Chapter 1

INTRODUCTION

1.1 Natural Language Processing

It goes without saying that communication is an integral part of any language. Language is the basic means of communication for human beings. The inception of electronic and information media in global communication was a huge respite to communication. By and large it is a modern means of communication. Natural language texts in electronic form assumed the status of a principal medium for communication. We have been coming across billions of pages of textual data online. This lead to the requirement of a simplified means of relevant access, useful for well designed information. Electronic documents are constraints to natural language. It has become pertinent that ‘language specific’ (knowledge) information is a must for processing these documents. This leads the indispensability to associate with computerization of human languages.

NLP or Natural Language Processing is an area which interacts with computational aspects of human language. At the outset to achieve a set goal of NLP aimed to visualize and comprehend the level of understanding the content of natural languages are kept to usage by one and all. The objective is to enable encoding of linguistic knowledge into principles of rules or other forms of representation. NLP further includes NLU viz. Natural Language Understanding and Natural Language Generation. A part of speech (POS) tagging is one of the most popular and extensively un-researched assignments in the area of NLP, especially when it
demands a requisite to cover far fetching areas having several complex assignments as complicated tasks.

NLP deals with the language of a document enabling/providing a semantic view. Language is proactive by nature and admits additions and deletions in substance to improve a set of illustrated rules. As part of an exercise, to translate out of vocabulary (OOV) words, translation is done to facilitate NLP applications. This has rules of reference with specific means of culture as well as regional language models gain importance generally in non-English languages [DB97]. The complexity of morphological differences, pertinent to our languages, are yet to be established or explored with the intake or adoptively depends on statistical and language models.

The NERC has a commendable role to play in the application of NLP in machine translation, information retrieval, automatic text brief up, question-answering systems and others. The natural language understanding and natural language generation are the two aspects handled and carried by NLP. Natural language comprehensive understanding is associated with the process of using languages soon after the words are recognized. The purpose is to signify a model of computation that suits the people’s requirements and linguistic tasks of listening, speaking, reading and writing. To bring out a natural model of language understanding it is important to involve knowledge from most of the disciplines inclusive of linguistics, Psycholinguistics, Philosophy computational linguistics etc. It is a relevant requirement to understand the working of a language. It is conceptual to bring in and unify all approaches in order to create theories of complexity and to appreciate and realize such theories with complex nature to enact computer programs. In order to improve efficiency and performance of programs, clues can be obtained by testing
such cases that fail to enable further improvement by repeatedly enacting the process; we could ultimately know how human language processing occurs.

The important constituent, the meaning of the text is enabled by computing the representation. It has become necessary to exemplify the thought of presentation in direction. In the absence of clarity it would rather become a hassle or hindrance. The precision of a language is necessary to bear representation of the meaning. The representation of language should be exact and bear clarity and capture the intuitive structure of the natural language sentences. Language understanding based on grammar does not make a shrewd or a sound sense of decision. An advanced way of learning should enable one to make out or understand every sentence with a mistake without reference to any accomplishment of Grammar (i.e. an advanced system should be able to understand even a sentence with mistakes). Syntactic presentation of the languages, normally depend or rely on the thought of context free grammar. This represents sentence structure with reference to the phrases that are constituents of other phrases, generally found in a tree form.

The context independent meaning of a sentence is presented by the reference of the logical form. The semantic relationship between words and logical form phrases are encoded by a possible word sense which is identified by the logical form. The abstract sets of semantic relationships arise between the verb and its noun phrases. Once the semantic association is finalized or materialized, word senses may become feasible and therefore could be discarded without giving room for thought. The important aspect of main works in semantic interpretation is to consider whether combination of the specific word meanings can unify into a combination to innovate coherent sentence meanings. Manipulating such interconnection between word
meanings can greatly reduce the number of feasible word sense for each word of a given sentence.

- **Natural Language Understanding**

  In the Natural Language Understanding systems TACITUS Hobbs et al. in the year 1993 made use of Coherence relation with an aim to get a total representation of the message [BR92]. To a greater extent, the system is leased on consistent association as put forward by Hobbs in the year 1985.

- **Natural Language Generation**

  Natural Language Generation was tested for effective relations by Marcu in 1997. Based on the strong relations the preferences on discourse segments, was secured [DM95], [AZ96]. The greater the score, the higher the discourse structure was found to be coherent.

