International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 9 No. 2 Sep. 2014, pp. 297-306

http://www.ijisr.issr-journals.org/

A Comprehensive Study on Natural Language Processing and Natural Language Interface to Databases

Umair Shafique and Haseeb Qaiser

Department of Information Technology,
University of Gujrat,
Gujrat, Pakistan

ABSTRACT: It was highly desirable for a machine to interact more friendly with the users so that the field of Natural Language Processing (NLP) emerged and Natural Language Interface to Databases (NLIDBs) systems are built and design. A major problem faced by the users of the data bases is that the databases generally make use of special purpose languages familiar only to the trained users like Structured Query Language (SQL). Natural Language Interface to Databases provides the interface in which queries are written in the form Natural Language. These queries are passed through the machine, machine translates these queries. There are different levels of it, after passing these levels machine produce relevant results. This paper will provides comprehensive understanding about Natural Language Processing and Natural Language Interface to Databases.

KEYWORDS: Machine Translation, Semantic, Pattern, Query, Lexical, Linguistic.

1 INTRODUCTION

Natural language interfaces to databases (NLIDBs) is the branch of Natural Language Processing (NLP) and the Natural Language processing is becoming most active technique in Human-Computer Interaction [1] and it is also be the branch of Artificial Intelligence [2]. The purpose of NLP research is to create such environment in which database usage does not require any programming skills and need little or no prior training.

In comparison of Natural Language Interfaces other user interfaces are less natural to interact. In their databases management systems quires are written in the form of complex language like SQL, SPARQL etc which is difficult for the causal and non-technical users that limits the access to the databases whereas a Natural Language Interfaces to Databases (NLIDBs) helps the users to enhance their performance by providing them the access to the information stored in a database naturally and conveniently. The users can perform flexible and easy queries that are expressed in some natural language and get the answer in the same language (e.g. English). For this purpose there are different techniques that are used e.g. semantic grammar that interleaves semantic and syntactic processing, pattern matching and syntactic grammar.

The purpose and the aim of this paper is to provide a comprehensive knowledge about NLP and NLIDBs, different levels of NLP and approaches and techniques which are used, NLIDBs history, development and future.

2 NATURAL LANGUAGE PROCESSING

Natural Language Processing is the area of research and application of different computational techniques for processing, understanding and manipulating the input and output to the system in human like natural language and this is also the main goal of it. It is the subset of both linguistics and computer science. Mainly Natural Language Processing is the branch of Artificial Intelligence (AI) that used for different domain like machine translation, information retrieval and expert systems etc.

Corresponding Author: Umair Shafique

2.1 LEVELS IN NATURAL LANGUAGE PROCESSING (NLP)

There are different levels in natural language processing and all these levels are depends upon each other.

2.1.1 MORPHOLOGICAL ANALYSIS

In NLP, the level of Morphological Analysis [3] is related to the internal structure of the words and a word is made up of small unit called morphemes e.g. “editing” is a word which is made up of two morphemes “edit” and “ing”.

2.1.2 LEXICAL ANALYSIS

In NLP, Lexical Analysis [4] is the process of converting string of word into tokens and tokens are the meaningful character string. A Lexical grammar is also associated with this which includes set of rules to define lexical syntax.

2.1.3 SYNTACTIC ANALYSIS

In NLP, Syntactic Analysis or parsing [5] is the analysis of the words in a string that how they are related to each other according to the rules of formal grammar.

2.1.4 SEMANTIC ANALYSIS

In NLP, Semantic Analysis determines possible meanings of the words in a string as this focus on the meanings of the words in a string that is way this analysis is related with syntactic analysis.

2.1.5 PRAGMATIC ANALYSIS

In NLP, the level of Pragmatic Analysis [6] is about the use of the sentences in different contexts and how that context affects the meanings of the sentences for this purpose this level may require vast amount of world knowledge.

