International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 7-12
e-ISSN:2278-621X

KNOWLEDGE BASE INTELLIGENT QUERY PROCESSING

Raji Sukumar A[1], Dr. Babu Anto P[2]

Abstract- In the field of Artificial Intelligence (AI) there exists a long-standing goal for an expert system. Intelligent Query Processing is concerned with in the context of automatically processing NL text query with several Knowledge Base (KB) technique of AI-literature to develop an expert system. An intelligent system requires understanding of related terms like wisdom, knowledge, reasoning, thought, cognition and language learning. Intelligence can be defined as the ability to acquire, understand, and apply knowledge or ability to use thought and reason. This paper propose a clear architecture to automatesexpert systems as a knowledge base engine through natural language interface.

Keywords –AI, KB, NLP, IR, NLIDB.

I. INTRODUCTION

Knowledge extraction from an NL text is possible by transforming into a representation which is manipulated by computers. Natural language processing (NLP), researchers traditionally attempted to build these knowledge manually. The knowledge acquisition is a bottleneck and particularly difficult in the case of NLP, where the amount of information to encode is comparatively large and it is difficult to make it complete and correct. It is estimated that every year about 160 terabytes of NL text type information is produced. Although NLP is difficult, its potential benefits have caused researchers to investigate the depth of both syntactic and semantic processing.The general NLP techniques are discussed with number of experiments based on NLP applications in Malayalam language. This paper proposes a Intelligent Query Processing Architecture of Malayalam language through various levels of NLP. Query processing is evaluated on two NLP applications. The KB techniques Ontology and a novel approach Knowledge Dictionary is applied in two NLP application areas, Information Retrieval (IR) and Natural Language Interface to Database (NLIDB).

II. BACKGROUND STUDY

Like most AI systems, NLP, requires substantial amount of knowledge that is difficult to acquire. A measure of semantic similarity is presented in taxonomy based on the notion of shared information context. Systems have been developed, mainly for languagessuch as English; some examples of these systems are California Restaurant Query, ExpediaHotels, GeoQuery[1], JobQuery[2], SQ-HAL [3], andSystemX[4].NLP can play a role in both the retrieval and storage of documents which can be used to build a friendly user interface that allows free language query submission and hence eliminates the need for mastering a formal query format.The different development stages of NLP, emphasis on MT by the influence of AI. This study briefly reviews some of these techniques. Between these technologies we should mention the Semantic Technologies as part of AI, Rule-based systems, logic-based inference and decision support systems [5].

[1] *Department of Information technology Kannur University, Kannur, Kerala, India*
[2] *Department of Information technology Kannur University, Kannur, Kerala, India*

III. INTELLIGENCE QUERY PROCESSING

Developing programs that simulate human intelligence is the most difficult task faced by AI researchers. Human mind react with situations from its ability to reason symbolically. Intelligence is the capability to acquire, understand and apply knowledge i.e., aptitude to exercise thoughts and reason. Knowledge is the information registered in human mind by learning. Knowledge is stored as symbols in human mind; these symbols are manipulated by reasoning. Knowledge representation is a central topic in AI, to develop a system that can understand human knowledge is also challenges faced by AI researchers. Language is the communication medium to express human thought, especially complex and require conscious inference. Research on knowledge representation propels the information age from the elementary stage, mainly with data processing, to the high level stage with knowledge processing. Although now there are many methods for knowledge representation like production rules, semantic network, logic, frame and script. To explore new methods for knowledge representation is still one of the important subjects in AI. NLP spans to three sub areas: NLIDB, IR system and KB system which gives a clear explanation to the intelligent QP. Query will be processed with the intelligent technique of KB system with various NLP applications.

