International Journal of Computer Science and Applications,
Vol. 5, No. 3, pp 70 - 90

NATURAL LANGUAGE INTERFACE USING SHALLOW PARSING

RAJENDRA AKERKAR and MANISH JOSHI†

Technomathematics Research Foundation
Kolhapur, India
raakerkar@yahoo.com

This paper deals with a natural language interface, which accepts natural language questions as inputs and generates textual responses. In natural language processing, key-word matching based paradigm generate answers, however these answers frequently affected by certain language dependant phenomena such as semantic symmetry and ambiguous modification. Available techniques, described in the literature, deal with these problems using in depth parsing. In this paper, we will present rules to tackle linguistic phenomena using shallow parsing and discuss advantages of a novel Natural Language Interface comprising of shallow parsing based algorithms in conjunction with some intelligent techniques to train the system. Experimental results show that this approach can analyze a wide range of questions with high accuracy and produce reasonable textual responses.

Keywords: Natural language interface; semantic symmetry; ambiguous modification; shallow parsing. .

1. Introduction

Database management systems (DBMS) have been widely used thanks to their efficiency in storing and retrieving data. However, databases are often hard to use since their interface is quite rigid in cooperating with users. To get information from a data-base, the user has to fill some search criteria into a predefined form and receive results as a table or a fixed report. Such an interacting method is inconvenient for users who do not know the structure of the database being used. Using natural language to interact with a database is a better choice for users, especially non-expert ones.

The research on natural language interface to databases (NLI2DB) has recently received attention from the research communities. The purpose of natural language interfaces is to allow users to compose questions in natural language and to receive responses under the form of tables or short answers. Due to implicit ambiguity of natural language, current natural language interfaces are often implemented in a specific domain and can only understand a subset of a natural language. This paper discusses our approach to tackle linguistic phenomena occurring at semantic level using shallow parsing.

In order to understand language, language is studied at different levels. Liddy (2003) propose following seven levels for linguistic analysis. Bhattacharyya (2003) also

† *Current Address: University of New Brunswick, Fredericton, Canada. E-mail: joshmanish@gmail.com.*

elaborated how these levels of linguistic analysis are useful in the Interlingua approach for Machine Translation.

Phonological

Phonetics is the interpretation of speech sounds within and across words. It is the study of language in terms of the relationships between phonemes whereas phonemes are the smallest distinct sound-units in a given language (Matthews, 1997). Phonetic knowledge is used, for example, for building speech-recognizing systems. Though most Natural Language Processing systems do not need to operate at this level, speech recognition systems heavily depend on this analysis.

Morphology

It is the study of the meaningful parts of words. It deals with componential nature of words, which are composed of morp hemes. Morphemes are the smallest elements of meaning in a language. Morphological knowledge is used, for example, for automatic stemming, truncation or masking of words.

Lexicology

Lexicology is the study of words. According to Matthews (1997) lexical level of analysis is defined as, 'of or relating to words or the vocabulary of language as distinguished from its grammar and construction'. This level refers to parts-of-speech tagging or the use of lexicons (dictionaries, thesauri, etc.). Lexicons are utilized in IR system to ensure that a common vocabulary is used in selecting appropriate indexing or searching terms / phrases.

Syntactic

The syntactic level of linguistic analysis is concerned with how words arrange themselves in construction. Syntax is the study of the rules, or "patterned relations", that govern the way the words in a sentence are arranged. Syntactic rules are used in parsing algorithms. Meaning can be derived from word's position and role in sentence. The structure of a sentence conveys meaning and relationship between words, even if we do not know what their dictionary meanings are. All this is conveyed by the syntax of the sentence. Natural Language Processing systems, in their fullest implementation, make good use of this kind of structural information.

Semantics

Semantics involves the study of the meaning of word. This is more complex level of linguistic analysis. The study of the meaning of isolated words may be termed lexical semantics. The study of meaning is also related to syntax at the level of the sentence and to discourse at the level of text. By using both syntactic and semantic levels of analysis, Natural Language Processing systems can identify automatically phrases of two or more words that when looked at separately have quit e different meanings (Hunt, 2000).

Discourse Analysis

Although syntax and semantics work with sentence-length units, the discourse level of NLP works with units of text longer than a sentence" (Liddy, 2003). This level relies on the concept of predictability. It uses document structure to further analyze the text as a whole. By understanding the structure of a document, Natural Language Processing

systems can make certain assumptions. Examples from information science are the resolving of anaphora and ellipsis and the examination of the effect on proximity searching.

Pragmatics

Pragmatics is often understood as the study of how the context (or "world knowledge") influences meaning. This level is in some ways far more complex and work intensive than all the other levels. This level depends on a body of knowledge about the world that comes from outside the document. Though it is easy for people to choose the right sense of the word, it is extremely difficult to program a computer with all the world knowledge necessary to do the same.

The above levels of linguistic processing reflect an increasing size of unit of analysis as well as increasing complexity and difficulty as we move from phonological level to pragmatics. The larger the unit of analysis becomes, (i.e., from morpheme to word, to sentence, to paragraph and to full document) the less precise the language phenomena. It decreases in precision results in fewer discernible rules and more reliance on less predictable regularities as one moves from the lowest to the highest levels. Additionally, higher levels presume reliance on the lower levels of language understanding, and the theories used to explain the data move more into the areas of cognitive psychology and artificial intelligence. As a result, the lower levels of language processing have been more thoroughly investigated and incorporated into Natural Language Processing related systems (Liddy, 1998).

1.1 Natural Language Interface to Structured Data

Using Natural Language to communicate between a database system and its human users, has become increasingly important since database systems have become widespread. In order to facilitate full use of the database systems, its accessibility to non-expert users is desirable. BASEBALL (Green, 1961) and LUNAR (Woods, 1972) were the first usable Natural Language Interface to Database (NLI2DB), which appeared in late sixties. The BASEBALL system was designed to answer questions about baseball games which were played in the American league during any one season. LUNAR NLI2DB contained chemical analyses of moon rocks and had a significant influence on subsequent computational approaches to natural language.

