Technology for Telugu

Kavi Narayana Murthy

Department of Computer and Information Sciences
University of Hyderabad, India
email: knmuh@yahoo.com

Abstract

While Telugu, like other Indian languages, has an ancient and very rich and vibrant literary history, quantitative, statistical and technological studies are a recent phenomenon. In fact very little is known about the quantitative nature of Telugu speech and script and tools and technologies have only just started appearing. There is a lot that can be and should be done to promote technology for Telugu. In this paper we describe our efforts at the Department of Computer and Information Sciences, University of Hyderabad, in developing a variety of tools, technologies and resources for Telugu. Included are brief descriptions of our Text and Speech Corpora, Dictionaries, Thesauri, morphological Analyzers and Stemmers, Spell Checkers, POS tagging and Chunking, Named Entity Recognition, Automatic Text Categorization, Referential Entity Resolution, Language Identification, Optical Character Recognition, Speech Recognition and Synthesis, etc. References are included where the reader can find more details.

1 Introduction

Telugu, like many other Indian languages, has a rich and ancient literary history but is technologically ill developed. In fact technology development in Indian languages is a recent phenomenon. Quantitative analysis of linguistic data on a large scale has been rare and in many cases we do not even know in precise enough terms basic things such as how many words are there in a language or how many word forms can be obtained from a given root. Technology development can lead to useful applications and hopefully also add life to languages which are on the decline, even if not yet considered endangered [Murthy, 2003].

Computers have no commonsense or world knowledge and therefore we think manual analysis should always give us better results. However, manual work is slow and no match at all for computers when it comes to speed. More importantly, to err is human and errors and inconsistencies can often be seen in manual work. Of late, many tools and technologies have been developed which enable computers to analyse large quantities of language data and automatically learn interesting and insightful patterns or rules

that characterize language use. When properly used, therefore, technology can be a great help, amplifying and enriching human expertise in language processing. This paper describes our ongoing efforts in developing a variety of tools and technologies for Telugu.

2 Corpus Development and Analysis

The starting point for automated technologies is language data. We need large and representative collections of appropriate language data, called corpora when made available in electronic formats suitable for mechanical processing by computers. We can design suitable tools and technologies based on the quantitative and statistical analyses of such corpora, using the insights we can thereby gain about language use. Until recently, the only corpus available for Telugu was the DoE-CIIL corpus, a plain text corpus spread over 776 files containing a total of 3,669,322 words. Frequency analysis shows that this corpus includes 636,022 types, or different word forms. Note that these are surface forms of fully inflected words, not the roots. High performance morphological analyzers are not yet available for Telugu and hence we do not as yet know the statistics at the level of root words. The type/token ratio is thus 17.33%, much higher than for languages such as English (which is about 3.42%) [Kumar, Murthy and Chaudhuri, 2007] Telugu, like other Dravidian languages, is very rich in morphology and a single root can lead to the formation of a very large number of surface word forms. Many possible word forms have not occurred even once in this corpus. We need much larger corpora. Hence our efforts in building and analyzing larger Telugu text corpora.

The Telugu corpus developed at the Language Engineering Research Centre (LERC), Department of Computer and Information Sciences, University of Hyderabad, India, hereafter referred to as LERC-UoH corpus, adds up to nearly 39 Million words, perhaps one of the largest corpora for any Indian language today. This corpus includes 3 major sub-corpora known as the CLC, NP1 and NP2 corpora. The CLC corpus includes 221 full books carefully selected by a panel of experts to include a wide variety of Telugu writings including a variety of genres, types and styles - modern and ancient, prose as well as poetry. This corpus has been checked and validated by a two-stage proof-reading process. The CLC corpus includes over 8 Million words. The NP1 corpus has been developed from the *iinaaDu* newspaper, one of the widely read newspapers in this region. *iinaaDu* runs a school of journalism of its own and its editors and sub-editors are well trained in maintaining some degree of uniformity. The NP1 corpus is about 26 Million words in size, spread across nearly 9,400 articles. This corpus was created by downloading selected articles from the on-line version of the newspaper and converting to standard ISCII [BIS, 1991] notation using tools developed by us here. The NP2 corpus was created similarly from the *aaMdhra-prabha* newspaper, another popular newspaper of the region. The NP2 corpus is smaller - about 1.3 Million words in size. All these corpora are ISCII encoded and are seen to be reasonably clean, although the NP1 and NP2 corpora have not been fully manually checked. UNICODE versions can be easily obtained since UNICODE for Indian languages has been designed with ISCII as a basis and ISCII and UNICODE have

nearly one-to-one correspondence. Together with the DoE-CIIL corpus, we thus have a nearly 39 Million word corpus for Telugu.

Type-Token growth rate analysis has been carried out separately for each sub-corpus as also for the entire corpus. It has been seen that over the entire corpus of about 39 Million tokens, 3,318,717 types have been obtained and still the curve shows no signs of bending down. We should therefore expect many more types in the language. The overall type-token ratio is 8.51 %, much larger than for Indo-Aryan languages (which averages to 5.812 %) and English (3.42 %). Repetition analysis shows that even if we keep all the 3,318,717 types in a dictionary, every tenth word or so we come across will be a new, hitherto unseen word form. We can see that about 3700 most frequent words are sufficient to give about 50% coverage of the corpus. 60% coverage can be obtained by just the first 9000 words or so. 95% coverage, however, requires 1.37 Million types, far higher than for English (For example, the most frequent 20,000 words give a coverage of about 95% on the British National Corpus of English which totals to nearly 100 Million words). 95% coverage can be obtained in Indo-Aryan languages with about 100,000 to 150,000 words. Words in Dravidian languages in general (and Telugu in particular) are an order of magnitude more complex than those in Indo-Aryan languages.

