DOI 10.29042/2018-3863-3867

Helix Vol. 8(5): 3863- 3867

Agent Based Multilingual Information Retrieval System: A Design Approach

[*1]Mangala Madankar, [2]Manoj Chandak, [3]Dr. U. N. Shrawankar
[1]Research Scholar, Department of Information Technology,
[2]Supervisor, Department of Information Technology,
[3]Associate Professor, Department of Information Technology,
G.H.Raisoni College of Engineering, Nagpur
**Email: msmadankar@gmail.com*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
As India is a Multilingual Country and on internet multilingual database content increases day by day. Hence multilingual information retrieval is very curious in a country like India. In this paper, a new research design of agent based multilingual information retrieval system is proposed. Agent based architecture is designed for the system architecture development. Here system will support the queries in Hindi, Marathi and English language. This paper briefly describes the multilingual information retrieval design and development system and the role of various agents in IR system.

Keywords: Multilingual Information Retrieval, Agents, Multilingual Data Processing.

Introduction

The utilization of web surfing is developing now daily, different data necessities of clients in dialects unique from English are expanding quickly. The quickly quickening pattern of globalization of the achievement of e-Governance arrangements and organizations expect information to be put away and recovered in a wide range of regular dialects. The essential information stockroom for such applications, should be proficient as for multilingual information. Productive multilingual information stockpiling and inquiry preparing of information straddling over in excess of one regular dialects are of vital significance in the present globalize world. CLIR manages asking inquiries in a single dialect and interest for recovering archives in another dialect. MLIR tolerating that making inquiry in various dialect and recovering archives in various (more than) the question dialects.

Our framework contain immense measure of information and information can be controlled in the database by a few counterfeit dialects. Numerous Natural Language frontends have been created as a win Work created in the region of artificial intelligence, yet the vast majority of them utilize the English as a characteristic dialect. India being a Multilingual Nation and just 5% of the population brag the training up-to matriculation level and there utilize is constrained. The information retrieval in local languages from storage database has majority of the impact. [1] This work proposed a framework of an agent based architecture that utilizes the various intelligent agents in the multilingual IR system. In the proposed research design multilingual database like resource information in Hindi, Marathi and English language of an education and sport domain will be stored in cloud environment. The user can give his query to the system in his native language with the help of user interface agent. Analyzer agent plays an important role for the multilingual query processing. User can ask the query in any language and information stored in database will be of any language from Hindi, English and Marathi, such as user can ask the query in Marathi and access the information in Hindi. The agents will be written in the programming language and they will communicate via Inter-agent Communication Language (ICL). Various agents will be involved here such as User interface agent, Database agent, Analyzer agent, Server agent, Parser agent and SQL agent.

In the proceeding section, we present the brief literature survey on agent based Information Retrieval system, section III gives the details about the proposed IR system and last section will be the concluding section

Related Work

Different strategies of multilingual information retrieval is describes in [1,2,3, 4,8,10,14,16,17]. The research is done on the various Indian languages are invoked here. The language-independent indexing technology is utilized to process the content accumulations of English, Telugu and Hindi dialects. Here author has utilized multilingual lexicon based word-by-word query interpretation approach. [3,8]. Various methods of cross language information retrieval is elaborated in paper [4,5,6,7,11,12,15,17,18,22,25] here various Indian languages are considered for the CLIR. Dictionary based methods, maximum entropy method are used for the translation of one language to another. Multilingual query processing system using various software agents is elaborated by Suman Mary Idicula in [19] Here system handles the query in Hindi and Malayalam languages. Meaning extraction from the plain query is done by the NLP techniques and retrieved information is given back to the user in its

native language. Author Bin Xue discussed the multi-agent Information retrieval on intelligent evolution. Various agents are designed for the accurate and efficient precision and recall [20]. In [21] author QuJubao, Liu Sheng, Ye Qiusun discussed about the multi-agent based web information retrieval system. In [26] Ahmad Al-Daraiseh, et.al developed an Intelligent Agents framework that use to extract Multi-lingual Web News.

Motivation

The inspiration of this research is that, Multilingual Nation like India with decent variety in different dialects and just 5% of the population gloats the education up-to matriculation level and there utilization is constrained. In planting and agribusiness field if the data is in farmer's local dialect can be available on the web then our development rate consequently increments. Furthermore, thus the data recovery in local dialects from storage database has dominant part of the effect. This MLIR framework can be viably use in application zones like e-administration, instruction, national asset arranging, debacle administration, agribusiness, country well-being, data stands and so on.