PLANES (Waltz, 1975), LADDER (Hendrix, 1978), and REL (Thompson, 1975) systems were developed by late 1970s. Some of these systems used semantic grammars. This is an approach in which non-terminal symbols of the grammar reflect categories of world entities (e.g. student_name, Design ation_of_employee) instead of purely syntactic categories like noun phrase, verb phrase etc.

Subsequently several NLI 2DB had appeared with different approaches towards handling Natural Language. By the mid-1980s, NLI2DB had become a very popular research area, and numerous research prototypes were implemented. Systems like CHAT-80 (Warren,

1981), DIALOGIC (Moore, 1981), incorporated some novel and ingenious techniques, and its implementation responded to queries very promptly. DIAGRAM (Robinson, 1982), TELI (Ballard, 1986) and JANUS (Bobrow, 1990), were also among numerous research prototypes of that period.

Historically, computing scientists have divided the problem of natural language access to a database into two sub-problems: the Linguistic component and the Database component (Turoff, 1985). The Database component performs traditional Database Management functions, whereas the Linguistic component is responsible for translating natural language input into a formal query and generating a natural language response based on the results from the database search. A lexicon is a table that is used to map the words of the natural input onto the formal objects (relation names, attribute names, etc.) of the database (Allen, 1995). Both parser and semantic int erpreter make use of the lexicon. A natural language generator takes the formal response as its input, and inspects the parse tree in order to generate adequate natural language response. Natural language database systems make use of syntactic knowledge and knowledge about the actual database in order to properly relate natural language input to the structure and contents of that database. Of course, the system expects the user to ask questions pertaining to the domain of the database, which in turn represents some aspect of the real world. Syntactic knowledge usually resides in the linguistic component of the system, in particular in the syntax analyzer whereas knowledge about the actual database resides to some extent in the semantic data model used. Knowledge about the user and goals of his speech act are required if a user-friendly dialogue is to be carried through. Questions entered in natural language translated into a statement in a formal query language. Once the statement unambiguously formed, the query is processed by the database management system in order to produce the required data. These data then passed back to the natural language component where generation routines produce a surface language version of the response.

NLI2DB were seen as a promising way to make databases accessible to users with no programming expertise, and there was a extensive optimism about their commercial prospects. Linguistic Technology's ENGLISH WIZARD was among some of these systems claimed to have been commercially successful.

The use of NLI2DB, however, is much less widespread than it was once predicted, mainly because of the development of alternative graphic and form-based database interfaces (Zloofs "query by example" technique). But these alternative interfaces are less natural to interact with and queries that involve quantification, or that require multiple database tables to be consulted are very difficult to formulate with graphic or form-based interfaces, whereas they can be expressed easily in natural language.

In order illustrate to provide generalized idea of working of Natural Language Interface to Database, we present one of the successful NLI2DB developed. (Stratica, 2005) discussed their results with CINDI NLI2DB developed by him and his colleagues. They

developed a template-based system for translating English sentences into SQL queries for a relational database system. The input sentences are syntactically parsed using the Link Parser, and semantically parsed through the use of domain-specific templates. The system is composed of a pre-processor and a run-time module. The pre-processor builds a conceptual knowledge base from the database schema using WordNet. This knowledge base is then used at run time to semantically parse the input and create the corresponding SQL query. The system is meant to be domain independent and has been tested with the CINDI database that contains information on a virtual library. Generalized architecture of the system is shown in Figure 1.

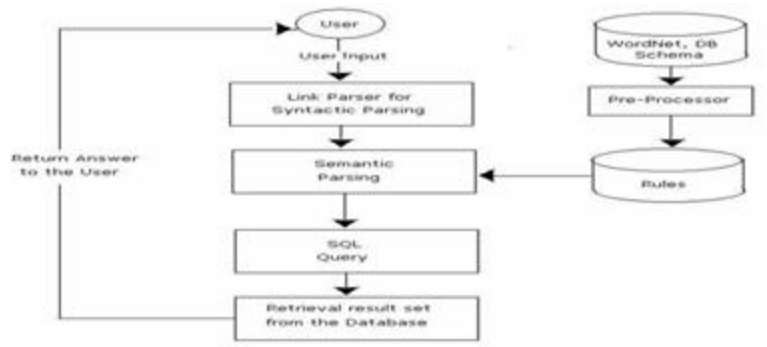

Figure 1: Generalized architecture of the NLI2DB system.

In order to tolerate lexical variations in the input questions, the preprocessor builds a semantic knowledge base composed of interpretation rules and semantic sets for all possible relation and attribute names in the database. This helps to build the same semantic representation, and hence, the same SQ L query for various questions.

In the above NLI 2DB example also we saw use of Link Parser which collects information at syntactic level and it can be used at semant ic level. Focus of NLI 2DB nowadays shifted to handle problems at higher level of linguistic analysis. Therefore development of NLI2DB systems that handle language related phenomena is an active area of research (Bandyopadhya, 2004).

1.2 Natural Language Interface to Unstructured Data

The process of establishing interaction between human being and machine was made successful in 1966 by ELIZA system, which was developed by Joseph Weizenbaum (Weizenbaum, 1966). ELIZA worked by simple parsing and substitution of key words into phrases stored in knowledge base. Though ELIZA did not employed any language related phenomena still it remains a milestone simply because it was the first time a programmer had attempted such a human-machine interaction with the goal of creating the illusion of human-human interaction.

Dialogue systems were historically the domain of AI researchers. These systems are sort of the natural Language Interface to unstructured data. These systems do not restrict themselves to interact with data in database tables only. Data from various sources can be used and accumulated.

Moving forward through the history of Natural Language Interface to unstructured data research brings us to SHRDLU (Winograd, 1972) and GUS (Bobrow, 1977). Both of these systems are dialogue systems interacting on information about a restricted domain. The difference between these systems and systems such as LUNAR (discussed in section 2.2.1) are their dialogue capabilities. GUS was designed to simulate a travel advisor and had access to a database containing limited information about airline flight times. SHRDLU is probably the better known of these two systems. It controlled a robot arm in a virtual micro-world consisting of a table top strewn with coloured blocks of varying shapes and sizes and a box into which the blocks could be placed. Whilst example conversations with SHRDLU are generally impressive, the system is still severely limited to only discussing the micro-world it inhabits.

