ISSN:0975-9646

Mallamma V. Reddy et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3236 - 3238

Semantical and Syntactical Analysis of NLP

Mallamma V. Redd[#1], Hanumanthappa M.[2]

[1]*Department of Computer Science, Rani Channamma University, Vidyasangam,Belgaum-591156,India*

[2]*Department of computer science, Bangalore University, Jnanabharathi Campus,Bangalore-560056,India*

***Abstract*— Natural language processing describes the use and ability of systems to process sentences in a natural language such as English or any other Indian Languages, rather than in specialized artificial computer languages such as C, C++. This paper deals with Syntactical and Semantical analysis of Indian languages such as Kannada for machine translation, which plays a vital role in accurate machine translation for NLP. The accurate machine translation leads to an accurate cross language information retrieval. The Syntactical and Semantical structures for machine translation are presented with an example.**

***Keywords*— Cross language information retrieval (CLIR), Information Retrieval (IR), Natural Language processing (NLP).**

I. INTRODUCTION

Kannada language is highly agglutinative language with three gender forms namely masculine, feminine and neutral and Word order plays an important role in positional languages like English which normally follow right-branching with Subject-Verb-Object orders where as In Kannada language is verb final language and all the noun phrases in the sentence normally appear to the left of the verb, hence it is 'Left branching language' and the adjectives, genitive and relative clauses precede their head nouns in a sentence. The subject noun phrase may also appear in many different positions relative to other noun phrases in the sentence.

South Dravidian language like Kannada is having almost 40 million speakers and has its own independent script and long documented histories. Even though Kannada is a language rich in literature, its resources are poor when viewed through the prism of computational linguistics. The development of NLP in Kannada language is not explored much and is in the beginning stage compared to other Indian languages. Moreover, Kannada is a highly agglutinative and morphologically rich language. Syntactic and semantic variance makes the problem much harder for creating full fledged computational linguistic tools and machine translation (MT) system for Kannada language. With the absence of computational linguistic tools and English to Kannada MT system so far, even a reasonable domain specific development can find its immediate applications in government and educational sectors.

In natural language processing, syntactic parsing or more formally syntactic analysis is the process of analyzing and determining the structure of a text which is made up of sequence of tokens with respect to a given formal grammar. A syntactic parser is an essential tool used for various NLP applications and natural language understanding. Syntactic parsing of sentences is considered to be an important intermediate stage for semantic analysis in NLP application such as Information Retrieval, Information Extraction and Question Answering.

The study of the structure of a sentence is called syntax and it attempts to describe the grammatical order in a particular language in terms of rules which details an underlying structure and a transformational process. Syntax provides rules to put together words to form components of sentence and to put together these components to form meaningful sentences [1]. Because of the substantial ambiguity present in the human language, whose usage is to convey different semantics, it is much difficult to design the features for natural language processing tasks. The main challenge is the inherent complexity of linguistic phenomena that makes difficult to represent the effective features for the target learning models.

II. NATURAL LANGUAGE PROCESSING

All Natural language processing (NLP) is primarily concerned with getting computers to perform useful and understanding tasks with human languages. In natural language processing [2] system first the words are placed into a structured form that leads to syntactical correct sentence. Typical applications for natural language processing include the following.

- A better human-computer interface that could convert from a natural language into a computer language and vice versa.
- A translation program that could translate from one human language to another (English to Kannada, for example). Even if programs that translate between human languages are not perfect, they would still be useful in that they could do the rudimentary translation first, with their work checks and corrected by a human translator. This cuts down on the time for the translation.
- Programs that could check for grammar and writing techniques in a word processing document.
- A computer that could read a human language could read whole books to stock its database with data

Allen mentions the varied types of knowledge relevant to natural language understanding:

1) *Phonetic and phonological knowledge***:** how words are related to sounds.The field of phonetics is a multilayered subject of linguistics [2] that focuses on

speech. In the case of oral languages (phonetics) as a research discipline has three main branches or areas of study:

- Articulatory phonetics is concerned with the articulation of speech: The position, shape, and movement of articulators or speech organs, such as the lips, tongue, and vocal folds.
- Acoustic phonetics is concerned with acoustics of speech i.e. the study of the physical transmission of speech sounds from the speaker to the listener: The spectro-temporal properties of the sound waves produced by speech, such as their frequency, amplitude, and harmonic structure.
- Auditory phonetics is concerned with speech perception: the perception, categorization, and recognition of speech sounds and the role of the auditory system and the brain in the same. In other words it is the study of the reception and perception of speech sounds by the listener.

