese stop-words list has been prepared in consultancy with the Linguistic experts.

Assamese Dictionary: The root word list named as dictionary consists of 15,750 words was created.

Named Entity List: A list comprising of 105905 (One lakh five thousand nine hundred five) Assamese named entities was created.

Informative Resources

Query Set: We prepared a set of Assamese query comprising of Regional and General Queries to evaluate the IR system.

URL data base: A database comprising of seed and blog URLs were stored for crawling purpose.

Results and Discussion

To determine the effective performance of an IR system various evaluation methods are used. Evaluation are done in different phases of the IR system- query processing evaluation, ranked retrieval results evaluation and snippet evaluation. After finding an acceptable accuracy for each modules we use the evaluation metrics [12] P@K (Precision at K results) to measure the performance of our system. Figure: 2 shows the overall results after calculating P@5 and P@10 for 20 regional queries.

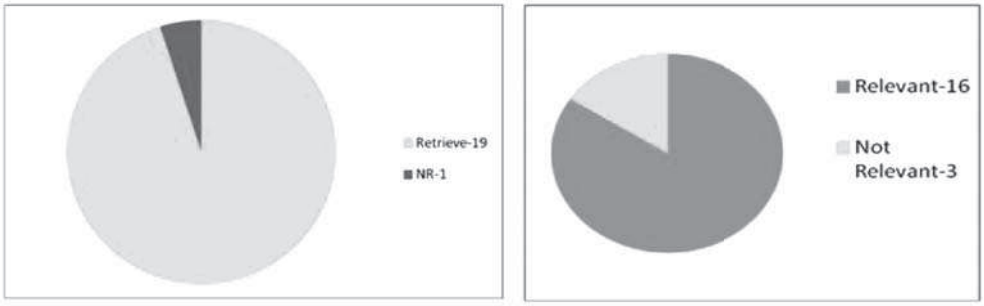

Figure 2. IR Results

Where NR: No Results found means for that query no URL is retrieved and Not Relevant indicates that all the p@5 and p@10 average value is 0. The 1st part of figure 2 shows that for only one query no result found and the 2nd part depict the relevant and irrelevant content among the retrieved URLs.

Conclusion

Information Retrieval system retrieves relevant documents based on the user query. The development of IR system for Assamese language is discussed in this paper. Various NLP components in the form of language resources are added as a plug-in to the IR system for improving its effectiveness. This work summarizes the fact of developing monolingual IR System achieving state-of-art accuracy for a resource scarce language along with the various NLP components integral to IR.

References:

1. Vannevar Bush. As We May Think. Atlantic Monthly, 176:101–108, July 1945.
2. Gerard Salton, editor. The SMART Retrieval System–Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.
3. Apache Solr(2011)http://lucene.apache.org/solr/
4. Apache Lucene(2011)http://lucene.apache.org/
5. Apache Nutch(2005)http://nutch.apache.org/
6. Apache Heritrix(2012) https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
7. Barman A.K, Sarmah J, Sarma S.K., Automatic Identification of Assamese and Bodo Multiword Expressions. In Proceedings of Second International Conference on Advances in Computing Communications and Informatics (ICACCI2013), IEEE, Mysore, India.
8. Barman A.K, Sarmah J, Sarma S.K., POS Tagging of Assamese Language and Performance Analysis of CRF++ & fnTBL Approaches, In Proceedings of the 15th International Conference on Computer Modelling and Simulation, UKSim 2013, IEEE, Cambridge, UK.
9. Sarma S. K, Medhi R, Gogoi M, Saikia U, Foundation and Structure of Developing an Assamese Wordnet. In Proceedings of GWC 2010.
10. Sharma P, Sarma U, Kalita J., The first Steps towards Assamese Named Entity Recognition, Brisbane Convention Center Brisbane Australia, 2010.
11. Sarma S. K, Bharali H, Gogoi A, Deka R, Barman A.K., A Structured Approach for building Assamese Corpus: Insights, Application and Challenges, In Proceedings of 10th Workshop on Asian Language Resource.
12. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999.