1.2 Applications and uses of Natural Language Processing

We have furnished the list in the given manner. It shows some of the frequent tasks on which research was carried on NLP. Some tasks have straight and real world applications while others are more formally used to serve as subtasks or to resolve bigger tasks [CR98]. The outstanding issue is to distinguish these tasks from other NLP tasks. The basis to distinguish is not the research above in terms of its quantity, but the focus converged on them has to be taken into account. Besides, there is a well defined problem setting, a standard metric for validating the task, standard corpora on whose basis the task could be evaluated and the competitiveness extended to the specific task.
Automatic Summarization

The NLP community has investigated the sub field of summarization for about half of the last century. In the year 2002, Redev et al. defined the summary as a text that is created as a production from different corpus. It explains distinct information through the actual; but it is nothing larger than half of the actual texts, however distinctly shorted than that. This goes with three important aspects getting characterized by the process of research on automatic summarization. In simple terms, the following could be understood:-

- Summaries are outputs of one single document and thereafter form multiple ones.
- Summaries should enable preservation of important information.
- Summaries should be brief in size.

In the light of literature, any trial or effort to provide an illustration or definition would conclude in disagreement within the community, though initially we agree in consensus on the points. In reality, most of the approaches take a different stand on the manner of their line of difficulties as problem formulations. We commence to familiarize or to introduce some of the usual terms in the summarization dialect. Extraction enables us to identify relevant chapters of the text so as to create verbatim. Abstraction is to facilitate production of important material in a new manner.

Co-Reference Resolution:

Reference is a unique concept with complexity. It is not out of play to add as to say it is beyond the scope of this presentation on the research area to give away a complete fine grained classification with reference to its types. Further to the details on the philosophical aspects of reference, I would like to recommend to Moore
Anaphora with a root origin of Greek Avatopa. “Carrying back” is a unique kind of anaphora concurred to express another expression preceding it in the discourse as referred to by Cermak in the year 1997.

The phrase, “Pointing Back” is an anaphora, resolution is the process of deciding the precedent or antecedent of an anaphor as pointed out by Mikhov in 2002. When there are two kinds of referrals to the expression viz. an anaphor and its antecedent have the same reference they are names as co refer as to be a co referential.

The corresponding relation (equivalent) is known as co reference. Post occurrences of co-referential expressions constitute a name sale co referential chain. This excerpt comprises an example of a co-referential chain (anaphors are indicatives as in Italics), their antecedents are projected in captions and then finally a co-referring expression bears matching indices.

• Discourse Analysis

A discourse is an enlarged stretch of wordy utterance having consistency. Together these utterances form a well intended organized message. The conferring of coherence can be made and presented in many values, but the outcome is always a kind of relationship between items on different levels in the discourse. In spite of the above, if they want to maintain relationship while in discourse implies assuming units between which this association holds and leads us to units between which this relation holds good and in turn in to the field of discourse breaking and the (Hierarchal) discourse structure.

Discourse markers (but, yet, still etc) are seen towards abridging a stance or a relation between parts of discourse, E.g. // the marker ‘BUT’ will form the relation contrast between two utterances or discourse constituent part or segments. It is
apparently seen as the relation marked on the surface level. The relation ‘Contrast’
can be further signaled unimaginatively only. Relations might be more under pressure
or force (stress) than by E.g. // Grosy & Sidner (1986), whereas the relations of
discourse can be described as inferences or results drawn up on the basis of the
utterances as said during the discourse. The discourse meaning gains strength by the
inferences. It would at times work as an adherent between utterances/segments while
the inferences extend support to the meaning of the message beside the structures. The
inferences can be laid on the basis of the proximity however they remain semantic in
nature. That is how it is given to understand that they are not dependant on the
adjacency (or proximity) in the same manner as the surface signaled discourse
bondages are closely related. The inferences machinery as well as the inherence skills
representation is very crucial.

- **Machine Translation**

  Machine Translation, the term is prominently used to translate, one language
to another. The objective of the MT system is beneficial in creating or enabling
feasible means of translation without any external help. Primarily, even MT systems
require programs meant for translation and machine made dictionaries and grammar
to help in work on translation. For improving the machine translation systems pre-
editing of the input enables the quality of translation. It is by means of
synchronization, the input is ensured for the purpose of pre-editing. Adjusting or
synchronizing the input is done by marking prefixes, suffixes, clause boundaries and
so on. Vocabulary control enables improving the quality of translation. The output of
the MT invariably is subjected to post-editing to bring about perfection. Most post-
editing is needed, in health and hygiene oriented data or information.
MT’s that generate translation between any two specific languages are known as ‘Bilingual’ systems and translation for any pair of languages is known as ‘Multilingual’ systems. Besides, the multilingual systems could be unidirectional or bidirectional. They (Multilingual systems) are opted to be bidirectional and bilingual as they have the ability to translate from one given language to the other and vice versa.