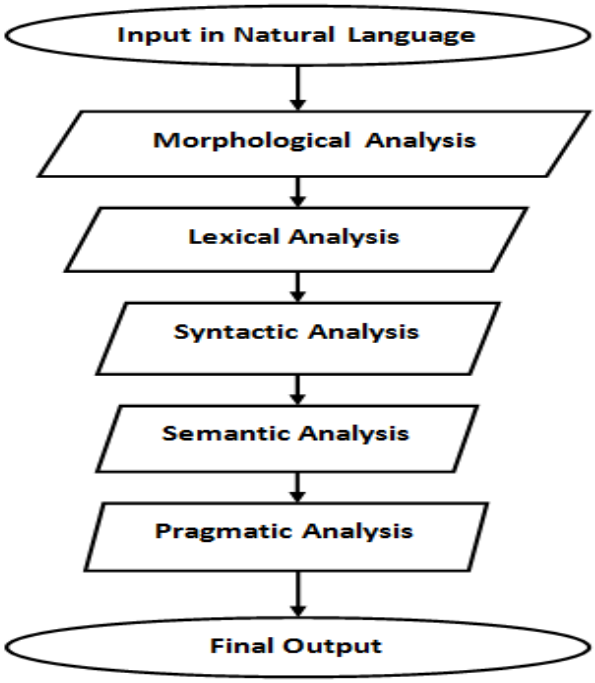

Fig. 1. Levels in Natural Language Processing

3 NATURAL LANGUAGE INTERFACE TO DATABASES (NLIDBS)

Natural language interfaces to databases (NLIDBs) are systems that translate a natural language question into a database query. Their major focus is to provide effortless, effective, user friendly and robust communication to the users regardless of their experience and expertise. Different NLP approaches are used to develop NLIDBs.

3.1 NATURAL LANGUAGE PROCESSING APPROACHES USED TO DEVELOP NLIDBS

Different approaches are used to develop NLIDBs these approaches are written below.

3.1.1 SYMBOLIC APPROACH (RULE BASED APPROACH)

For several decades Natural Language Processing research has been dominated by this symbolic approach [7] because in this approach the words are the symbols that stand for objects and concepts in the real worlds they are put together into sentences that follow well specified grammar rules. Language is analyzed at various levels of this approach to obtain information and certain rules are applied to achieve linguistic functionality on this obtained information. It is supported well as human language capable to include rule-base reasoning. Rules are formed for every level of linguistic analysis and it tries to capture the meaning of the language based on these rules.

3.1.2 EMPIRICAL APPROACH (CORPUS BASED APPROACH)

Empirical approaches [8] are based on statistical analysis and other data driven analysis of raw data which is in the form of text corpora and a corpus is collections of machine readable text. A number of techniques have been emerged to enable the analysis of corpus data. Recent researches in computational linguistics indicates that empirical or corpus based approach is currently the most promising approach to developing robust and efficient natural language processing (NLP) systems.

3.1.3 CONNECTIONIST APPROACH (USING NEURAL NETWORK)

In the recent years, the field of connectionist or neural network [9] processing has seen a remarkable development. There has been significant research applying neural network methods to language processing. As human language has capabilities that are based on neural network in the brain, the artificial neural networks or connectionist network provides on essential starting point for modeling language processing. This approach is based on distributed representations instead of the symbols which correspond to statistical regularities in language.

3.1.4 MAXIMUM ENTROPY APPROACH

Maximum entropy approach [10] is also based on statistical analysis that combines diverse pieces of contextual evidence in order to estimate the probability of linguistic. This approach allows the unrestricted use of contextual features and combines them in a principled way.

4 HISTORY OF NLP AND NLIDBS

The history of Natural Language interfaces to Databases as old as the history of Natural Language Processing (NLP) research and NLP history starts from the late 1940's that may also be called the first phase of NLP that focuses on the Machine Translation (MT) and MT was the first computer based application for natural language. In 1946 the two scientists While Weaver and Booth starts one of the earliest MT projects. In 1957 Chomsky [11] introduces the idea of generative grammar by publishing Syntactic Structure. In 1967 W.A WOODS introduced procedural semantics [12] for question-answering machine and in 1970 he introduced the ATNs i.e. Augmented Transition Networks [13]. From 1980's towards a lot of research have been done by many researchers some of them including Discourse Representation Theory "a theoretical framework for dealing with issues in the semantics and pragmatics of anaphora and tense" in 1981 [14], MUMBLE system [15] by David D. McDonald in 1980, Non-context-freeness of NL syntax proven by Stuart M. Shieber in 1985 [16] etc. if we look at the history of Natural language processing, we will found that it has different application over the period including in Machine and Automatic Translation, Information Retrieval (IR), Information Extraction (IE), Dialogue Systems and Question-Answering etc. As Natural Language Interface is the most easy and helpful way for the novice and casual users to access database who do not understand the complex database query languages, so that a lot of effort has been done by the researchers to provide the intelligent Natural Language Interface to Databases (NLIDBs) to these users.

