CHAPTER 6

Conclusion & Future Work

6.1 Work Contribution

6.2 Future Work

6.3 Concluding Remark
6.1 WORK CONTRIBUTION

The degree of correct interpretation and processing of Natural Language Query relies on the existence of exhaustive domain-specific lexicon. Despite of this limitation, the proposed N-ELIDB system provides a general framework to handle natural language queries pertaining to non-overlapping domains.

The major benefits of the system are:

- User can query in their (English) language without spending time to learn formal database language such as SQL, that is, it provides means of accessing information in the database independent of its structure and encoding knowledge
- Easy to understand and maintain data for naïve users
- Flexible and generalized framework for all domains
- It provides rich metadata
- Allows relaxation in grammatical rules while entering query
- It supports spell checker
- It supports context resolution and word sense disambiguation
- It provides multiple data format
- Knowledge can be preserved for reusability
Some of the observations noted during the implementation of various methodologies used in present work are as follows:

- Morphology method helps to match words having same meaning but different morphological form
- Stopwords removal expedite the process
- Word based n-gram processing helps in identifying multi-word token
- Fault tolerance method makes system more user friendly
- Domain identification method is useful to map query to appropriate domain
- Knowledge base allows to reuse previously used queries, thereby saving user’s effort
- The well designed heuristic rules for tokenization helps in identifying the token as person, place, date, attribute token etc.
- Multi-Liaison algorithm assists to extract the subject, object and relationship
- WordNet helps in deriving semantic similarity between words
- Answer Class Identification helps in generating more appropriate SQL for a given query
- Data Conversion algorithm supports flexibility of using multiple data formats.
- Although complete word sense disambiguation is difficult to achieve, it is partially supported.
Some of the observations related to different classes of Query are as under:

- Single condition query gives high accuracy result
- Composite conditions query supports reasonably large number of conditions
- Group query and Order by query support keyword based approach
- DDL and DML queries are supported by using template approach
- Temporal queries are handled effectively
- The paraphrases designed for each query category are found to be effective.

The accuracy of the system is measured in terms of precision percentage with three classes that identifies query response as: Correct Queries, Incorrect Queries and Partially Correct Queries. The results found are encouraging and the overall efficiency of system is observed to be more than 70%.
6.2 FUTURE WORK

The following areas of NLIDB can be considered for future work:

- To accept queries in vernacular languages. Any natural language such as English can have several dialects/lingos whose word does not match with standard dictionary words. To include support for these words further processing is required along with appropriate external resources.

- To include more fuzzy questions using fuzzy terms like bad, very bad, not good, not very bad, not very good, etc. Here the fuzzy terms needs to be scale. For example, 1-10 measurement (1-very bad and 10 – very good).

- To support multimedia data such as image, sound and graphics can be attempted.

- To include question based on prediction in case of Information Retrieval system. For example, user can ask question like: “when the student puja will complete the final year of her studies?”, “what will happen if a student fail”, etc.

- To include anaphora resolution concept. The Anaphora denotes the act of referring (in broad sense). That is, an expression refers to another contextual entity. For example, in sentence “Akash helped Ram, he was happy”, here, the tem “he” refers to the “Akash”. It is a challenge to process anaphora (pronoun) since reference identification may be difficult in certain sequence of text. It requires understanding of how discourse is constructed.
• To include computational phonology and text-to-speech. The phonology refers to the study of sound pertaining to the system of language. The phonology component may allow user to interact with database using speech rather than through text input.

• To include discourse structure. The system can include discourse analysis component to deal with discourse structure defined in terms of coherent sequences of sentences, propositions and speech. It can analyze the 'naturally occurring' language use, and not invented examples [46]. It can also reveal socio-psychological characteristics of user.
6.3 CONCLUDING REMARK

The thesis has been proposed a Natural-English Language Interface to Database. It successfully demonstrated the integration of various techniques that can enhance the accuracy of the (Natural to Formal language) translation process. It also tries to improve user interface so that user can enter the query in free form without any restriction of template or grammar. The algorithm can be applied to different domains provided that it is supported with appropriate external resources.

The basic objective of the study is to improve human computer interface while accessing the structured data from the database by using various NLP techniques.

However, if the user enters very few words in query and if those words are difficult to analyze for words sense disambiguation due to overlapping domain, then the system either may not be able to handle such queries or it may produce partial output.

Though, this work demonstrates how different NLP techniques and methodologies can be used in a specific domain using domain knowledge and other external resources. It also shows a method to assist user in database interaction by suggesting/providing the previously used queries from the knowledge base.