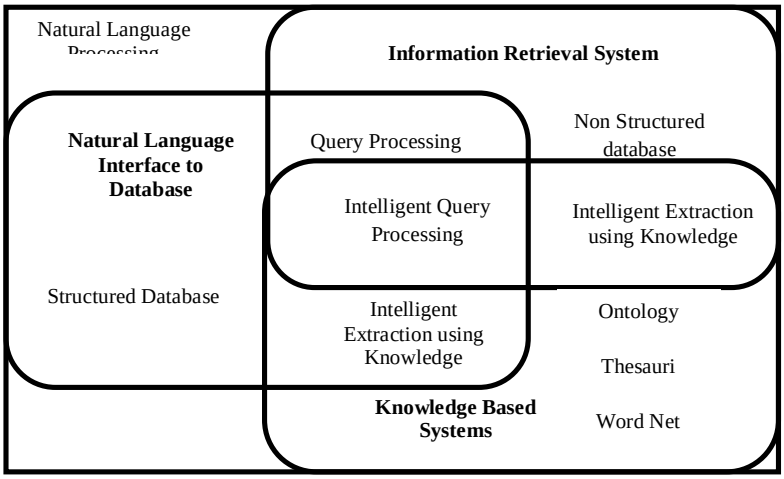

Figure 1: Intelligent Query Processing Under NLP

III. PROPOSED SYSTEM

A. Query Processing Based on NLP–

User's information request will be transferred to a query which will be processed by a QP system or any retrieval system, which has a retrieval unit like a file, a document, a web page, a paragraph or some other structural unit which contain an answer to the search query. The different kinds of queries normally posed to text retrieval system are shown in the Figure 2. The types of query depend on the QP system based on the users experience with different query languages. If the user knows exactly what he wants the retrieval task is easier. The different techniques meant to improve the effectiveness of the queries are also discussed. The query text contains different words, the expansion of a word to the set of its synonyms or the use of a thesaurus and stemming to put together all the derivatives to the same word.Some words which are very frequent and do not carry meaning (such as 'the', 'an'), called stop words are removed. Malayalam language has stop words like (*vendi, koodi, ninnu, allenkil, porenkil*). The ultimate goal of this system is to design and implement a NL query processor named Natural Language Query Processing System (NLQPS), for querying a search engine.

1) Query Languages

A query will have the capability to retrieve the information behind the document either 'subject matter' or the 'content' of some text. Communication is taking place through NLs. This work concentrates on text query processing and system architecture with various categories of query languages is suggested.Retrieving information usingsemantic matching of words is a cognitive activity that callsfor operation of accessing information from memory. Most human knowledge is coded in NL, which is difficult to use as knowledge representation language for computer systems. The type of query a user might formulate depends on the retrieval model. Based on this principle different query languages can be formed.The main query languages are:

a) Keyword-based querying

The simplest form of a query is composed of keywords; the document containing such keywords is searched for. Keyword based queries include simple words and phrases as well as Boolean operators which manipulate sets of documents. They are popular because they are intuitive, easy to express, and allow fast ranking and a keyword query iscategorized as:
i) *Single word*: The most elementary query that can be formulated in a text retrieval system is a word. The word model can be further divided into letter model, which can be used in pattern based queries. Since a word carry a lot of meaning in natural language, many models are completely structured on concept of words. The result of word queries are set of documents containing at least one of the words of the query.