- **Morphological Segmentation:**

  The application of morphological segmentation was first invoked by F.Zwicky, a professor at the California Institute of Technology. This analysis was first experimentally applied by him to the Aerospace industry. He opted to analyze the structure of the jet engine technology. He defined the parameters of this technology that was incorporated with thrust mechanism, oxidizer and fuel type. When all the possibilities were exhausted under each parameter at the apex, alternative approaches were brought out to assemble all feasible permutations. For E.g. // a Ramjet utilizing atmospheric O2 and a solid fuel, Zwicky opined that the permutations showing “Empty cells” reacted for creativity by saying “Why not a nuclear powered ceramic fan jet?” as aptly viewed by Zwicky. Morphological analysis is a manifestation that leads to ‘Organized invention’. The technique permits two important elements.

  i. A systematic analysis of the present and futuristic structure of a domain along with prominent spaces in that structure.

  ii. A strong reaction as stimulant for the location or invention of alternatives to fill the spaces and comply with the compiled accomplishments.
To quote as saying, “Essentially, morphological analysis is a method of identifying and investigating the total set of possible relationships contained in any given multi dimensional problem complex that can be parameterized”.

- **Questions Answering System:**

  Ever since December 1993, the world’s first web-designed question answering method has been enabled by ‘START’. The system has been online and ceaselessly in operation. START undertakes the supply users with “Just the right information”. Presently, the system can respond to millions of English queries about places, locations (Viz. cities, countries, lakes, weather, maps, demographics, political and economic systems), movies (E.g. // titles, actors, directors), people (E.g. // birth, dates, biographies, Lexicon definition and many more details.

- **Optical Character Recognition**

  This system enables to provide an alphanumeric identification of printed or manually written characters at par with electronic speed by a simple means of scanning. In recent years, the intelligent character recognition–ICR has been in use to explain the process of interpreting image data in a specific alphanumeric text.

  The images can be scanned and there after recognition machine (engine) or the OCR system interprets these formed images of manual type of printed characters into ASCII data (readable characters by machine). Therefore, OCR allows users to get enable a fast automated data capture from forms, remove key strokes to minimize data entry costs but maintain a high precision of accuracy and speed that is necessary in form processing applications.

- **Parsing**

  It is a cute process to split a level order or a sequence of characters or their values into minimal sized parts. It would be useful to recognize characters or values
that come across a specific sequence/order or a sequence of characters or their values into minimal sized parts. Additionally it enables the provision of a strong, powerful, readable and maintainable approach to patterns of expression. This method enables to innovate or create one’s own custom language for specific purposes.

The Parse also undertakes acceptance of two refinements: /all and /case. This /all refinement parses all the characters within the string, with all delimiters such as space, tab, new line, semi colon and comma. The case refinement parses a string based on a case when case is not specified upper and lower cases.

- **Parts of Speech Tagger:**

  Tagging is an automatic assignment of descriptors to the given tokens. The descriptor is known as a tag, the indication of the parts of speech and semantic information etc. Tagging is a synonym for classification. The process of allotting one of the parts of speech with reference to the given word is called parts of speech tagging and more simply referred to as POS tagging. Parts of speech cover and include nouns, verbs, adverbs, adjectives, pronouns, conjunctions besides their sub categories. Parts of speech tagger or POS tagger is a program that undertakes this task. Taggers make use of many kinds of information, lexicons, dictionaries, rules and so on and have categories of a specific word. To exemplify that a word may belong to more than one category E.g. // Run is both a noun and a verb such is the area of a tagger where their usage enables probabilistic information to solve this ambiguity. Two types of taggers are