Multilingual Information Retrieval Systems

Multilingual Information Retrieval System flow is illustrated in below figure 1. Here three main steps of MLIR are discussed i.e. indexing, translation and matching.

Indexing

This process is at the server side, while making the database or the content from where one can retrieve the relevant information. In indexing process, there are six steps.

Step 1: Preprocessing: In preprocessing remove all duplicates, header and footer etc from the documents, and unify format and coding.

Step 2: Language identification: In language identification step identify the common words, frequencies, bigrams and trigrams. Same dialect may utilize distinctive coding, and same data may be available in different language, so need to identify all of them. It is important to apply the appropriate stop word in language identification step.

Step 3: Granularity: Here we find what the granularity of recovered thing. Like Entire Document, sub-document i.e. sentence, paragraph, chapter, passage, or super document i.e. aggregation of document, linked document folder, or retrieved item is logically related.

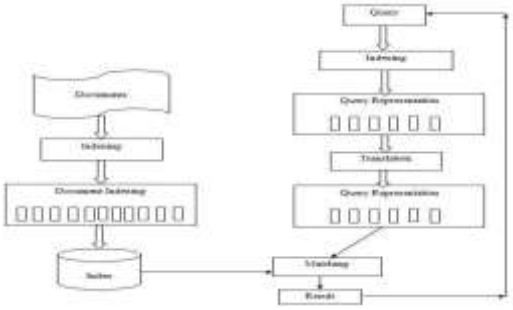

Figure 1: Multilingual Information Retrieval System

Step 4: Segmentation: (tokenization-language independent Approach, n-gram indexing)

Archive is part into substantial tokens. The tokens are appropriate to frame the record structure. Unfortunate tokens are dispensed with. Stage 5: Normalization: Tokens are standardized with a specific end goal to achieve highlights which are appropriate for recovery controlled vocabulary. - Consider " judgment"- vs - "judgement" this will be normalize orthographic variations -Case normalization-(e.g., Moon vs. moon) -Consider e.g., "analyzing" vs "analysis" this will be lexical variants - comparable terms that are synonymous in importance - (e.g., "film", or "movie")

Step 6: Enrichment

- Records are improved with additional highlights, or with more specific highlights
- Named Entity recognition
- Thesauri for expansion
- Anchor text from inlinks
- Contextual data (from user profiles, from linked pages, from clustering, ...)

Translation

Machine Translation is one of the parts of dialect preparing inside Computational Linguistic. The machine-interpretation technique is utilized to decipher either the reports or client asked for inquiry by utilizing a standard machine interpretation framework.

a. Bilingual dictionary: Bilingual lexicons are utilized as a part of a word reference based approach. For interpreting content and word starting with one dialect then onto the next dialect, bilingual lexicon can be utilized.

b. Parallel Corpora: Corpus based interpretation normally gives much upgraded execution, when contrasted with lexicon based approach. The arrangement of parallel corpora is entangled and very costly.

c. Morphological Analyzer: Analyzing morphology of given content is called as Morphological Analyzer, which is a product segment. It faculties or creates morphemes of an info word.

d. Transliteration: If question words not found in the bi-lingual lexicon at that point go for transliteration. For the transliteration, administer based approach can be utilized for the dialect like Devanagari.

e. Word sense disambiguation: Word sense disambiguation (WSD) is depicted as the activity of looking through the feeling of a word in circumstance. WSD is a center issue in numerous assignments identified with language handling.

Matching

The last phase of multilingual data recovery is matching. The matching stage needs to appoint weights to question (and record) terms.

For matching some Assumptions to be follow:

Words having similar vocabulary tend to have the similar significance

1. More questions terms match → more relevant
2. Questions terms more continuous in doc → more relevant
3. Questions terms bunched firmly in doc → more relevant

4 Others (visit inlinks, event in title, and so forth.)

Methods of evaluating IR System

1) *Precision:* Precision is the portion of the records recovered from the database which are significant to the client's data require

P= Tp/(Tp + Fp).

Where Tp=relevant record retrieved, Tp + Fp= relevant retrieved record and Tp=true positive, Tp=false positive.