2) *Morphological knowledge:* how words are built from more primitive morphemes (e. g., how "friendly" comes from "friend." Morphology deals with the different inflections of a word, the forms it can take: a noun can be singular or plural; a verb can have different tenses, and so forth. Morphemes include "run," "laugh," "non-," "-s," and "es." Programs can be written to process tokens of words or even the more basic level of morphemes.

3) *Syntactic knowledge:* how sequences of words form correct sentences. Knowledge of the rules of grammar.

4) *Semantic knowledge:* how words have "meaning"; how words have reference (denotation) and associated concepts (connotations).

5) *Pragmatic knowledge:* "how sentences are used in different situations and how use affects the interpretation of the sentence," this involves the intentions and context of the conversation.

6) *Discourse knowledge:* how preceding sentences determine the meaning of a sentence, such as in the case of the referent of a pronoun.

7) *World knowledge:* general knowledge about for example, other user beliefs and goals in a conversation.

A. Syntactical Analysis

Syntactic analysis [3] gets at the structure or grammar of the sentences. Processing a sentence syntactically involves determining the subject and predicate and the place of nouns, verbs, pronouns, etc. Given a lexicon telling the computer the part of speech for a word, the computer would be able to just read through the input sentence word by word and in the end produce a structural description. But problems arise for several reasons.

First of all, a word may function as different parts of speech in different contexts (sometimes a noun, sometimes a verb, for example).

For example, "the fox runs through the woods" treats "fox" as a noun, whereas "the fox runs through the woods were easy for the hounds to follow" uses it as an adjective. You don't know how "fox" is used until you read the entire sentence. So we have to determine which part of speech is relevant in the particular context at hand.

Second, and related to this, there may be several possible interpretations of the structure of a sentence. How are we to decide which is the correct analysis?

Third, in searching for the interpretation of a sentence, there may be different ways to do this.

Ex: cat eat rat

Here it follows the syntactical rule as

Subject-cat
Verb-eat
Object –rat

B. Syntactical Analysis for Kannada Language

Kannada Language being one of the major Dravidian languages of India and it has 27th place in most spoken language in the world. But still it does not yet have computerized grammar checking methods for a given Kannada sentence. When Computational Linguistic is concerns Kannada is lagging far behind compared to Telugu and Kannada. Writing the grammar production for any south Indian language is bit difficult. Because the languages are highly inflected with three gender forms and two number forms. In most of the Indian languages including Kannada a verb ends with a token which indicates the gender of the person (Noun/ Pronoun.

A Kannada syntactic parser tool recognizes a Kannada sentence and assigns a syntactic structure to it. For example the input to the syntactic parser is a Kannada sentence, ‘‘ರಾಮ ಚೆಂಡನ್ನು ಎಸೆದನು’’. ‘‘ Rama threw the ball.‘‘

The syntactic parser recognizes the sentence by solving lexical and attachment ambiguities and assigns a syntactic structure to it in the form of a parse tree as shown in Fig. 1.

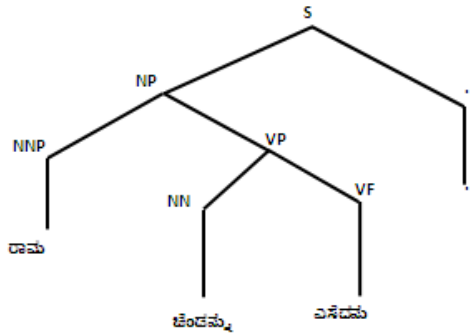

Fig 1: Syntactic tree structure

C. Semantic Analysis

Semantics, as a branch of linguistics, aims to study the meaning in language. As one knows that a language exhibits a meaningful message because of the semantic interaction with the different linguistic levels phonology, lexicon and syntax. However, the field of semantics, too, contributes towards stylization. That means any discussion of the semantic features of literary style implies a discussion of the nature of the semantics in literary texts.