  i. Rule Based

  ii. Stochastic

  The first type uses manual rules to find and locate the tag ambiguity. The second type are either HMM based where the tag sequence is chosen which enhances
the product word possibilities and tag sequence probability or cue based using decision trees or maximum entropy models to combine probabilistic features [CH02]. An ideal tagger should be well defined, efficient, accurate, tunable and recyclable [BA97]. Actually taggers are capable of identifying the tag for an earmarked word or to offer a guess as to the availability of information. Sometimes the task assumes complication as the natural language is complex and the resultants face the difficulty where the taggers make precise decision about the tags. Routine hassles coming by should not be viewed seriously and hence periodic errors in tagging not taken as a constraint to research. The evaluations made out at Message understanding conferences MUC proved that extraction of information from documents are enabled by this method. It would be first used to identify the classes of type of information sources referred to in the text. This led to establishing the NE task, where the systems are updated accordingly to identify the dates, events, times, numerical status and names. NER has retained a vital component of information extraction and connected NLP tasks [AI93]. These include the web based corpus development from the archive of leading news daily in Telugu. NE annotations of part of this corpus are drafting gazetteers, inflection lists and bilingual dictionaries either in manual form or partial semi-automatically verifying the corpus. Besides, a number of (POS) parts of speech tagging models have been developed with utilizing a different machine learning algorithm for usage in the NERC systems.

1.3 Telugu Language and its Complexity

Morphologically, Telugu is enriched like any other Dravidian Language. A good number of surface word forms originated from a simple root of this language. With emphasis on Telugu any generic Dravidian language has an order of complexity...
in its entirety more than those in the Indo-Aryan languages. Telugu words are by and large are more complex and long in composition.

In terms of bytes (Characters) the mean word length is 11.61 and the standard deviation is 4.41, whereas an English word has a mean length of 8.18 with a standard deviation of 3.12. These measures are taken as a base on a 3 million word English corpus derived from the British National Corpus the ‘BNC’

The richness in morphology of Telugu is chiefly attributed to the Grammar mitigated by syntax in English (besides other similar languages) and is handled within the morphology in Telugu (and other Dravidian languages). Several words or phrases (Tokens) in English would be mapped on to a single word in Telugu. Thus Vaccaadu (he came), Vastaadaa (will he come?), Vaste (if he/she/it/they/I/we/you come), ragalagutaaDu (he will be able to come), raalekapoyadu Du (he was unable to come) VaccinavaaDu (the person (3p.sl) who came) raaDanu-konnavaa (did you think he would not come?) are all single words (Tokens) in Telugu written and spoken as atomic units. Verbs may include aspectual auxiliaries apart from tense and agreement. There are many kinds of non-finite forms too. A single verbal root can lead to formation of a few hundred thousand word forms. Nouns are inflected for number of and case. Derivation being more productive, more forms becomes possible when we consider full word forms. Thus vaccinaavadiki (to the person (3p.sl) who came) is a noun in singular, dative case derived from the verb root ‘Vaccu’ (to come). External Sandhi (that is conflation between two or more complete word forms) and compounded to the numbers. Naturally we will see very large number of types and the type token ratio should be expected to be very high too. These are not simple concatenations or juxtapositions of complete words written without intervening spaces as is the convention in some languages of the world. These words are made up
of several mono phonemes conjoined through complex morpho-phonemic processes. Telugu in particular and Dravidian language in general are amongst the most complex languages in the world at the level of morphology, perhaps comparable only to Finnish and Turkish.

1.4 Telugu Morphology

Word structure is studied by morphology which is a sub discipline of linguistics. While words are generally accepted as the smallest units of syntax, it is evident that in most languages, words can be related to other words by rules. E.g. // English speakers recognize that the words dog, dogs ad dog catcher are closely associated. English speakers recognize these relations by virtue of the unconscious linguistic awareness they have of the rules of word formation processes in English. Therefore, these speakers assume that dog is to dogs is just as cat is to cats, or encyclopedia is to encyclopedias. Similarly, dog is to dog catcher is as dish is to dishwasher. These rules apprehended by the speaker in each case reflect specific patterns in the way words are formed from smaller units and how those smaller units interact in speech. In this way, morphology is the branch of linguistics that studies such patterns of word formation across and within languages and attempts to explicate formal ruled reflective of the knowledge of the speakers of those languages.

1.5 Free Word Order

Word order is an important aspect found in its relative independence and becomes its own characterization of that particular language, as in English language has a fixed (Constant) order that displays specific positional difficulties or constraints for different ingredients in a sentence and precisely these positions are arranged or displayed in a syntactic set up (set of rules for the formation of a sentence). Linear order of words and phrases out of a given sentence as seen in fixed word languages
display is very significant that any attempt to modify the order will result in disturbing or disfiguring the grammatical structure of that sentence. However, some of the world languages have been found to have not attached any significance to the word order pattern. That is order or words bear very little importance to structure. The permutations of words in Sanskrit sentence are valid grammatically which stands to mean precisely the same thing. Literally a sentence is in fact not a sequence of words but a set of words.

