MULTI-LINGUAL INFORMATION RETRIEVAL USING DEEP LEARNING

Sonam Sanjogkumar Dodal
Government College of Engineering, Aurangabad
sonamdodal@gmail.com

Pallavi V. Kulkarni
Government College of Engineering, Aurangabad
pallavi.k11@gmail.com

Abstract*— The task of finding data files related to an information need from a group of information resources is known as Information Retrieval. In this work, the author propose a multi-lingual information retrieval system using deep learning. Input to the system is a question in sentencing form that can be processed by NLP tools. In the preprocessing phase, part-of-speech tagging of the input sentence is performed. A three layer neural network is used for creating word to vector representation. The word2vec model continuous-bag-of-words (CBOW) is used for this purpose. Then related words are obtained via word-2-vec using deep learning RNN. RNN is the recurrent neural network. Finally, results are obtained by calculating the cosine similarity score. For multi-lingual results, bilingual mapping is performed using CFILTs bilingual corpus. The tourism dataset is used for experimentation purposes.

Index Terms— Natural Language Processing, Information retrieval, Recurrent neural network, Learning.

I. INTRODUCTION

NLP is an essential part of artificial intelligence. It enables the system to understand human speech as it spoken. There are several applications of NLP, out of which information retrieval (IR) is a hot research area. These information retrieval systems are also recognized as text retrieval systems. It helps user to obtain information related to or close to the information they wants. Information Retrieval utilizes several NLP methods such as stemming, compounding, word sense disambiguation, de-compounding, part-of-speech tagging, chunking and so on [1].

An IR process starts with a query entered by user into the system. Queries are nothing but sentences of data needs. In IR the question does not uniquely recognize a solo object in the group. Instead numerous objects may match the question with various grades of relevancy. It is contradictory to the traditional SQL queries of a database, in which outcomes returned may or may not match the query. Therefore results are usually ranked.

There are two types of information retrieval: The first is cross-lingual information retrieval (CLIR), in which the language used for querying is different than the language of documents retrieved [2]. The second is Multi-lingual information retrieval (MLIR) which includes asking questions and retrieving documents in more than one languages. The proposed system explained section III is also a MLIR system.

The remaining paper is arranged as: Section II contains history of existing systems. Section III gives a short explanation of the developed system. Section IV shows system architecture. The algorithm used by the author is given in Section V. Section VI shows the experiments and results obtained from the system. At last, section VII focuses on the conclusions of this work.

II. LITERATURE REVIEW

Early seventies was an active era for database research. Meanwhile many systems have been developed. Few of the developed systems are deliberated below:

2.1 LUNAR System

This system is developed by W.A. Woods in early seventies (1973). As in [3] LUNAR uses two databases: first for chemical analysis and second for referencing literature. LUNAR is based on ATN parser [3]. According to [4] the accuracy of LUNAR is 78% which can be improved to 90% if errors were modified.

2.2 CHAT-80

In the eighties, CHAT-80 system is a very famous NLP system [9]. This was implemented using Prolog [7]. It was an inspiring, effective, cultured system. As in [8] CHAT-80 uses the database of facts like rivers, seas, oceans and cities.

2.3 NALIX

The NALIX developed in 2006. As in [9] this system uses XML database and X-Query as query language [9]. It

uses the concept of keyword searching. NALIX builds a reversed-engineering method [9].

2.4 GINLIDB

A system named GINLIDB was invented in 2009 [10]. As in [11] UML is used for designing and VB.net 2005 is used for developing this system.

All the systems explained above uses structured database. Structured databases are easy to maintain for small set of information. As data size increases tremendously, it becomes difficult to store and maintain the data in relational databases. In such situation unstructured database such as document or text file plays an important role. Information retrieval helps us to obtain needed data from these text files. The IR systems developed until now takes input in English and returned documents in English. Such systems are difficult to use for those people whose domain or mother-tongue is not English. Thus there must be a system which support searching files in multiple languages. To fulfill this need, the author propose a multi-lingual information retrieval which supports three languages: English, Hindi and Marathi. This system overcomes the challenges posed by relational databases by querying over unstructured data using NLP and deep learning techniques.

III. PROPOSED SYSTEM

The proposed system is a Multi-Lingual Information Retrieval using Deep Learning. It supports three languages, English, Hindi and Marathi. User gives input in sentencing form which is processed by NLP tools. In the preprocessing step, user's query is transformed into annotated form with part of speech tags. The word to vector representation is created using three layer neural network. Then related words are obtained via word-2-vec using deep learning RNN. Finally, results are obtained by calculating the cosine similarity score. For multi-lingual results, bilingual mapping is performed before calculating cosine similarity score. The tourism dataset is used for experimentation purposes. Section IV will explain the detail architecture of the proposed system.