There are seven types [4] of meanings namely; logical or conceptual, connotative, stylistics, affective, reflected

collocative and thematic. He made a significant distinction between two meanings: one is conceptual meaning which is known as denotative and tied down to the grammatical structures of a sentence and the second one is stylistics, i.e., connotative meaning which while depending on denotative meaning, gives readers additional information about the utterance. This indicates that denotative or linguistic meaning is direct whereas stylistic meaning is implicit and is dependent on the literary context of usage [5].

Semantic analysis [6] deals with the meaning of words and sentences, the ways that words and sentences refer to elements in the world. "Meaning" in these discussions is usually associated with semantic. Semantic knowledge is a method of representing knowledge. The goal is to reduce the syntactic structures and provide the meaning.

D. Semantical Analysis for kannada

Once the computer has arrived at an analysis of the input sentence's syntactic structure, a semantic analysis is needed to ascertain the meaning of the sentence. The syntactic structure of a sentence, the NLP system will attempt to produce the logical form of the sentence. Logical form is context-free in that it does not require that the sentence be interpreted within its overall context in the discourse or conversation in which it occurs and logical form attempts to state the meaning of the sentence without reference to the particular natural language. Thus the intent seems to be to make it closer to the notion of a proposition than to the original sentence.

The basic or primitive unit of meaning for semantic [7] will be not the word but the sense, because words may have different senses, like those listed in the dictionary for the same word. Thus different senses can be organized into a set of classes of objects; this representation is called ontology such as substance, quantity, quality, relation, place, time, position, state, action, and affection, events, ideas, concepts, and plans.

Actions and events are especially influential. Events are important in many theories because they provide a structure of organizing the interpretation of sentences. Actions are carried out by agents. Also important is the already mentioned notion of a situation.

Ex: rat eat cat

This sentence follows the Syntactical rule, but semantically the sentence is meaningless. There for the analysis performs the following steps to make it meaningful sentence:

1) The parts of speech of each word in the input sentence and atom format definitions attached to each predicative/attributive word such as verbs, adjectives, prepositions and the others are fetched from the knowledge base.
2) The input sentence is partitioned into a set of clauses by considering the positions of conjunctions, relative pronouns and punctuation marks.
3) Words in a clause are put into four groups preserving the word order in the clause, each of which may contain subjects, the main verb, direct objects/complements respectively. For this purpose, the correspondence relation between the NL syntax and the atomic formula syntax is utilized.
4) Each phrase in the four groups is decided whether it indicates qualification of nouns or of predicates.
5) If there are AND/OR conjunctions in a clause it is decided whether they are in the scope of a preposition or not.
6) After all of the words in a clause have been interrelated each other, the remaining clauses are processed by repeating 3 to 5.
7) After all of the clauses have been processed, substantiation of each personal and demonstrative pronoun is established.
8) In consequence, the extended formula deduced from the input sentence is obtained.

III. MT FOR ENGLISH TO KANNADA LANGUAGE

English to Kannada MT is the application of computers to the task of translating texts from English language to Kannada language. For example, for the input English sentence “Rama threw the ball”, the MT system should produce the equivalent Kannada sentence as “ರಾಮ ಚೆಂಡನ್ನು ಎಸೆದನು”. this sentence is semantically and syntactically meaningful.

IV. CONCLUSION

This paper has outlined an approach for Cross Lingual Information Retrieval which emphasizes pre and post processing strategies for the queries entered in a source language using semantic and syntactic analysis. With the above example we clearly say that without the semantic and syntactic analysis the machine translation result may be ambiguous.

REFERENCES

[1] http://en.wikipedia.org/wiki/Kannada_language
[2] Rada Mihalcea and Dragomir Radev, “Graph based natural language processing and information retrieval” www.cse.unt.edu/~rada/CSCE5290/Lectures/Intro
[3] Roxana Girju, “Introduction to Syntactic Parsing”, 2004.
[4] Geoffrey N. Leech “Semantics : The Study of Meaning” by (1974, Paperback)ISBN-10:-13:9780140216943
[5] Gargesh, R.”Linguistic Perspective on Literary Style”. Delhi: Delhi University Press.1990
[6] http://www.universalteacher.org.uk/lang/semantics.html
[7] Wood, G.C.,” Lecture on Introduction to Semantics at the University of Sheffield”.