There are umpteen number of world languages which have infinite scope for amending the order of words in a sentence without affecting its basic meaning of expression. The present day languages could be cited as example in this direction. English is a SVO (Subject Verb Object) language, where as Indian languages are SOV (Subject Object Verb) languages but with a free lancing word E.g. // The following Telugu sentences pertinently or essentially mean the same thing.

E.g.

1.రవి బడికి రాధతో వెళ్ళాడు        (Telugu sentence)
Ravi baDiki Raadhatoo VeLLaaDu        (Transliterated form)
Ravi school-to Radha with went        ( glass)

2. బడికి రాధతో రవి వెళ్ళాడు
baDiki Ravi Raadhatoo veLLaaDu
School-to Ravi Radha-with went

3. రాధతో బడికి రవి వెళ్ళాడు
Raadhatoo Ravi baDiki veLLaaDu
Radha-with Ravi school-to went
1.6 Proper Nouns in Telugu

Amongst all the Indian languages Telugu occupies a holistic position as a highly inflectional and also as an adherent language. Each and every word is structurally built on sound grammatical perception to ease out a large number of word forms. It is basically a suffixing language and these suffixes are enjoined to the right side of the word. It is also verb final in general and poised with a word free order language.

1.7 Major Problem in Telugu Comparison with English

Telugu language suffers with the line of difficulties of NER which is the significant and conspicuous absence of capitalization. Owing to its agglutinative nature, each word in Telugu is inflected for a very large number of word forms. Some are indecisively ambiguous in nature.

E.g.  

**Person Name Vs Organization Name**

**Person Name Vs Place**

**Person Name Vs Common nouns**

At times words appear in various forms like Prajaa raajyaM Paartii (For (Praja Rajyam Party)). Pi aar Pi (P.R.P), Pra raa paa, Pi Aar pi, pra raa paa, Piaarpi, Praraapa.

1.8 Motivation and Problem Definition

In Information Extraction (IE) systems, accurate detection and classification of named entities is a very important task given that named entities can help us to extract knowledge from the texts such as when and where the event has happened and who were involved. Named entities are valuable in several language engineering applications like Machine Translation, Question and Answering systems, Automatic
Text Summarization and Information Retrieval in building more accurate internet search engines. Named entity recognition for Indian languages is a challenging task. Telugu is a morphologically rich language. The concept of capitalization does not exist for Indian languages and it makes this problem difficult. So far work has been done on token level named entity recognition but many named entities in Telugu are multi word named entities. These multi-word named entities when looked as single token may not appear as a named entity but when taken collectively indicate presence of a named entity. This has been the motivation behind developing this thesis.

In this work a rule based Named Entity Recognition and Classification system has been developed for Telugu language. Our focus is mainly on identifying person, location and organization entities. The issues which make named entity recognition for Telugu a tougher task are discussed. In this thesis we describe the identification and classification of Named Entities using word level features, list lookup features and contextual information. Further classification of identified Named Entities and ambiguity resolution is done through contextual information and rules. In this Thesis we are using some transliteration schemes (please see appendix-A)

1.9 Organization of the Thesis

This thesis is organized into five chapters.

In Chapter 1, Introduction to NLP, applications of NLP and NER are presented. Need for NER, and their applications are also assessed. The characteristics of Telugu language and motivating factors for this work are dealt with in Chapter 1.

In Chapter 2, a detailed survey is carried out on various NER systems. This chapter focuses on the systems with respect to English language as also European languages, Asian languages and Indian languages. The salient features and
performance evaluations as projected by various researchers have been studied. In this chapter the survey made from various NER systems is given a thought to develop more efficient NER systems for Telugu.

We have presented a novel technique by adopting Rule-based methods for identification of Named Entities at word level using context features and gazetteers in Chapter 3.

The detailed approach using Rule-based techniques is presented in this chapter whereas in Chapter 4 Heuristic based technique is used. This chapter gives an account of the efficiency of our NER system. Recall and Precisions are taken as measures to evaluate the system. We have used some special sentence level context features and different types of filters like Morphology, Sandhi and compound words are used for identification of sentence level Named Entities.

Our final conclusion regarding the performance of the system is summarized in Chapter 5.

We also present future scope and propose enhancements to the system of our model.

In Appendix-A, Transliteration scheme, Telugu Unicode chart and Pronunciation for Telugu are shown.