5 HISTORICALLY DEVELOPED NLIDBS

In this section an overview is given about some of the historically developed NLIDBs.

5.1 LUNAR

The LUNAR [17] was the system that answers questions about samples of rocks brought from the moon. The system uses Augmented Transition Network Parser and Wood's procedural Semantics**.** Two different databases were used in this system one for chemical analyses and one for literature references. The performance of this system was quite impressive. It can handle about 78% of requests without any errors and this ratio increase up to 90% when dictionary errors were corrected.

5.2 RENDEZVOUS

In this RENDEZVOUS system users were able to access databases through unrestricted natural language. In this system, special emphasis given on query paraphrasing and in engaging users in clarification dialogs when there is difficulty in parsing user input.

5.3 PHILIQA

This system was known as Philips Question Answering System [18] this system uses a syntactic parser which runs as a separate pass from the semantic understanding passes. The system is mainly involved with problems of semantics and there for has three separate layers of semantic understanding. First layer is about English Formal Language, second layer is about World Model Language and third layer is about Data Base Language and appear to correspond roughly to the "external", "conceptual", and "internal" views of data.

5.4 LIFER/LADDER

This LADDER [19] system was designed about US Navy ships information as a natural language interface to a database. The system uses the technique of semantic grammars that interleaves syntactic and semantic processing. The question answering is done through parsing. This system support only simple one-table queries or multiple table queries with easy join conditions.

5.5 PLANES

This system includes an English language front-end with the ability to understand and explicitly answer user requests. PLANES (Programmed LANguage-based Enquiry System) [20] carried out the database based upon information of the U.S. Navy 3-M (Maintenance and Material Management) which contains complete records of aircraft and flight data.

5.6 CHAT-80

CHAT-80 [21] system was developed and implemented in Prolog language. The CHAT-80 was an impressive, efficient and sophisticated system in which English text is transferred into prolog expressions, prolog is basically logic programming language that is associate with artificial intelligence and computational linguistic , which were evaluated against the Prolog database.

5.7 ASK

The ASK system allowed the end users to teach new words and concepts to the system at any point during the interaction. ASK was basically a complete information management system that provides its own built-in database and have the ability to interact with multiple external databases, electronic mail program and other computer applications. All the applications connected to this system were accessible to the end-user through natural language request.

5.8 EUFID

The EUFID (End-User Friendly Interface to Data Management) [22] is a flexible system. This system consists of three major modules that are analyzer module, map per module and translator module. The system is written in FORTRAN and C programming languages.

5.9 DATALOG

This system DATALOG (Database Dialogue) [23] is highly portable and extendable. This is a Natural Language in English database query system which based on Cascaded ATN grammar and provides separate representation schemes for linguistic knowledge, general world knowledge, and application domain knowledge.

5.10 TEAM

The system TEAM [24] was designed to be easily configurable by database administrators with no knowledge of NLIDBs because at that time a large part of the research was devoted to portability issues. In TEAM, an English query transform into a database query in two steps. In First step the DIALOGIC system constructs a representation of the literal meaning or logical form of the query. In Second part the data-access component translates the logical form into a formal database query.

5.11 JANUS

This system JANUS [25] was similar to multiple underlying systems (databases, expert systems, graphics devices, etc). The system used a hybrid approach to representation, employing an intentional logic for the representation of the semantics. JANUS system also supports temporal questions.

Table 1. Historical Developed NLIDBs

System Name	Year	Domain
LUNAR	1972	Scientific : Rocks sample from the moon
RENDEZVOUS	1977	General
PHILIQA	1977	General
PLANES	1978	U.S Navy 3-M
LADDER	1978	US-Navy Ships
EUFID	1980	General
CHAT-80	1980	General
DATALOG	1985	General
TEAM	1987	General
JANUS	1989	General
ASK	1996	Information Management

6 RECENTLY DEVELOPED NLIDBS

In this section an overview is given about some of the recently developed NLIDBs.

6.1 PRECISE SYSTEM

PRECISE [26] introduces the idea of semantically tractable sentences which are sentences that can be translated to a unique semantic interpretation by analyzing some lexicons and semantic constraints. The model used for this is called Semantic Tractability Model. The system adopts a heuristic based approach. It was evaluated on two database domains. The first one is the ATIS domain, the second domain is the GEOQUERY domain.