A program called PARRY (Heiser, 1980) was also developed in that period. Presently there is a vast amount of NLP-based research carried out for the development of such systems. One modern Natural Language System is Jupiter (Zue, 2000), probably best described by its product page at MIT: *"Jupiter is a conversational system that provides up-to-date weather information over the phone. Jupiter knows about 500+ cities worldwide (of which 350 are within the US) and gets its data from four different Web-based sources"*.

Clearly Jupiter is more complex than systems such as SHRDLU as the system is dealing with input via the telephone and hence has to cope with the added problem of robust speech recognition to provide a reasonable input to the dialogue system. Note, however, that just as SHRDLU was limited to questions about the block world it inhabited to, Jupiter is limited to questions about weather reports for the cities it is aware of.

Boris Katz at MIT's Artificial Intelligence Laboratory developed the START (SynTactic Analysis using Reversible Transformations) Natural Language System. It is a software system designed to answer questions that are posed to it in a natural language. START uses several language dependant functions like parsing, natural language annotation to present the appropriate information segments to the user.

Since large information is available in unstructured manner, retrieving out relevant documents containing the required information was the primary goal of the interfaces of this category. This task is known as Information Retrieval. Pinpointing exact information called as Information Extraction is the next step and development of Question Answering System is advanced step of human machine interaction.

Information extraction is the task of locating specific pieces of data from a natural language document, and has been the focus of DARPA's MUC program (Lehnert, 1991). The extracted information can then be stored in a database which could then be queried using either standard database query languages or a natural language database interface. (Califf, 1998)

However, a difficulty with information extraction systems is that they are difficult and time-consuming to build, and they generally contain highly domain-specific components, making porting to new domains also time-consuming. Thus, more efficient means for developing information extraction systems are desirable.

Question answering system establishes non-formal communication with the user and tries to answers to the natural language queries asked by the user. These Question Answering systems are sort of Natural Language Interface to unstructured data.

The *Text REtrieval Conference (TREC)*, co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 and being organized regularly. Question Answering is one of the important aspects of Text Retrieval and several groups present their current work in this conference. KUQA system presented in TREC-9 (2000) developed by Soo-Min Kim and his colleagues categorized questions based on expected answer and then used NLP techniques as well as Wordnet for finding candidate answers which suits in corresponding category. They however not handled any linguistic phenomena (Kim, 2000).

In TREC-13 (2004) Michael Kaisser and Tilman Becker presented QuALiM system that used complex syntactic structure. Based on certain syntactic description question patterns were identified. Syntactic description of prospective answers was also maintained and accordingly from documents retrieved using google search engine system generated answers (Kaisser, 2004).

The concept of combining Natural Language Processing (NLP) techniques with large-scale Information Retrieval and Information Extraction is not new, is yet not successful to the desired extent. Fagan (1987) experimented with indexing noun phrases and prepositional phrases. (Croft, 1987), (Smeaton, 1994), (Strzalkowski, 1996), (Zhai, 1996) and (Arampatzis, 1998) experimented with indexing syntactically derived word pairs; the types of constructions examined in the context of indexing include linguistically motivated pairs such as head/modifier and adjective/noun. In addition, (Smeaton, 1994), (Croft, 1987) tried full linguistic trees and case frames as units of indexing. However, none of these experiments resulted in dramatic improvement in precision or recall, and often even resulted in degraded performance. In all of theses studies, the word-level index was directly augmented with linguistically-derived representations. Often, this caused performance issues because the creation of an index is limited by the speed of the parser, and because sophisticated linguistic representations were not amenable to large-

scale indexing. The current generation of question answering systems that employ NLP alleviate performance problems by delaying linguistic analysis until the corpus has been narrowed down to a small set of candidate documents or passages. The MURAX System (Kupiec, 1993) is an early example of such an approach. Litkowski (1999) described a system that utilizes a combination of syntactic relations, like subject-verb-object, and some semantic relations, like time and location. After retrieving a set of candidate documents, the system then parses both the question and the passages and attempts matching at the relation level. Unfortunately, this and similar techniques that depend heavily on syntactic analysis, for example, PIQASso (Attardi, 2001), yielded relatively poor performance. A drawback of this two step paradigm is low recall: if the keyword-based document retrieval system does not return *any* relevant documents due to such problems as synonymy, anaphora, or argument alternations, any amount of additional processing is useless. The current work-around to this problem is to implement feedback loops that relax the query set if the results are too restrictive (Moldovan, 2002). Not only does this introduce complex dependencies in a system's architecture, but it also necessitates the addition of new modules to assess the quality of the result sets.

But Katz and Lin of MIT showed that NLP technology in general is not powerless (Katz, 2003). Performance drop or performance improvement of the interface, on the contrary depends on the manner in which NLP techniques has been applied. Katz and Lin identified two broad linguistic phenomena that are difficult to handle with the simple keyword matching driven paradigm. They tackled the linguistic phenomena of semantic symmetry and ambiguous modification using ternary relations.

The literature survey reveals the fact that Natural Language Interface is an area of interest for number of researchers. Large numbers of experiments are being carried out worldwide to develop effective Natural Language Interface. Though several different approaches are exploited to achieve success, every system has certain shortcomings.
We will now present various approaches available to deal with language.

2. Various Approaches

Natural language is the topic of interest from computational viewpoint due to the implicit ambiguity that language possesses. Several researchers applied different techniques to deal with language. Next few sub-sections describe diverse strategies that are used to process language for various purposes.

2.1 Symbolic Approach (Rule Based Approach)

Natural Language Processing appears to be a strongly symbolic activity. Words are symbols that stand for objects and concepts in real worlds, and they are put together into sentences that obey well specified grammar rules. Hence for several decades Natural Language Processing research has been dominated by the symbolic approach (Miikkulainen, 1997).

Knowledge about language is explicitly encoded in rules or other forms of representation. Language is analysed at various levels to obtain information. On this obtained information certain rules are applied to achieve linguistic functionality. As Human Language capabilities include rule-base reasoning, it is supported well by symbolic processing. In symbolic processing rules are formed for every level of linguistic analysis. It tries to capture the meaning of the language based on these rules.