IV. SYSTEM ARCHITECTURE

System architecture is the detail view of the system. Figure 1 shows the system architecture. The system works in five phases. After passing these five phases, the input is converted into the output.

- Language selection and querying
- Preprocessing
- Bi-lingual mapping and multi-lingual searching
- Searching for the results
- Displaying results

In the first phase, user select language in which he want to obtain the results. After selecting language, he enter the query in human interaction i.e. sentencing form.

In the preprocessing phase, initial training is given to the system. All the dictionaries are loaded in this phase. The synset dictionary consist of all the nouns, verbs, adverbs and adjectives is obtained. Then part-of-speech tagging (POS) of input is performed. To perform part of-speech tagging of English query, Stanford NLP framework is used; whereas for Hindi and Marathi query, Indo wordnet framework ILT is used. In this system we focus on ten POS tags: NP, NNP, NNS, NN, NNPS, JJ, JJS, JJR, RBS, RB and RBR for proper nouns, plural nouns, singular nouns, common nouns, proper noun plural, adjectives, adjective superlative, adjective comparative, adverb superlative, adverb and adverb comparative respectively. Steaming is used to choose one single form of a word instead of different forms.

4.1 Word to Vector Representation:

Word2vec converts text into numerical form that deep neural network can understand [12]. After performing part-of-speech tagging, word to vector representation is created for each word in the input sentence. This is done to increase the accuracy of system. The word2vec representation module represents the words in the input query with a specific numerical value. The deep neural network takes this as input for completing the learning process.

There are two models for performing word2vec representation: CBOW model and Skip-gram model [13]. CBOW is Continuous Bag of Words model which predict a word when a context is given. On the other hand, skip-gram predict the context when an input word is given. For the proposed work skip-gram is less efficient than CBOW. Because CBOW model gives more weights which will be used for training neural networks. The author uses CBOW model in the proposed system. Figure 2 shows the simple CBOW model [14].

4.2 Deep Learning Process:

Deep learning comes under machine learning. It trains to the system using deep neural networks for taking decisions like human. There are various kinds of neural networks like feed-forward NN, recurrent NN, convolutional NN, recursive NN [15]. Each of these neural network have their own pros and cons. The proposed system uses RNN for deep learning process.

RNN is the recurrent neural network. This means it uses the functionality of recurrent neural networks. The architecture in figure 1 shows that after word2vec representation, words related to the user query are obtained using deep learning RNN. This is done via topic modeling.

