The world as we know it is our interpretation of the observable facts in the light of theories that we ourselves invent.
- Immanuel Kant

Abstract

The availability of electronic document is increasing every day. Explosive growth of electronic data increased the need of intelligent techniques for data retrieval that allows user to distil the most valuable information from mountains of accumulated data. This thesis investigates several research problems arising in Natural Language Processing (NLP) for the purpose of extracting intelligence from a Natural Language (NL) text query. NLP is among the most heavily researched areas of Artificial Intelligence (AI), which has taken a noticeable boom in recent years. The investigator had made a sound understanding of NLP techniques with its application areas like Information Retrieval (IR) and Natural Language Interface to Databases (NIDB) Systems. Automated Intelligent Query Processing System (AIQPS) accepts text queries in NL and processes the query. This is possible by analyzing the query through various levels of NLP. Extracting the knowledge from NL text is one of the challenges in NLP.

Query processing problem addressed here is in Malayalam a south Indian language with rich of linguistic features. India is blessed with diversity in NLs. This work concentrates in the study of Malayalam text query understanding by applying NLP techniques. Language is the tool with which one can express a greater part of human ideas and emotions. The intelligence behind NL is explored by understanding the representation and inference of language in human mind. The investigator had done a detailed study on Kerala Panineeyam grammar and identified that, NLP of any language requires full-fledged Lexical Resources (LR) and Linguistic Tools (LT). This work proposes models for different LRs and LTs for Malayalam Language and a Malayalam Query processing system is developed under restricted domain. The complete NLP works under this thesis is presented as ‘Bhashathilakam’ oru dheebodhi thadyam means a truth of Intellectual knowledge.
Abstract

Information Retrieval (IR) and Natural Language Interfaces to Database (NLIDB) are two application areas of NLP. Both the systems can accept a user query to structured database and non-structured documents. To develop an intelligent system which understands human language query is the main objective of this thesis. The text query in Malayalam is accepted by the system and processes with different NLP techniques under a restricted domain. The NL Query Processing System (NLQPS) developed in this work permits the semantically level understanding of the query by using background knowledge. NLQPS is a rule based system which will enable text understanding with various knowledge relations like polysemy, homonymy and synonymy. Automation of any NL is possible by processing and understanding of a NL text, by having various Linguistic Tools, Lexical Resources and Knowledge Base systems.

The NLQPS on a non-structured database comes under the area IR based on various retrieval strategies. An IR system understands the user query and retrieves ranked relevant documents. A Malayalam IR system is developed as MLIR-Veg, which is tested on the restricted domain of documents related vegetables. MLIR-Veg follows Vector Space Model (VSM) as IR strategies, which assigns a measure of similarity coefficient. The experiments with different similarity measures are conducted, which includes: Inner product (dot product), Cosine, Dice, Jacard and N-Jacard over a Malayalam corpus. The results are discussed with IR evaluation models. It is identified that N-Jacard gives better performance with Malayalam Language corpus. Different IR utilities can be applicable over an IR strategy to improve the result obtained by submitting an improved query based on the Malayalam language dependent retrieval utilities like thesauri and semantic network to provide Knowledge base relations.

The NLQPS on a structured database is the relational model, NLIDB which is an application area of NLP. The NLIDB system developed for railway time enquiry in Kerala as MLTTES-Kerala which accepts NL queries. Architecture is proposed for the system which accepts Malayalam queries and meaning extraction using NLP techniques. This is designed in a structured way such that it can be applicable for other regional languages. A deeper grammar analysis is performed for the construction of parsed syntax tree with the ontology based lookup in order to extract knowledge from a NL query. The efficiency is evaluated by analyzing user queries submitted within the railway time enquiry corpus of distinct words.
Abstract

This corpus contain all station names in Kerala, words related to time, day, direction and question words to make a query.

A full-fledged LRs and LTs are not available in Malayalam. The investigator had done a thorough examination over Malayalam Grammar, proposes language models for various levels of NLP and models for different LTs and LRs required for Malayalam NLP. This thesis focuses mainly on the knowledge required to process NL in written form with its various levels, and makes a support to the theory of knowledge representation by means of graphs comes under the area of semantic networks. The results obtained from both IR and NLIDB shows that the performance of a query processing system can be enhanced by applying different intelligent techniques over a NL query. The intelligent techniques are various promising solutions of knowledge representation like Knowledge Graph (KG) Conceptual Graph (CG) and Ontology. Based on KG principle every word should have a word graph and sentence should be represented by a sentence graph. The level of semantic details to be captured within CG and the type of relations used to relate concepts depend on the nature of the application and the domain is represented by Ontology. Having an information on all linguistic and knowledge relations of the vocabulary system. This thesis proposes a solution to the Knowledge extraction problem is by developing a knowledge dictionary KnowNet (Bhashamakudam) is a relational model of words, which includes the function of a Knowledge Graph, conceptual network, Ontology, linguistic dictionary, WordNet and Thesaurus.