2.2 Empirical Approach (Corpus Based Approach)

Empirical approaches are based on statistical analysis as well as other data driven analysis, of raw data which is in the form of text corpora. A corpus is collections of machine readable text. The approach has been around since NLP began in the early 1950s. Only in the last 10 years or so empirical NLP has emerged as a major alternative to rationalist rule-based Natural Language Processing.

Corpora are primarily used as a source of information about language and a number of techniques have emerged to enable the analysis of corpus data. Syntactic analysis can be achieved on the basis of statistical probabilities estimated from a training corpus. Lexical ambiguities can be resolved by considering the likelihood of one or another interpretation on the basis of context.

Recent research in computational linguistics indicates that empirical or corpus-based methods are currently the most promising approach to developing robust, efficient natural language processing (NLP) systems (Church, 1993; Charniak, 1993). These methods automate the acquisition of much of the complex knowledge required for NLP by training on suitably annotated natural language corpora, e.g. tree-banks of parsed sentences (Marcus, 1993).

Most of the empirical NLP methods employ statistical techniques such as n-gram models, hidden Markov models (HMMs), and probabilistic context free grammars (PCFGs). Given the successes of empirical NLP methods, researchers have recently begun to apply learning methods to the construction of information extraction systems (McCarthy, 1995), (Soderland, 1995), (Riloff, 1993, 1996), (Huffman, 1996). Several different symbolic and statistical methods have been employed, but most of them are used to generate one part of a larger information extraction system. (Majumder, 2002) experimented N-gram based language modeling and claimed to develop language independent approach to IR and Natural Language Processing.

2.3 Connectionist Approach (Using Neural Network)

Since human language capabilities are based on neural network in the brain, Artificial Neural Networks (also called as connectionist network) provides on essential starting point for modeling language processing (Wermter, 1997). In the recent years, the field of connectionist processing has seen a remarkable development. The sub-symbolic neural network approach holds a lot of promise for modeling the cognitive foundations of

language processing. Instead of symbols, the approach is based on distributed representations that correspond to statistical regularities in language.

There has also been significant research applying neural-network methods to language processing (Reilly, 1992; Miikkulainen, 1993). However, there has been relatively little recent language research using sub-symbolic learning, although some recent systems have successfully employed decision trees (Magerman, 1995; Aone, 1995), transformation rules (Brill, 1993, 1995), and other symbolic methods (Wermter, 1996). SHRUTI system developed by (Shastri, 1997) is a neurally inspired system for event modeling and temporal processing at a connectionist level.

3. Semantic Level Problems

Any Natural Language Application based on the concept of keyword matching has to face some common problems at semantic level. We have decided to put emphasis on problems caused by linguistic phenomena of semantic symmetry and ambiguous modification. This section describes problems raised due to these phenomena. We further elaborate shallow parsing based algorithms to deal with it.

Semantic symmetry and ambiguous modification are linguistic phenomena, which occur quite often. Semantic symmetry is a linguistic phenomenon in which word order matters. Two sentences with same keywords may have different meaning. It means at word level (lexical level) two sentences looks to be the same but at semantic level these two sentences vary. An Interface for Natural Language or a Question Answering System has to consider this fact and answer generated by the system must drop out sentence with wrong meaning and must present sentence with correct meaning only.

Ambiguous modification is the process which needs to be handled carefully. Instead of modifying actual expected noun from the sentence, adjective in the question modifies other unexpected noun which results in infiltration of wrong sentences into the answer. Due to this performance of the Interface degrades.

Both these linguistics phenomena cause problems which directly affects on overall performance of the system. Both these phenomena occur at semantic level. Most of the systems tackling any problem at semantic level (like Machine Translation, Writing theme of the document etc.) need information collected at syntactic level (syntactic analysis). Even the only system, tackling problems caused by semantic symmetry and ambiguous modification (Katz, 2003) collects information at syntactic level and builds ternary relations based on the output obtained from a Minipar parser.

(Katz, 2003) have implemented Sapere, a prototype natural language question answering system that retrieves answers by matching ternary expressions derived from the question with those derived from the corpus text. They compared the results obtained from Sapere

system with a simple boolean retrieval engine that uses a standard inverted keyword index to index documents at the sentence level.

Sapere (Katz, 2004) is primarily a relations-indexing engine; it stores and indexes ternary expressions extracted from the corpus text and perform matching at the relation level between questions and sentences stored in its index. Ternary expressions are generated from text by postprocessing the results of Minipar, a fast and robust functional dependency parser. Sapere system detects the following types of relations: subject-verb-object (including passive constructions), adjective-noun modification, noun-noun modification, possessive relations, predicate nominatives, predicate adjectives, appositives, and prepositional phrases.

Ternary expressions are similarly derived from the question, with the *wh*-entity left as an unbound variable. Sapere attempts to match relations in the question with those found in the corpus text, thereby binding the unbound variable in the question with the actual answer. If such a match occurs, the candidate sentence is returned.

System developed by Katz and Lin generates appropriate answer but at the cost of time. Because time required for Minipar parser followed by time required for developing ternary relations delays the response to the question asked by the user.

4. Shallow Parsing Based Algorithms

This section helps understanding effect of semantic symmetry and ambiguous modification phenomena, and describes shallow parsing based algorithms to overcome problems caused by these phenomena.

4.1 Semantic Symmetry

Semantic Symmetry occurs when an entity is used as subject as well as an object in different sentences and because of which in Question Answering System, selectional restriction (keyword matching) in different sentences, based on such entities; generates wrong answer. Following Example (Joshi 2005) illustrates the phenomenon of semantic symmetry and demonstrates problems caused thereof.

Question : Who killed militants ?

Candidate Answer 1: National army soldiers killed six militants.
Candidate Answer 2 : Militants killed 13 bus passengers.

In above sentences 'Militants' is an entity (POS – Noun) which acts as subject in sentence 2 and as an object in sentence 1. The selectional restriction for the subject of 'kill' is word 'Militants' in one sentence and the selectional restriction for the object is also word 'Militants' in another sentence. Thus, a semantically symmetric relation involves sentences where one can swap the subject and object and still end up with a sentence that makes sense.

Hence, candidate answers fetched on the basis of keyword matching needs to be monitored carefully as these sentences may have different meaning altogether. In above example, system returns two sentences based on mere keyword matching. Question in above example is referring to the sentences, which contains semantic symmetry relationship.