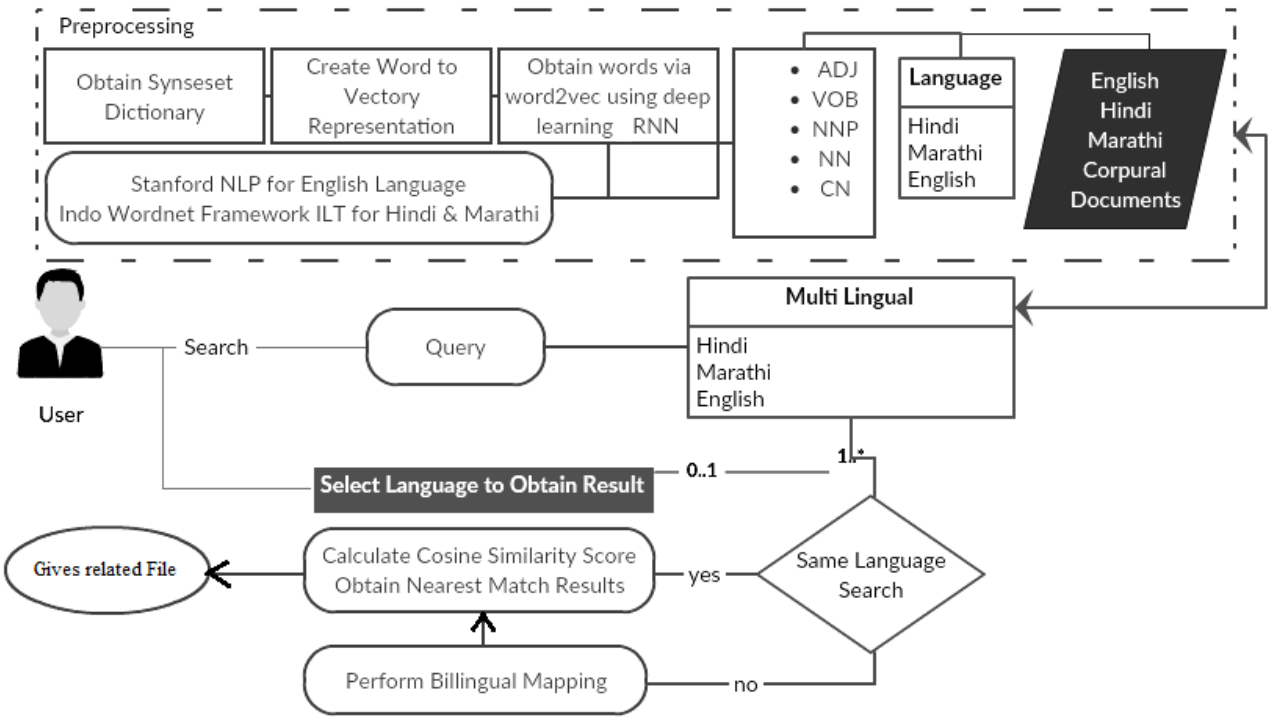

Figure. 1: Architecture of proposed system

Topic modeling comes under text mining. If the words of the input query does not exactly matches with the words in the documents, then the system retrieves result based upon the words obtained through trained RNN model. Topic modeling is done by two ways: First is LDA topic modeling and second is stopword modeling. The deep learning model used in this system is trained with the help of LDA topic modeling. LDA technique focuses on topic words. It finds abstract topics from the set of files. Stopword modeling removes all the stopwords from the sentence and assigns equal weight to each of the topic words.

The third phase bi-lingual mapping is an optional phase. If user desired same language search, then there is no need of any bi-lingual mapping and this phase becomes omitted. On the other hand, if user demands multi-lingual results then it is mandatory to perform bi-lingual mapping. In the proposed system bi-lingual corpus developed by CFILT [16] and wordnet is used for performing bi-lingual mapping. The task of wordnet is to find the synonyms of words in the input query, so that the system performs matching based upon these synonyms to retrieve better results.

In the fourth phase, cosine similarity score is calculated for obtaining nearest matching results. In case of infrequent words cosine similarity is very useful notion. The formula to calculate cosine similarity score is as in equation 1.

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}||\,||\vec{b}||} \qquad \text{……..(1)}$$

Finally, the fifth phase is the output phase. In this phase end user obtain the list of files ranked accordingly from highest matching score to lowest matching score as result.

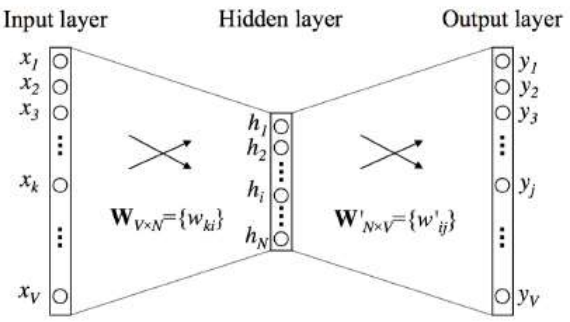

Figure 2: A simple CBOW model

V. ALGORITHM

The algorithm used for extracting topics from the datafile is shown in following figure 3. Input to the algorithm is the path of the directory where all the dataset files are placed. The algorithm returns a text file consisting of all the topics that are used to retrieve the documents of

Algorithm

Input:

The path of directory where the training dataset D is stored

where the directory contains n number of files f1, f2,...., fn.

Output:

A text file T containing set of all topics from D that are used

for searching results.

Procedure:

```
Begin
    File T
    Create word2vec Representation   //Assigns weight to each word all files
    while (i<=n)
        if (getTopic(fi) != stopword)   //Removes stopwords from all files
        T = getTopic(fi)
    Obtain the output T.
End
```

Figure 3: Algorithm used for extracting topics

user's need. In the procedure, word to vector representation of all files is created. Then LDA modeling removes all the stopwords from the files and assigns equal weight to each of the topic word. Finally the file with all the topics is successfully created.

VI. EXPERIMENTS AND RESULTS

We have implemented the proposed system in Java with NetBeans IDE framework. For testing purpose we have used Tourism dataset. This dataset is obtained from CFILTs official website [16].