6.2 NALIX SYSTEM

NALIX (Natural Language Interface for an XML Database) [27] can be classified as syntax based system. The transformation processes are done in three steps, generating, validating and then translating the parse tree to an XQuery expression. The way in which system was built is implements a reversed-engineering technique by building the system from a query language toward the sentences.

6.3 WASP SYSTEM

WASP (Word Alignment-based Semantic Parsing) [28] is a system that learns to build a semantic parser given a corpus a set of natural language sentences annotated with their correct formal query languages. The whole learning process is done using statistical machine translation techniques and it does not requires any prior knowledge of the syntax. WASP was also evaluated on the GEOQUERY domain. The system was evaluated on different natural languages including English, Japanese, Turkish and Spanish.

6.4 STEP SYSTEM

The STEP (Schema Tuple Expression Processor) system [29] is for natural language access to relational databases. This system use bidirectional grammar to paraphrases queries back to the natural language. It uses structured semantic grammar through coupling phrasal patterns.

6.5 GENERIC INTERACTIVE NATURAL LANGUAGE INTERFACE TO DATABASES (GINLIDB)

The Generic interactive natural language interface to databases (GINLIDB) system [30] is designed and developed by the use of UML and Visual Basic.NET-2005. This system is generic in nature. This system has also uses Augmented Transition Network and ATN grammar that can add to knowledge bases if new queries are entered.

6.6 C-PHRASE SYSTEM

The C-Phrase system [31] is a web-based natural language interface and the database made available to the users through a web-based query interface. It runs under LINUX operating system.

Table 2. Recently Developed NLIDBs

System Name	Year	Domain
PRECISE	2004	ATIS (Air-Travel) & GEO-Query (U.S. Geography)
NALIX	2005	General
WASP	2006	General
STEP System	2007	General
GINLIDB System	2009	General
C-Phrase System	2010	Web-based

7 NATURAL LANGUAGE INTERFACES OTHER THAN ENGLISH LANGUAGE

This section includes of some of the recently developed NLIDBs that are in other natural languages than that of English.

7.1 URDU NATURAL LANGUAGE INTERFACE

The interface [32] that is proposed for this system is based on formal semantics and use AV mapping algorithm to deal with the tokens. This algorithm maps the particular value of token into corresponding attribute token. This converts the query from Urdu language to SQL query. There is about 85 % accuracy in results produces by this system.

7.2 HINDI NATURAL LANGUAGE INTERFACE

This interface [33] maps the query in Hindi language to SQL query and produces the result in same Hindi language by using semantic matching. The architecture for this system have different phases, first it tokenize the query (i.e. in Hindi) and then map it and form SQL query and execute it.

Table 3. NLIs other than English Language

System Name	Year	Domain
Efficient Transformation of a Natural Language Query to SQL for Urdu	2009	General
Hindi Language Interface to Database using Semantic Matching	2013	Employee Database

8 TECHNIQUES FOR DEVELOPING NLIDBS

Here are the techniques that are following in different NLIDBs

8.1 PATTERN MATCHING

Some of the early developed NLIDBs rely on pattern matching techniques [34] to answers the questions of the users. The advantage of using pattern matching technique is that its simplicity that is no elaborating parsing and interpretation modules are needed and systems easily implemented. This technique also helps the systems to come up with reasonable answers even if the questions are out of the range for which the systems are designed.

8.2 SYNTAX BASED

The syntax based systems are those in which the users questions are parsed (i.e. syntactically analyzed) and resulting parse tree is directly mapped an expression in some database query language. It is normally very difficult to devise mapping rules that will directly transform parse tree into some database query language expression. This system uses a grammar that describes the possible syntactic structure of the question by the users.

8.3 SEMANTIC GRAMMAR

The Semantic grammar system also handle question answering by parsing input and mapping the parse tree to database query. The only difference in this technique is that the grammar's categories that are non-leaf nodes that will appear in the parse tree do not necessary correspond to syntactic concepts. This introduced as an engineering methodology, which allows including semantic knowledge into the system easily. Systems that are developed on this semantic grammar are difficult to port into other knowledge domain because semantic grammars contain hard-wired knowledge about a specific knowledge domain.

8.4 INTERMEDIATE REPRESENTATION LANGUAGES

This technique Intermediate Representation Languages [35] is used by most current NLIDBs, they first transform natural question into an intermediate logical query that expressed in some internal meaning representation language. In this approach the system can be divided into two portions. The first portion starts from a sentence up to the generation of a logical query in this part the use of logic query languages makes it possible to add reasoning capabilities to the system by embedding the reasoning part inside a logic statement, and the second portion starts from a logical query until the generation of a database query.