Two sentences 'National army soldiers killed six militants.' and 'Militant killed 13 bus passengers.', are similar at the word level, but they have very different meanings and should be presented as answer appropriately by considering meaning of these sentences. In these cases, lexical content is insufficient to determine the meaning of the sentence.

4.1.1 Pattern of Sentences and Semantic Symmetry

All sentences are categorized as Active voice sentences or Passive voice sentences. Sentences in Active voice follow the structure of SVO that is Subject followed by Verb followed by Object, whereas Passive voice sentences have OVS structure that is Object followed by Verb followed by Subject. The sentence in Active voice can also be presented in Passive voice by changing position of Subject and Object and changing verb to past participle form.

Questions may be of type XVO where X is the subject entity which we want to find out as an answer, V is the verb and O is the object entity. 'Who killed Militants? ', is XVO type question. Now only those active sentences are correct in which object entry do not come before verb entry that is the order 'VO' is maintained. If this correct sentence available in passive voice form (OVS) then sequence of Verb and Object in question shall not match, but we can understand that this sentence is in Passive voice form by looking at the verb. If the verb in the sentence is in Past Participle form it indicates that the sentence is in Passive voice form. Hence sequence is not matched and POS of verb not matched points to the correct answer. So, we can formulate the rule which emphasizes on sequence and tense of verb.

This same logic is applicable for the questions of type – SVX that is X is the Object entity which we want to obtain as an answer, V is the verb and S is the subject entity. 'Militants killed whom?' is an example of such category. Any sentence in which word 'Militants' appears after Verb 'killed', is an object entity whereas we are looking for sentences in which word 'Militants' is placed as Subject. It emphasizes the importance of sequence of words in sentence. In Passive voice sentence SVO sequence is not followed but the passiveness of the sentence can be determined by the tense of the verb, which helps in deciding the correctness of the sentence. This fact underlines importance of tense of verb (Past tense or Past Participle). If the question is in Passive voice form then too the above mentioned logic works. XOV* type of questions falls in this category. 'By whom Militants were killed? ', is the example of this type of questions. X – is the

Subject which we expect from the Question Answering System, O is the object and V* is the past participle tense verb used in the question.

4.1.2 An Algorithm

Based on the study of patterns of questions and the sentences, we have formulated rules that pick up exact sentences as answer among from number of candidate answer sentences. Following algorithm shows how the problem caused by semantic symmetry is solved.

Two important factors considered are -

i) sequence of keywords in question and in the candidate answer sentence
ii) POS of keywords (especially verb keyword)

The algorithm scans each candidate answer sentence and applies following rule to check whether that sentence is correct answer sentence or not.

```
Rule 1 -
If (sequence of keywords in question and candidate answer matches) then
      If (POS of verb keyword are same) then Candidate answer is Correct
Otherwise -
  Candidate Answer is wrong

Rule 2 -
  If (sequence of keywords in question and candidate answer do not match)
  then
      If (POS verb keyword are not same) then Candidate answer is Correct
  Otherwise -
  Candidate Answer is wrong
```

The detailed discussion on this algorithm can be found in (Joshi (2006), Joshi (2008)).

4.2 Ambiguous Modification

Adjectives are often ambiguous modifiers. If a paragraph contains a pool of adjectives and nouns, any particular adjective could potentially modify many nouns. Under such circumstances, a Natural Language Interface System cannot achieve high precision without exactly identifying the association between adjective and nouns.

Ambiguous modification related with adjective occurs when an entity behaves in *un*restrictive manner and can associate to more than one noun in a particular sentence.

4.2.1 An Algorithm

We have formulated rules, after studying structure of sentences, which are based on shallow parsing. Candidate answers are fetched based on keyword matching can be tested

for correctness using these rules. Every sentence that is amenable for problem due to Ambiguous Modification contains one adjective and more than one noun. One of these nouns is used for defining the scope whereas the other is pointing to the identifier which we are looking for. These nouns can easily be distinguished and sequence of these two nouns and adjective is the important factor that is used in formulation of rules.

The algorithm scans each candidate answer sentence and applies following rule to check whether that sentence is correct answer sentence or not. We have identified the adjective as Adj, scope defining noun as S_N and the identifier noun as I_N.

Rules –
If the sentence contains keywords in following order –
Adj a S_N, where a indicate string of zero or more keywords,
Then

R_1-a →If a is I_N ==➔ Correct Answer
Or
R_1-b →If a is Blank ==➔ Correct Answer
Else
R_2 →If a is Otherwise ==➔ Wrong Answer

If the sentence contains keywords in following order –
S_N a Adj ß I_N Where a and ß indicate string of zero or more keywords,
Then

R_3 →If ß is Blank ==➔ Correct Answer
Else
R_4 →If ß is Otherwise ==➔ Wrong Answer.

5 The System

We have developed a system called int**E**lligent **N**atural **L**anguage **I**nterface - ENLIGHT that incorporates these algorithms. It incorporates algorithms that deal with semantic symmetry and ambiguous modification, the frequently occurring phenomena in English Language, using information collected at lexical level. Therefore, we present results of comparison from two different perspectives. Firstly, comparing results of ENLIGHT system with basic keyword matching system, in order to prove effectiveness of the new approach proposed by us and secondly, comparing the ENLIGHT with Sapere system proposed by Katz et al to check the efficiency of the system.

The test corpus used in carrying out experiments on the system was electronic versions of news from newspapers. Approximately 2000 news extracts, information broacher of several educational institutions and documents of TREC-2005 was provided to the basic system. As for the collection of experimental questions, we questioned employees at the inquiry counter about the questions asked by the people to them. They have provided us

some questions from that domain. Questions for news extract are formulated manually by us whereas we have taken some questions from TREC question databases. Questions formulated by (Greenwood, 2005) are also used. For detailed experimental results and system features refer (Joshi (2005), Joshi (2006), Joshi (2008), Joshi (2008a)).
The current prototype of our system has achieved the following results:

- A friendly interface has been provided to users: they can query the database using natural language and receive results in the form of textual responses.
- The system accepts quantified questions and negative questions, which are very difficult to express.
- The system can assist users to rephrase questions correctly to his/her intention.
- Incomplete queries can be automatically corrected without asking users to pick a choice.
- The system is portable to other domains.