6.1 Performance Analysis Formulae:

We have done the performance analysis on four parameters- Precision, Recall, F-score and Accuracy. In IR, precision is the part of retrieved cases that are related. Precision is also known as positive predictive value. Recall is the part of the related cases which are retrieved [17]. It is also called sensitivity. Recall as well as precision are based on computing relevance.

Let P → Precision,
R → Recall,
F→ F-score,
and A → Accuracy

Then the formulae of precision, recall, f-score and accuracy are as follows in equation (2) to (5) [18].

1) $P = Tp / (Tp + Fp)$ ...……(2)

2) $R = Tp / (Tp + Fn)$ ...……(3)

3) $Fs = (2*(P * R)) / (P + R)$ ...……(4)

4) $A = (Tp + Tn) / (Tp + Tn + Fp + Fn)$ ……(5)

Here, Tp is true positive, Tn is true negative, Fp is false positive and Fn is false negative.

In simple terms, high accuracy means an algorithm returns meaningfully additional related than unrelated results, whereas a high recall shows that algorithm gives the most relevant results.

6.2 Results Obtained from Proposed System:

We have tested our system for 100 different queries. The queries were asked in English and the average performance of system obtained for each English-to-Hindi and English-to-Marathi mapping is shown in following table 1.

Table 1: Performance of system

	Precision	Recall	F-score	Accuracy
Hindi	79.12%	75.90%	77.48%	63.33%
Marathi	77.79%	74.15%	75.92%	61.22%

6.3 Results in Graphical Form:

Table 1 shows the average values of precision, recall, f-score and accuracy for 100 queries in tabular form. Figure 4 gives the graphical representation of these results.

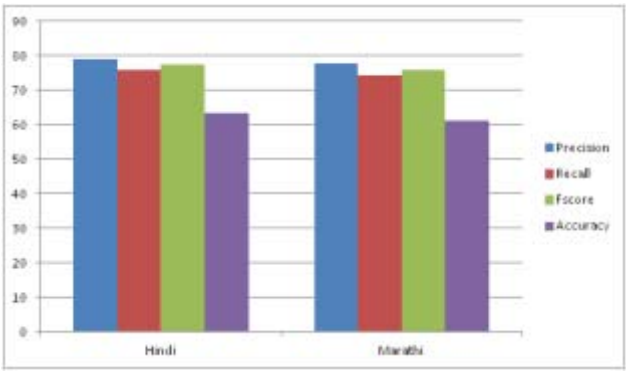

Figure 4: System results in graphical form

6.4 Comparison of Proposed System:

Comparison of the system is done on the basis of whether deep learning (DL) is used or not. The system is tested for English-to-Marathi results. It is observed that multi-lingual searching part of the system which uses deep learning technique gives better results than the bi-lingual searching part which doesn't use deep learning.

Table II: Comparison of system

English-Marathi	Precision	Recall	F-score	Accuracy
Without DL	84%	80%	82.35%	70%
With DL	90.86%	99.44%	94.95%	91.19%

Above table II shows the comparison of proposed system based upon deep learning usage. Figure 5 gives the graphical representation this comparison.

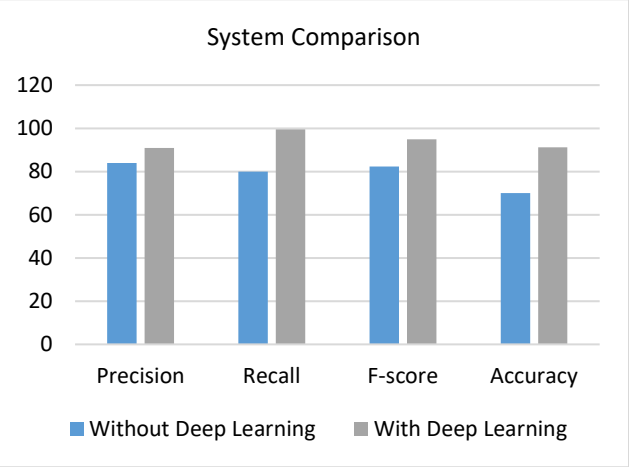

Figure 5: Comparison of System in graphical form

6.5 Limitations of Proposed Work:

There are two limitations of the proposed work. First is limited language support. This work supports only two Indian languages i.e. Hindi and Marathi. So it is of less use for those users who are unaware about Hindi and Marathi. The second limitation is the perfectness of input query. The input query provided by the user must be in correct grammatical structure in order to obtained perfect result.