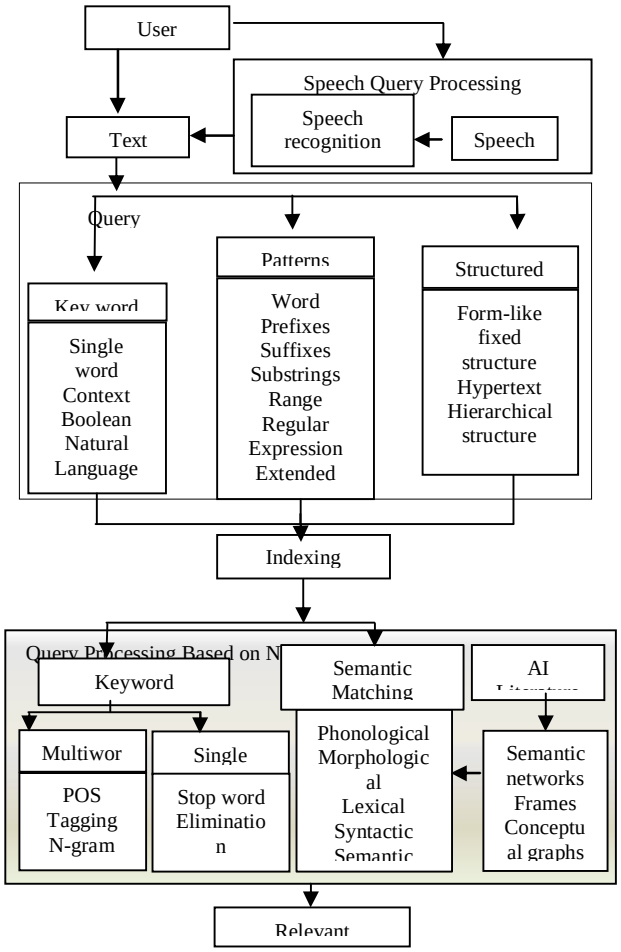

Figure 2: Query types in different query languages

ii) *Context*: The Context based queries complement single-word queries with the ability to search a word in a given context. Words which appear near each other may signal a higher likelihood of relevance than if they appear apart. The query can be phrases of words or words which are proximal in the text.
a) *Phrases*: Phrases are sequence of single word queries in a given context.
b) *Proximity*: A relaxed version of the phrase query is the proximity query. Here a sequence of single words or phrases is given, together with a maximum allowed distance between them.
iii) *Boolean queries*: The combining keyword queries are Boolean queries. A Boolean query has a syntax composed of atoms that retrieve documents. The Boolean operators work on their operands and deliver sets of documents. The most commonly used operators are OR, AND, BUT etc.
iv)*Natural Language Query*: A key word query can be processed as a natural Language based query with its linguistic and semantic level. A word can be recognized with its linguistic level with its morphological variations and corresponding semantic variations.

b) Pattern based querying

Pattern matching which includes more complex queries are generally aimed at complementing keyword searching with more powerful data retrieval capabilities. The query formulations are based on the concept of a pattern, which allow retrieval of pieces of text that have some property. The data retrieval queries are useful for linguistics, text statistics and data extraction. A pattern is a set of syntactic features that must occur in a text segment. The most used types of pattern are:

i) Words : A sequence of characters which is the most basic pattern in a text.
ii) Prefixes : A string which must form the beginning of a text word.
iii) Suffixes : A string which must form the termination of a text word.
iv) Substring: A string which can appear within a text word.
v) Ranges: A pair of strings which matches any word between lexicographical orders.
vi) Errors threshold: The search pattern retrieves all text which are 'similar' to the given word. The errors occur due to spelling mistake, and finding the most similar word with the concept of similarity with Lavenshtein distance or edit distance which is the minimum number of character insertion, deletions and replacements needed to make them equal.
vii) Regular Expression : A Regular Expression is rather a general pattern built up by simple strings, which are meant to be matched as substrings.
viii) Extended Patterns : An Extended Patterns are subsets of Regular Expression which are expressed with simple syntax.

IV.EXPERIMENTS

In this work there is a MLTTES-Kerala corpus for the possible words in the railway time enquiry system. The user query from MLTTES-Kerala corpus will be processed in various NLP levels. This it requires the linguistic information about the reflected form of the words. The intelligent technique ontology enables the system an intelligent query interface. Query building is together with a query language involves various stages. A precise and unambiguous query expression is generated by means of a diagrammatic interface. Ontology has its own lexicon set and which can be extracted with corresponding knowledge tag. This is possible for each domain of lexicons. This is implemented with the help of knowledge tagging. With the concept map technique, using the knowledge tagging, a domain specific ontology can be designed. The ontology learning process involves two basic tasks- domain specific concept identification and construction of concept hierarchy by establishing the knowledge tagging. Identifying top level concepts and creating a good concept hierarchy are the major challenges involved in the ontology learning tasks. Time ontology is designed with the time domain of Malayalam vocabulary.