To improve the system performance, future work includes:

- Enriching the knowledge sources of the system in order to increase the system efficiency;
- Researching methods to improve the coherence and the fluency of output texts.

6. Conclusion & Future Work

Algorithms developed for ENLIGHT system are useful to improve precision of search engines. It is expected that search engines should return documents by considering meaning of the query rather than just matching keywords. Most of the search engines operate on keyword matching paradigm.

We have tested a few well known search engines and found that they did not take care of such phenomena. We pre-processed the result displayed on first page, returned by *Google* search engine. This data is processed so that it can be used by ENLIGHT system. When the same query is asked to ENLIGHT system, it initially returned all earlier presented sentences. But when we run 'answer rescoring' module, ENLIGHT system eliminated those sentences that are correct at word level but irrelevant at meaning level.

Functioning of search engines is at the verge of drastic change. Upcoming third generation search engines are concentrating on linguistic aspects rather than mere keyword matching. In this situation, we feel that the approach proposed in the paper is more effective because of two reasons – (i) Problems caused by linguistic phenomena at semantic levels are tackled without carrying in-depth parsing and (ii) while dealing with the problems, system does not compromise with response time.

References

[Ahn, 2004] Ahn D., Jijkoun V., Mishne G., Muller K., de Rijke M., Scholobach S., "Using Wikipedia at the TREC QA Track", *TREC-2004*.

[Allen, 1995] Allen J. F., "Natural Language Understanding", *Menlo Park, CA: Benjamin/Cummings*, 1995.

[Androutsopoulos, 1995] Androutsopoulos I., Ritchie G., Thanisch P., "Natural Language Interface to Database: An Introduction", *Journal of Natural Language Engineering*, 1/1, pp. 29-81.

[Androutsopoulos, 2002] Androutsopoulos I., Ritchie G., Thanisch P., "Database Interface", *Handbook of Natural Language Processing*, Dale, Moisl and Somers (Eds.), pp. 209-233.

[Arampatzis, 1998] Arampatzis A., Th. P. van der Weide, Koster C. H. A., P. van Bommel, "Phrase-based information retrieval", *Information Processing and Management*, 4(6), pp 693-701.

[Attardi , 2001] Attardi G., Cisternino A., Formica F., Simi M., Tommasi A., Zavattari C., "PIQASso: Pisa question answering system", *TREC 2001.*

[Ballard, 1986] Ballard B., Stumberger D., "Semantic acquisition in TELI", *Proceeding of the 24th Annual meeting of ACL*. New York, 20-29.

[Bandyopadhya, 2004] Bandyopadhya S., "Cross Language Database System: Multilingual Human Computer Interface", *Lecture Notes*, Jadavpur University, Kolkata.

[Bandyopadhya, 2003] Bandyopadhya S., "Natural Language Processing", *Refresher Course in NLP*, Jadavpur University, Kolkata.

[Barroso, 2003] Barroso L., Dean J., Holzle U., "Web search for planet: The Google Cluster Architecture", *IEEE Micro*, March-April 2003, pp. 22-28.

[Bhattacharyya, 2003] Bhattacharyya P., "Natural Language Processing", *Universal Networking Language Workshop*, IIT, Mumbai.

[Bobrow, 1977] Bobrow D., Kaplan R., Kay M., Norman D., Thomson H., Winograd T., "GUS, a Frame Driven Dialog System", *Artificial Intelligence*, 8(2), pp. 155–173.

[Bobrow, 1990] Bobrow R., Resnik P., Weischedel R., "Multiple underslying systems: translating user request into programs to produce answers", *Proceeding of the 28th Annual meeting of ACL*, Pittsburgh, pp. 227-234.

[Callif, 1998] Callif M., "Relational Learning Techniques for Natural Language Information Extraction", *Ph.D Thesis*, Report AI98-276, The University of Texas at Austin, Austin.

[Chaitanya, 2000] Chaitanya V., Akshar Bharati, Sangal R., "Natural Language Processing: A Paninian Perspective", *PHI Publication*, 81-203-0921-9.

[Chandioux, 1976] Chandioux J, "METEO: UN system".

[Charniak, 1993] Charniak E., "Statistical Language Learning", *MIT Press*.

[Church, 1993] Church K., Mercer R., "Introduction to the special issue on computational linguistics using large corpora", *Computational Linguistics*,19 (1), pp. 1-24.

[Cowie, 2002] Cowie J., Wilks Y., "Information Extraction", *In Handbook of Natural Language Processing*, Dale, Moisl and Somers (Eds.), pp. 241- 260.

[Croft, 1987] Croft B., Lewis D., "An approach to natural language processing for document retrieval", In *SIGIR-1987*.

[Cui, 2003] Cui H., Li K., Sun R., Chua Tat-Seng, Kan Min-Yen, "Question Answering Main Task", National University of Singapore at TREC-13.

[Duisburg, 1994] Duisburg G., "Information on Computational Linguistics," *Office note*, Gerhard- Mercator University Press.

[Easterbook , 2003] Easterbook S., "How thesis get written : Some cool tips", *Lecture Note*, University of Totronto.

[Fagan, 1987] Fagan J., "Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparisons of Syntactic and Non-Syntactic Methods", *Ph.D. thesis*, Cornell University.

[González , 2006] Juan J. González B., Rodolfo A. Pazos Rangel, I. Cristina Cruz C., Héctor J. Fraire H., Santos Aguilar de L. and Joaquín Pérez O, "Issues in Translating from Natural Language to SQL in a Domain-Independent Natural Language Interface to Databases," *Lecture Notes in Computer Science*, Volume 4293/2006, MICAI 2006: Advances in Artificial Intelligence.

[Green, 1961] Green B. F., Wolf A. K., Chomsky C., Laughery K., "BASEBALL: An Automatic Question Answerer", *Proceedings of the Western Joint Computer Conference*, Volume 19, pp 219–224. Reprinted in (Grosz et al., 1986), pp. 545–549.

[Greenwood, 2005] Greenwood M., "Open-Domain Question Answering", *Ph.D. Thesis*, University of Sheffield, UK.