VII. CONCLUSION

Thus we have implemented the proposed system successfully. It is multi-lingual information retrieval system developed using deep learning technique Multi-lingual searching results are obtained with the accuracy of 91.19 %. This results are obtained using deep learned model and similarity score. Bi-lingual Searching results are obtained with the accuracy of 70 %. This results are obtained by using cosine similarity score only. Hence the approach of using deep learning for information retrieval is better.

This system is very useful to those people whose mother-tongue is not English. Especially Indian people who knows Hindi very well can use this system to retrieve the information they need in Hindi as well as in Marathi language.

In future, the efficiency of this system can be improved by providing index to the tourism dataset. Work can be improved for joint-word queries. Also we can make multi-lingual sentiment analysis based upon the user's search, i.e. whether his/her search is positive or negative. Proposed work can be extend to support more Indian languages like Bengali, Tamil, Panjabi etc.

ACKNOWLEDGMENT

This is to acknowledge and thank all the individuals who played defining role for shaping this work. Without their constant support, guidance and assistance this work would not have been completed. I would like to convey my heartfelt thanks to our Head of Department, Dr. V. P. Kshirsagar for giving me the opportunity to embark upon this topic and for his constant encouragement. Also, I am sincerely thankful to Principal Dr. P. B. Murnal who created a healthy environment for all of us to learn in best possible way.

REFERENCES

[1] Thorsten Brants, "Natural Language Processing in Information Retrieval", Google Inc.

[2] Pothula Sujatha and P. Dhavachelvan, "A review on the cross lingual and multi lingual information retrieval", International Journal of Web and Semantic Technology, Vol.2, No.4, October 2011.

[3] W. Woods, "An Experimental Parsing System for Transition Network Grammars in Natural Language Processing", R. Rustin, Ed, Algorithmic Press, New York, 1973.

[4] W. Woods, R. Kaplan and B.Webber, "The Lunar Sciences Natural Language Information System", Bolt Beranek and Newman Inc., Cambridge, Massachusetts Final Report. B. B. N. Report No 2378, 1972.

[5] G. Hendrix, E. Sacrdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data", iACM Transactions on Database Systems, Vol. 3, No. 2, pp. 105– 147, 1978.

[6] Hendrix, G. (1977). "The LIFER manual-A guide to building practical natural language interfaces". SRI Artificial Intelligence Center, Menlo Park, Calif. Tech. Note 138.

[7] D. Warren and F. Pereira, “An efficient and easily adaptable system for interpreting natural language queries in Computational Linguistics” Vol. 8, pp. 3 – 4, 1982.

[8] T. Amble, “BusTUC - A Natural Language Bus Route Oracle.”, in 6 Applied Natural Language Processing Conference, Seattle, Washington, USA, 2000.

[9] Yunyao Li, Huahai Yang, and H.V. Jagadish, “Constructing a Generic Natural Languaoge Interface for an XML Database”, EDBT (2006).

[10] Dua, M.; Dept. of Comput. Eng., NIT kurukshetra, Kurukshetra, India; Kumar, S.; Virk, Z.S., "Hindi Language Graphical User Interface to Database Management System", IEEE.

[11] A. Faraj EI-Mouadib, S. Zubi Zakaria, A. Ahmed Almagrous and S. Irdess EI-Feghi “Generic Interactive Natural Interface to Databases”, International Journal of Computers issue 3, vol. 3, 2009.

[12] Shervin Minaee, Zhu Liu, “Automatic Question-Answering Using a Deep Similarity Neural Network”, August 5, 2017.

[13] Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anishka Rastogi, Shikha Jain, “Machine Translation using Deep Learning: An Overview”, ICCCE, July 01-02, 2017.

[14] David Meyer, “How exactly does word2vec work?”, July 31, 2016.

[15] Jiajun Zhang and Chengqing Zong, “Deep Neural Networks in Machine Translation: An Overview”, IEEE INTELLIGENT SYSTEMS, *Published by the IEEE Computer Society,* 2015.

[16] CFILT: - Center for Indian Language Technology at IIT, Bombay, “www.cfilt.iitb.ac.in”.

[17] C. Ferri, J. Hernandez-Orallo, R. Modroiu, “An Experimental comparision of performance measure for classification”, Pattern Recognition Letters 30 (2009) 27-38.

[18] Cyril Goutte, Eric Gaussier, “Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation”, IR Research (ECIR’05), LLNCS 3408 (Springer), pp. 345-359.