9 FUTURE OF NLP AND NLIDBS

The future of NLP and NLIDBs is very bright but we should notice that there are some limitations. If we somehow successful to elaborate more conveniently the grammars, semantic and syntactic rules etc during development we can achieve more accurate and reliable natural language interfaces. It may be thought that in future these systems gives further help and support to the users to interact with without hesitations. Those users may have the goal to always interact with the

system in natural language. All data processing qualities and upward capability may increase and may also give efficient and trustworthy support.

10 CONCLUSION

This paper has been attempted with the purpose to serve the reader with field of NLIDBs and NLP. A lot of research has been done from last few decades in this field. Although several NLIDBs system have been developed so far but the use of NLIDBs systems is not wide spreader and also it is not a standard option for interfacing to a database. Many new interfaces are also developed that bring some hope that in near future the domain of natural language interfaces will get impressive output from the users in terms of usage because the expert users are also fed up with traditional queries writing by keyboarding.

ACKNOWLEDGMENT

We are very thankful to our worthy and respected teachers including Dr. Muddesar Iqbal (Director of Computing&IT), Mr. Mubbashar Hussain (HOD of IT.dep), Mr. Fiaz Majeed, Miss Qamar-Un-Nisa and Miss Komal Zaman for their support and guidance.

REFERENCES

[1] Jonathan Grudin and Norman, “Language Evolution and Human-Computer Interaction”, Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society, 611-616, 1993.
[2] Bruce G. Buchanan, “A (Very) Brief History of Artificial Intelligence”, 25th Anniversary Issue, American Association for Artificial Intelligence, 2005.
[3] Geert Booij, “Morphological analysis”, Bernd Heine and Heiko Narrog (eds.), The Oxford Handbook of Grammatical Analysis, Oxford University Press, 563-589, 2009.
[4] P.Wittenburg, W.Peters and S.Drude, “Analysis of Lexical Structures from Field Linguistics and Language Engineering”, Third international conference on language resources and evaluation, 2002.
[5] Hilda Koopman, Dominique Sportiche and Edward Stabler, “An Introduction to Syntactic Analysis and Theory”, Wiley-Blackwell, 2013.
[6] Abdul Karim Bangura and Gandhi’s Satyagraha “A Pragmatic Linguistic Analysis of Its Meanings”, The International Journal of Language Society and Culture, Issue 20, 2007.
[7] Didier Dubois, Henri Prade, Lluis Godo and Ramon Lopez de Mantaras, “A Symbolic Approach to Reasoning with Linguistic Quantifiers”, Proceedings of the Eighth international conference on Uncertainty in artificial intelligence Pages 74-82, 1992.
[8] Roger Garside, Geoffrey Leech and Geoffrey Sampson, “The Computational Analysis Of English: A Corpus-Based Approach”, (eds.) University of Lancaster and University of Leeds London: Longman, xii+ 196 pp. 1987.
[9] Martin Redington and Nick Chater, “Connectionist and Statistical Approaches to Language Acquisition: A Distributional Perspective”, LANGUAGE AND COGNITIVE PROCESSES, 129–191, 1998.
[10] Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra, “A maximum entropy approach to natural language processing” Journal Computational Linguistics Volume 22 Issue 1, Pages 39-71, March 1996.
[11] Noam Chomsky, “ASPECTS OF THE THEORY OF SYNTAX”, THE M.LT. PRESS Massachusetts Institute of Technology Cambridge, Massachusetts, August 1965.
[12] W. A. WOODS, “Procedural semantics for a question-answering machine” AFIPS, fall joint computer conference, part I Pages 457-471, 1968.
[13] W. A. WOODS, “Transition Network Grammars for Natural Language Analysis”, Magazine Communications of the ACM, Volume 13 Issue 10, Pages 591-606, 1970.
[14] Hans Kamp, Josef van Genabith and Uwe Reyle, “Discourse Representation Theory”, Handbook of Philosophical Logic Volume 15, pp 125-394, 2011.
[15] David D. McDonald, “MUMBLE: a flexible system for language production”, IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2 Pages 1062-1062, 1981.
[16] Stuart M. Shieber, “Evidence Against the Context-Freeness of Natural Language”, The Formal Complexity of Natural Language Studies in Linguistics and Philosophy Volume 33, pp 320-334, 1987.
[17] Woods, W., Kaplan, R. and Webber, B. “The Lunar Sciences Natural Language Information System” Bolt Beranek and Newman Inc., Cambridge, Massachusetts Final Report. B. B. N. Report No 2378, 1972.