[Haiqinq, 2006] Haiqing H., "A study on Question Answering System Using Integrated Retrieval Method", *Doctoral Thesis* submitted to Tokushima University, Japan.

[Heiser , 1980] Heiser J., Colby K., Faught W., Parkison R., "Can Psychiatrists Distinguish a Computer Simulation of Paranoia from the Real Thing?", *Journal of Psychiatric Research*, (15), pp.149–192.

[Hendrix, 1978] Hendrix G., Sacerdoti E., Sagalowicz D., Slocum, J., "Developing a Natural Language Interface to Complex Data", *ACM Transactions on Database Systems*, 3(2), pp. 105-147.

[Hirschman, 2003] Hirschman L., Gaizaukas R., "Natural Language Question Answering: The view from Here", *Natural Language Engineering*, 7(4), pp. 275-300.

[Hunt, 2000] Hunt C., "Natural Language Processing and role of Linguistic Analysis", Lecture Notes.

[Joshi, 2005] Joshi M., Akerkar R., "Algorithm to Effectively Handle Semantic Symmetry in Question Answering Systems", *CSIT-2005*, 5-8080-631-7, pp. 246-250.

[Joshi, 2006] Joshi M., Akerkar R., "Shallow parsing based algorithms to improve precision of Question Answering Systems", *National Workshop on Artificial Intelligence*, SIGAI, G-DAC.

[Joshi, 2008] Joshi M., Akerkar R., "Solving Semantic Level Problems Using Shallow Parsing Algorithms", *International Journal of Computer & Communication Technologies, 1 (1)*.

[Joshi, 2008a] Joshi M., Akerkar R., "Algorithms to Improve Performance of Natural Language Interface," *International Journal of Computer Science and Application*, Volume (5), Issue 2.

[Jurafsky, 2002] Jurafsky D., Martin J., "Speech and Language Processing", *Pearson Education Publication*, 81-7808-594-1.

[Kaisser, 2004] Kaisser M., Becker T., "Question Answering by selecting Large Carpora with linguistic methods", *Proceedings of the thirteenth TExt REtrieval Conference (TREC-2004)*.

[Katz, 2002] Katz B., Lin J., "START and beyond", Proceedings of 6th World Multiconference on Systemics, Cybernetics, and Informatics.

[Katz, 2003] Katz B., Lin J., "Selectively using relations to Improve precision in Question Answering", *Proceedings of EACL*.

[Katz, 2004] Katz B., Lin J., "Sapere: From Keywords to Key Relations", MIT Computer Science and Artificial Intelligence Laboratory.

[Kennedy, 2000] Kennedy C., Branimir, Boguraev, "Anaphora for everyone: Pronominal Anaphora Resolution without a parser".

[Kim, 2000] Soo-min Kim, Dae-ho Baek, Sang-Beom Kim, Hae-Chang Rim, "Question Answering considering Semantic categories and co-occurance density", *Proceedings of the ninth TExt RErieval Conference (TREC-2000)*.

[Kothari, 2002] Kothari N., Sanghi D., "Intelligent Railway Informaion System".

[Krovetz, 1993] Krovetz R., "Viewing morphology as an inference process", *SIGIR-93 ACM*, pp. 191-202.

[Kupiec, 1993] Kupiec J.,"MURAX: A robust linguistic approach for question answering using an an on-line encyclopedia" , SIGIR

[Landaur , 1997] Landaur T., Laham D., Rehder B., Schreiner M., "How well can passage meaning be derived without using word order: A comparison of latent semantic analysis and humans", *COGSCI-97*, pp. 412-417.

[Lehnert, 1991] Lehnert W., Sundheim B., "A performance evaluation of text-analysis technologies", *AI Magazine*, 12 (3), pp. 81-94.

[Liddy, 1998] Liddy Elizabeth, "Enhanced Text Retrieval using Natural Language Processing", *ASIS Bulletin*.

[Liddy, 2003] Liddy Elizabeth, "Natural Language Processing", *Encyclopedia of Library and Information Science*, New York, Marcel Dekker.

[Lin, 2003] Lin J., Katz B., "Question Answering Techniques for World Wide Web", *11th Conference of European chapter of Association of Computational Linguistics*, EACL-2003.

[Litkowski , 1999] Litkowski K., "Question-answering usingsemantic relation triples", *TREC-8*.

[Lyman, 2003] Lyman P., Varian H., "How much Information?".

[Majumder, 2002] Majumder P., Mitra M., Chaudhari B., "N-gram: A Language Independent Approach to IR and Natural Language Processing", Lecture Notes.

[Marcus, 1993] Marcus M., Santorini B., Marcinkiewicz M.," Building a large annotated corpus of English: The Penn Treebank", Computational Linguistics, 19 (2), pp. 313-330.

[Matthews, 1997] Oxford Concise Dictionary of linguistics, Oxford University Press, Oxford, New York.

[McCarthy, 1995] McCarthy J, Lehnert W, "Using decision trees for coreference resolution", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1050-1055.

[McCoy, 1998] McCoy K., Penninngton C., Badman A., " Language, Comparison: From research prototype to practical integration", *Natural Language Engineering*, 4(1), pp. 73-95.

[Miikkulainen, 1993] Miikkulainen R., "Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory", *MIT Press*, Cambridge, MA.

[Miikkulainen, 1997] Miikkulainen R., "Natural language processing with subsymbolic neural networks", *Neural Network Perspectives on Cognition and Adaptive Robotics.*

[Mitra, 2003] Mitra M., "Open Domain Question Answering: An overview", Lecture Notes.

[Moldovan , 2002] Moldovan D., Pas M. ¸ca, S. Harabagiu, Surdeanu M., "Performance issues and error analysis in an open-domain question answering system", *ACL.*

[Moldovan, 2006] Moldovan D., Clark C., Harabagiu S., Hodges D., "Cogex: A semantically and contextually enriched logic prover for question answering", *Journal of Applied Logic, In Press, Corrected Proof.*

[Moore, 1981] Moore, R., "Problems in Logical Form", *Proceedings of the 19th Annual Meetings of the Association for Computational Linguistic*, California June 1981, pp. 117-124.

[Moss, 2005] Moss A., "Program transformation of embedded systems", *Ph.D. Thesis*, submitted to the University of Bristol, 2005.