Table -1 Experiment Result

Table 1: Performance According to knowledge Frame Type in NLIDB

Query Frame	No. of queries	Retrieved documents	TP	FP	FN	TN	Precision	Recall	F-Measure
ഇനി	5	4	4	0	1	0	100	80	88.9
രാവിലെ	10	8	7	1	2	0	87.5	77.8	82.4
അതിരാവിലെ	16	15	10	2	1	0	83.3	90.9	86.9
ഉച്ചക്ക്	17	16	10	4	3	0	71.4	76.9	74
ൈവകുന്നേരം	58	57	52	2	2	2	96.3	96.3	96.3
സന്ധ്യക്ക്	38	35	34	2	1	1	94.4	97.1	95.7
രാത്രി	28	27	20	3	2	3	87	90.9	88.9
പകല്	14	14	12	1	0	1	92.3	100	96
ഇന്ന്	7	6	5	1	1	0	83.3	83.3	83.3
ഇന്ന്രാവിലെ	7	5	4	1	1	1	80	80	80
Total	200	187	158	17	14	8	87.6	87.3	87.2

Referring to Table 1 it is clear that for 200 questions asked, 187 answers were retrieved out of which 158 answers were relevant to the query and 29 were non-relevant. Even though 13 other relevant answers existed in the corpus they were not extracted because some of the query patterns were not recognized properly or semantics was notsufficient enough to identify the required entity. In this work, retrieval scheme is purely based on named entities. Hence all the sentences with the candidateentities are retrieved. But all theseentities might not be the answers to the question

Questions :200
Relevant Answer:158
Retrieved Answers:187
Recall :87.3
Precision :87.6
F-Measure :87.2

The overall performance of MLTTES-Kerala isgiven in Table 1. Precision of 87.6% shows that the answers retrieved are correct answers and only very few non-relevant answers are retrieved. Percentage of recall is less than precision and is 87.3%.

An IR system is designed in MLIR-Veg, which searches for Malayalam documents under vegetable domain with NL query. This system retrieves documents that are relevant to user queries with the query-document communication using VSM retrieval strategy. NLP techniques have found a useful application in IR which means not only document retrieval, but also includes semantic retrieval by applying NLP techniques. A novel retrieval utility KnowNet is proposed to form the improved query with linguistic and semantic knowledge.MLIR-Veg conducted test using series of 4 query sets. Where first three are hypernym variations and the last one is an improved query containing all hypernyms which was present in other queries.

V. CONCLUSION

This work concentrates the problem of processing a Natural language text Query intelligently based on NLP. Query processing problem is analyzed and identified different query languages through which a normal query is posed by a user. A clear architecture of query processing problem is proposed and

identified the various intelligent techniques from AI-literature like Conceptual Graph, Ontology and Frame to process the queries. It is observed that Natural language is the most intelligent way of Knowledge representation. A detailed study is conducted on Malayalam language to extract knowledge syntactically and semantically. Two NLP applications IR and NLIDB are analyzed with Query Processing problem. The overall performance of MLTTES-Kerala is given in Table 1. Precision of 87.6% shows that the answers retrieved are correct answers. In the case of MLIR-Veg, an average F-measure performance for these individual 4 type of queries. Compare all methods of Information Retrieval performance improved for the improved query with all hypernyms with 87% accuracy.

REFERENCES

[1] Zelle J M, R J Mooney. "Learning to parse database queries using inductive logic programming. In: Proc." Thirteenth National Conference on Aritificial Intelligence, Portland (1996): 1050–1055.

[2] Thompson C A, R J Mooney, L R Tang. "Learning to parse natural language database queries into logical form." In: Proc. ML-97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition (1997).

[3] Flores V, J M, J M Matadamas H. "Computational Linguistics Laboratory: Project Sylvia-NQL." In: Proc. 7th. International Congress on Computer Science Research, Technological Institute of Cd. Madero, Tampico, Mexico (2000): 73–81.

[4] Cercone N. "Human-Computer Interfaces: DBLEARN and SystemX." International Workshop On Rough Sets And Knowledge Discovery (RSKD-93) (1993): 27-28.

[5] Locke W N and Booth A D . "Machine Translation of Languages." John Wiley (1955).