[18] R.J.H., Scha, “Philips Question Answering System PHILIQA1”, In SIGART Newsletter, no.61. ACM, New York, February 1977.
[19] Hendrix, G., Sacrdoti, E., Sagalowicz, D. and Slocum, J. “Developing a natural language interface to complex data”. ACM Transactions on Database Systems, Volume 3, No. 2, USA, Pages 105 – 147, 1978.
[20] David L. Waltz, “An English Language Question Answering System for a Large Relational Database”, Communications of the ACM, Volume 21, July 1978.
[21] Warren D and Pereira, F, “An efficient and easily adaptable system for interpreting natural language queries in Computational Linguistics.” Volume 8 pages 3 – 4, 1982.
[22] I. Kameny, J. Weiner, M. Crilley, J. Burger, R. Gates and David Brill, “EUFID: The End User Friendly Interface to Data Management Systems”, Proceeding VLDB '78 Proceedings of the fourth international conference on Very Large Data Bases - Volume 4 Pages 380-391, 1978.
[23] Carole D. Hafner, “Interaction of Knowledge Sources in a Portable Natural Language Interface”, Proceeding COLING '84 Proceedings of the 10th international conference on Computational linguistics Pages 57-60, 1984.
[24] Barbara J. Grosz., “TEAM: A transportable natural language interface system”, Proceedings of the Conference on Applied Natural Language Processing held at Santa Monica, California, ed. Association for Computational Linguistics, 39-45. Morristown, N.J.: Association for Computational Linguistics, 1983.
[25] Ralph Weischedel, “A hybrid approach to representation in the Janus natural language processor”, ACL '89 Proceedings of the 27th annual meeting on Association for Computational Linguistics Pages 193-202, 1989.
[26] Ana-maria Popescu , Alex Armanasu , Oren Etzioni , David Ko and Alexander Yates, “Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability”, COLING '04 Proceedings of the 20th international conference on Computational Linguistics Article No. 141, 2004.
[27] Yunyao Li, Huahai Yang, and H. V. Jagadish, “NaLIX: an Interactive Natural Language Interface for Querying XML”, ACM SIGMOD’05 international conference on Management of data Pages 900-902, 2005.
[28] YukWah Wong and Raymond J. Mooney, “Learning for Semantic Parsing with Statistical Machine Translation”, In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2006). pp. 439-446, New York City, NY, June 2006.
[29] Michael J. Minock, “A STEP Towards Realizing Codd’s Vision of Rendezvous with the Casual User”, VLDB '07 Proceedings of the 33rd international conference on Very large data bases, 2007.
[30] Faraj A. El-Mouadib, Zakaria Suliman Zubi, A. Almagrous and I. El-Feghi, “Generic Interactive Natural Language Interface to Databases (GINLIDB)”, Proceedings of the 10th WSEAS International Conference on EVOLUTIONARY COMPUTING., 2009.
[31] Michael Minock, “C-Phrase : A system for building robust natural language interfaces to databases”, Journal Data & Knowledge Engineering, Volume 69 Issue 3, Pages 290-302, 2010.
[32] Rashid Ahmad, Mohammad Abid Khan and Rahman Ali, “Efficient Transformation of a Natural Language Query to SQL for Urdu”, Proceedings of the Conference on Language & Technology 2009.
[33] Ashish Kumar And Kunwar Singh, “Hindi Language Interface to Database using Semantic Matching”, ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, Vol.6, No.2: Pgs.133-140, June 2013.
[34] Elaine Califf and Raymond, “Relational learning of pattern-match rules for information extraction”, American Association for Artificial Intelligence, 1999.
[35] Keerthi and Kamal Adusumilli, “Natural Languages Translation Using an Intermediate Language”, IAENG International Journal of Computer Science, 33:1, March 2007.

AUTHOR'S BIOGRAPHY

UMAIR SHAFIQUE- Pursuing M.Sc degree in Information Technology
from Department of Information Technology,
University of Gujrat, Gujrat Pakistan.

HASEEB QAISER- Pursuing M.Sc degree in Information Technology
from Department of Information Technology,
University of Gujrat, Gujrat Pakistan.