2) Recall: is the portion of the reports which are significant to the client inquiry that are effectively recovered

R= Tp/(Tp+ Fn)

Where Tp =relevant retrieved record

(Tp+ Fn) = Total number of relevant retrieved record in the database and

Tp = true positive,Fn= false negative

Agent Based Approach

Here research design will be based on the agent based architecture system. In the proposed research design multilingual database like resource information in Hindi, Marathi and English language of education and sport domain will be stored in cloud environment as shown in figure2. The user can give his query to the system in his native language with the help of user interface agent. Analyzer agent plays an important role for the multilingual query processing. User can ask the query in any (English, Hindi or Marathi) language and information stored in database will be of any language from Hindi, English and Marathi. If user can ask the query in Marathi language then intelligent agent find the related information from the database, query translation agent will be work on the same level, and access the information from all the three languages and then again translate the information in users native language i.e. marathi. Theagents will be written in the programming language and they will communicate via Inter-agent Communication Language (ICL)

Different specialists will be included here, for example, User interface operator, Analyzer specialist, Parser specialist, SQL specialist, Database specialist and Server operator. Move of UI operator is to enter the content to the framework. Analyzer operator will tokenize the approaching client question. Significant words changed over into units called tokens and learning substance of tokens will be put away in outlines. Parser operator will check the linguistically rectify sentences and after that parsing will be finished with the assistance of creation rules. Creation tenets will contain the characteristic dialect design as precursor and class as outcome. The yield of the parser is utilized by the SQL operator,

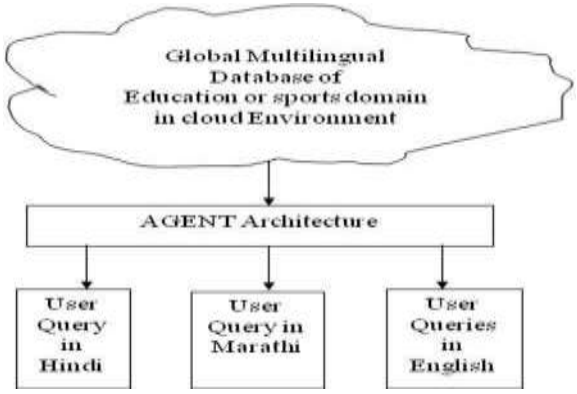

Figure 2. Research Design of Multilingual Information Retrieval System

This specialist created what might as well be called the Natural dialect inquiry entered by the client. Database agent interacts with the Cloud database such as MYSQL contain all the domain information. Server agent will be responsible for communication and control for providing the global data. Figure 3 shows the block diagram of multilingual information retrieval system, where how IR process will accomplish is given.

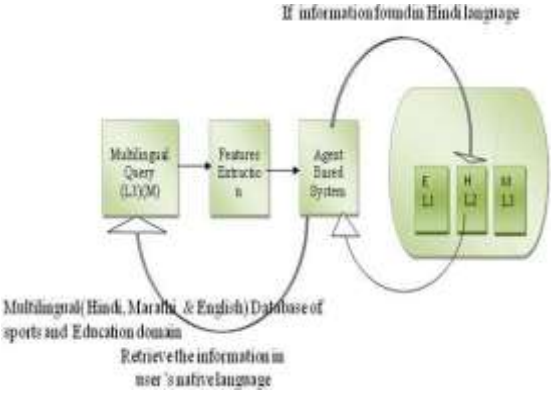

Figure 3. Block Diagram of Multilingual Information Retrieval System

Conclusion

Cross-lingual and Multi-lingual Information Retrieval System gives new pattern to seek records through various assortments of dialects over the world and this can be the new standards for looking among two dialects as well as in numerous. User can enter the query in Marathi language or the Hindi language but in the database information are stored in the English language, and then our agent based system will convert the Hindi or Marathi query into English and find the related information from the database and again convert the document into the user native language like Hindi or Marathi. We are used the precision and recall methods to find the exact relevant information in the users native language. Experimental set up is ready, Cloud design and interface modules are ready; results will be shown in next version of the paper.