[Newwell, 1998] Newwell A., Langer S., Hickey M., "The role of natural language processing in alternative and augmentative communication", *Natural Language Engineering*, 4(1), pp. 1-16.

[Palmer, 2002] Palmer D., "Tokenisation and Sentence Segmentation", Dale, Moisl, Somers (Eds.), pp. 11-33.

[Plath, 1976] Plath W. J., "REQUEST : A Ntural Language Question Answering", *IBM Journal of Research Development.*

[Porter, 1980] Porter, "An algorithm for suffix stripping", *Program* , Vol. 14, number 3, pp. 130-137.

[Reilly, 1992] Reilly R., Sharkey N. (Eds.) "Connectionist Approaches to Natural Language Processing", *Lawrence Erlbaum and Associates*, Hilldale, NJ.

[Rich, 2001] Elaine R., Knight K., "Artificial Intelligence", *Tata McGraw-Hill Edition*, Second Edition.

[Riloff, 1993] Riloff E., "Automatically constructing a dictionary for information extraction tasks", *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 811-816.

[Riloff, 1996] Riloff E., "Automatically generating extraction patterns from untagged text", *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 1044-1049.

[Robinson, 1982] Robinson J., "DIAGRAM: A Grammar for Dialogues", *Communications of ACM*, Vol. 25, No. 1, pp. 27-47.

[Schwitter, 1999] Schwitter R., Aliod D., Hess M., "ExtrAns – Answer Extraction from Technical Documents by Minimal Logical Forms and Selective Highlighting (forthcoming)", *The Third International Tbilisi Symposium on Language, Logic and Computation*, Batumi, Georgia, pp. 12-16.

[Seneff, 1998] Seneff S., Ed Hurley, Lau R., Pao C., Schmid P., Zue V.,"Galaxy-II: A Reference Architecture for Conversational System Development", *Proceedings of ICSLP 98*, pp.931–934.

[Shastri, 1997] Shashtri L., "A model of rapid memory formation in the hippocampal system", *Proceeding of Meeting of cognitive Science Society*, Stanford.

[Simon, 1983] Simon H., "Why should machine learn?", *Machine Learning, An Artificial Intelligence approach*, ed. R S Michalski, J G Carbonell, T M Mitchell. Pao Alto, *Tioga Press.*

[Smeaton, 1994] Smeaton A., O'Donnell R., Kelledy F., "Indexing structures derived from syntax", *TREC-3: System description.* In *TREC-3*.

[Soderland, 1995] Soderland S., Fisher D., Aseltine J., Lehnert, "Crystal: Inducing a conceptual dictionary", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1314-1319.

[Stratica, 2005] Stratica N., Kosseim L., Desai B., "Using semantic templates for a natural language interface to the CINDI virtual library", *Data & Knowledge Engineering* (55), *Elsevier Publications*, pp. 4–19.

[Strzalkowski, 1996] Strzalkowski T., Guthrie L., Karlgren J., Leistensnider J., Lin F., Perez-Carballo J., Straszheim T., Wang J., Wilding J., "Natural language information retrieval: TREC-5 report", In *TREC-5*.

[Sugimoto, 2004] Sugimoto T., Ito N., Iwashita S., Michio Sugeno owards, "Programming in Everyday Language: A Case for Email Management", *Computational Linguistics and Intelligent Text Processing*, Volume 2945/2004.

[Thompson, 1975] Thompson, F.B., Thompson, B.H., "Practical Natural Lansuage Processing: The REL System as Prototype", *Advances in Computers*. Rubinoff H., Yovlts M. (Eds.) pp. 109-168, 13, Academic Press, New York.

[Turoff, 1985] Murrey Turoff, Natural Language and Computer Interface Design, Report of computerized conferencing, NJIT, January.

[Uszkoreit, 2000] Uszkoreit H., "What is Computational Linguistics?", *A short non-technical review*, University of Saarland, Germany.

[Vicedo. 2001] Vicedo J., Antonio L., "A semantic approach to Question Answering", *TREC-9*.

[Waltz, 1975] Waltz, D., "Natural Language Access to a Large Data Base: An Engineering Approach", *Proc. 4th International Joint Conference on Artificial Intelligence*, Tbillsl, USSR, pp. 868-872.

[Wang , 2006] Wang F., "Knowledge Representation in a Behavior-Based Natural Language Interface for Human-Robot Communication", *Lecture Notes in Computer Science*, Volume 4114/2006.

[Wardhaugh, 2003] Wardhaugh D., "Understanding English Grammar – A Linguistic Approach", *Blackwell Second Edition*. 7(4), pp 123-156.

[Warren, 1981] Warren, D., "Efficient Processing of Interactive Relational Database Queries Expressed in Logic", *Proc. Seventh International Conference on Very Large Data Bases*", Cannes, France, pp. 272-283.

[Weizenbaum, 1966] Weizenbaum J, "ELIZA- A computer program for the study of natural language communication between man and machine", *Communications of the ACM*, 9(1), pp.36-45.

[Wermter, 1977] Wermter S., "Hybrid approaches to neural network-based language processing", *Technical Report TR-97-030*, International Computer Science Institute.

[Winograd, 1972] Winograd T., "Understanding Natural Language", *Academic Press*, New York.

[Woods, 1972] Woods, W., Kaplan R., N-Nebber B., "The Lunar Sciences Natural Language Information System", *BBN Report 2378*, Bolt Beranek and Newman, Cambridge, Massachusetts.

[Zarri, 1998] Zarri G., "Natural Language Processing : Associated with Expert Systems", *CRC Press*, 0-8493-3106.

[Zhai, 1996] Zhai C., Tong X., Milic-Frayling N., Evans D., "Evaluation of syntactic phrase indexing— CLARIT", *NLP track report*, In *TREC-5*.

[Zue, 1998] Zue V., "Galaxy-II: A Reference Architecture for Conversational System Development", p*roceedings of ICSLP- 98*, pp. 931–934.

[Zue, 2000] Zue V., Seneff S., Glass J., Polifroni J., Pao C., Hazen T., Heatherington L., "JUPITER: A Telephone-Based Conversational Interface for Weather Information", *IEEE Transactions on Speech and Audio Processing*, 8(1), pp. 100–112.