References

1. S.M.Chaware, SrikanthaRao. "Information Retrieval in Multilingual Environment," Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09, IEEE Computer Society, pp 648-652, (2009).
2. N. Swapna1 , N.Hareen kumar2, B. Padmaja Rani3, "Information Retrieval In Indian Languages: A Case Study On Cross-Lingual And Multi-Lingual", International Journal of Research in Computer and Communication technology, IJRCCT, ISSN 2278-5841, Vol 1, Issue 4, (September 2012 Raju Korra#1, PothulaSujatha*2, SidigeChetana*3, MadarapuNaresh Kumar, "Performance Evaluation of Multilingual Information Retrieval (MLIR) System over Information Retrieval (IR) System" IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 MIT, Anna University, Chennai. Pp-722-727, (June 2011)
3. Sung J. Shim "Using Cross-Language Information Retrieval Methods for Bilingual Search of the Web", IEEE Computer Society, International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce. (2005)
4. AnuragSeetha, Sujoy Das, M. Kumar, "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method", 10th International Conference on Information Technology, IEEE Computer Society, pp- 56-61, (2007)
5. Yue-Jie Zhang1, Tao Zhang2, "Research On English-Chinese Cross-Language Information Retrieval" Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007, IEEE Computer Society, pp- 3448-3453, (2007)
6. Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef, MalekBoualem, "Query Expansion for Cross Language Information Retrieval Improvement" 978-1-4244-4840-1/10/ IEEE-(2010).
7. Yilu Zhou, Jialun Qin, Hsinchun Chen, Jay F. Nunamaker , "Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal" Proceedings of the 38th Hawaii International Conference on System Sciences –pp-1-10, (2005)
8. Kumar Sourabh, VibhakarMansotra, "Query Optimization A Solution for Low Recall Problem in Hindi Language Information Retrieval", International Journal of Computer Applications (0975 – 8887) Volume 55– No.17, -pp-6-17 (October 2012).
9. k. Ganesan and G. Siva, "Multilingual Querying and information processing", Information Technology Journal ISSN-1812-5638, PP 751-755, (2007).
10. Sriram S. Parthatalukdar, Sameer Badskar, "Phonetic distance based cross lingual search",
11. Qin Chen1,Lei Liu2,Lin Ma "Application of Maximum Entropy Method in Chinese-English Cross Language Information Retrieval", pp-1192-1195, IEEE (2008).
12. JolantaMizera-Pietraszko, "Interactive Document Retrieval from Multilingual Digital Repositories", IEEE, pp- 423-428, (2009).
13. NikolaosAmpazis, Helen Iakovaki , "Cross-Language Information Retrieval using Latent Semantic Indexing and Self-organizing Maps", IEEE, pp-751-755,(2004).
14. Mohammad Shamsul Arefin*, Yasu, "Multilingual Content Management in Web Environment", Chittagong University of Engineering & Technology, Bangladesh, IEEE (2011).
15. B.Ashwin Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)", Second International Conference on Advanced Computing & Communication Technologies, IEEE, Computer Society, pp-64-68,(2012).
16. Bao-Quoc Ho, Van B. Dang, Minh V. Luong and Thuy T.B. Dong, "English-Vietnamese Cross-Language Information", IEEE, pp-107-113,(2008).
17. ZhaoRongying, "Visual analysis on the research of cross-language information retrieval", IEEE, pp-107-113,(2008).
18. Sumam Mary Idicula, David Peter, S "A Multilingual Query Processing System using Software Agents", Journal of Digital Information Management _ Volume Number 6, pp-385-390, (December 2007)

19. Bin Xue, "Research on Multi-agents Information Retrieval System Based on Intelligent Evolution", 2nd International Conference on Computer Science and Network Technology, pp-1042-1045
20. QuJubao, Liu Sheng, Ye Qiusun, "The Design of Web Information Automatic Retrieval System Based on Multi-Agent", International Conference on Multimedia Information Networking and Security, 2009
21. Manoj Kumar Chinnakotla, SagarRanadive, Pushpak Bhattacharyya and Om P. Damani, "Hindi and Marathi to English Cross Language Information Retrieval"at CLEF (2007).
22. Tan Xu1 and Douglas W. Oard, "Maryland: English-Hindi CLIR" FIRE-(2008) Technology Dr.M.Hanumathappa1,Mallamma V. Translation Tool for Natural Language Processing", International Journal of Science and Applied Information , Volume 1, No.4, , ISSN No. 2278-3083- pp-107-112, September – October (2012).
23. Mallamma V Reddy, Dr. M. Hanumanthappa, "Kannada and Telugu Native Languages to English Cross Language Information Retrieval" Mallamma V Reddy et al, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5), pp-1876-1880,(2011).
24. Ahmad Al-Daraiseh, Wafa' Haddoush, "Developing a Framework that Utilizes Intelligent Agents to Extract Multi-lingual Web News," IEEE,(2015).
25. Pinaki Bhaskar, Amitava Das, Partha Pakray and Sivaji Bandyopadhyay "Theme based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010", Forum for Information Retrieval Evaluation (FIRE) (2010).
26. Ashish Almeida, Pushpak Bhattacharyya, "Using Morphology to Improve Marathi Monolingual Information Retrieval" *FIRE-(2008).*