We see that Telugu words tend to be long and complex. The mean word length in terms of bytes (or 'characters') is 11.61 and the standard deviation is 4.41. In contrast, English words have a mean length of 8.18 with a standard deviation of 3.12 (based on a 3 Million word English Corpus derived from the British National Corpus (BNC) by random selection). Truly, *akshara*-s are appropriate units of writing in Indic scripts and there is really no such thing as 'alphabet' or 'character' in these scripts. There are over 20,000 akshara-s but the most frequent 5000 account for more than 95% of all words. In terms of *akshara*-s, the mean word length for Telugu is 4.99 and the standard deviation is 1.62. The global vowel-to-consonant ratio is 0.80. A preliminary analysis of sentence lengths (in words) has also been performed. The average length of a sentence in Telugu is 10.09 words, which is much smaller than the average sentence length for English, as can be expected (Average length of sentences in the British National Corpus is about 23 words). The distribution is skewed and the mode is 8.17. We have also carried out various other kinds of analyses including verification of Zipf's and Mandelbrot's laws, entropy and perplexity calculations, and bi-gram analysis at word level. See [Kumar, Murthy and Chaudhuri, 2007] for full details.

3 Dictionaries and Thesauri

We have obtained a word list of more than 3.3 Million words.These are fully inflected forms including derivations and saMdhi formations. Developing a high performance morphological analyzer is a challenge in itself and so automatically extracting root words from this list is non-trivial.

C P Brown's English-Telugu dictionary has been cast in electronic form, formatted using XML and web-enabled for on-line searching. This dictionary has about 31,000 entries. Telugu-English dictionary of C P Brown, of sim-

ilar size, has also been converted into electronic form but some more work is required to iron out formatting inconsistencies. Efforts are on to make available other published dictionaries in electronic form for easy access. An English-Telugu dictionary of equivalents, about 35,500 entries in size, suitable for machine translation has also been developed here. We also have several other monolingual and bilingual dictionaries for other Indian languages, all in electronic form.

Recently, we have compiled a comprehensive dictionary of closed class words, totalling to about 3,500 words, giving detailed morpho-syntactic tags. Efforts are currently on to develop a wide coverage and reliable electronic dictionary with detailed morpho-syntactic tagging for each word.

A dictionary and a thesaurus essentially contain the same information - words and their meanings, although they differ in terms of organization and structure. It is therefore an interesting computational problem to see if some kind of a thesaurus can be constructed automatically from available dictionaries. Thesauri are not available for many of our languages and developing thesauri by hand is not an easy task. We have shown that a kind of a thesaurus, which groups together words with similar meaning, can be automatically obtained from suitable bilingual dictionaries giving equivalents, such as those used for machine translation. For example, Telugu words which are given as equivalent to given a English word (or to equivalent English words) can be grouped together by a simple reverse indexing scheme. The thesaurus construction is fully automatic and it takes hardly one second to do this. Sample Telugu and Kannada thesauri have been developed to show this concept. See [Kumar and Murthy, 2007] and [Murthy, 2004] for more details. Development of high quality thesauri and word-nets will remain a thrust area in development of lexical resources for Indian languages.

4 Morphology and Stemming

Telugu words are long and complex. Dravidian languages such as Telugu and Kannada are morphologically among the most complex languages in the world, comparable only to languages like Finnish and Turkish. Below we only give a glimpse of the nature and complexity of Telugu morphology. See [Murthy, 1999], [Kumar and Murthy, 2007], [Murthy, 2006] for more details. The main reason for richness in morphology of Telugu (and other Dravidian languages) is, a significant part of grammar that is handled by syntax in English (and other similar languages) is handled within morphology. Phrases including several words 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), 'raaleekapooyaaDu' ((he) was unable to come), 'vaccinavaaDu' (the person (3P,sl) who came), 'raaDanukonnaavaa' (do you think he will not come?) are all single words in Telugu, written and spoken as atomic units without spaces or pauses. Verbs may include aspectual auxiliaries apart from tense and agreement. There are several types of non-finite forms too. A single verbal root can lead to the formation of hundreds of thousands of word forms. Nouns are inflected for number and case. Derivation being very productive, even more forms become possible when we consider full word forms. Thus

'vaccinavaaDiki' (to the person (3P, sl) who came) is a noun in singular, dative case derived from the verb root 'vacc' (to come). External saMdhi (that is, conflation between two or more complete word forms) and compounding add to the numbers. Naturally we will see very large number of types and the type-token ratio is therefore very high. These are not simple concatenations or juxtapositions of complete words as is the case in some languages of the world. These words are made up of several morphemes conjoined through complex morphophonemic processes. Developing wide coverage, high performance morphological analyzers and synthesizers is a challenging task. Small scale systems do exist but the performance claims made must be seen carefully as most of these systems have been developed and tested on only small scale data and often under idealized conditions.

One strategy that has been tried is to reduce detailed morphological analysis to lemmatization, where we are only interested in separating out the root from the rest of the word. The various affixes are not analyzed individually and the whole word is taken as a combination of just two components: the root and a combined affix complex. Words are grouped in paradigm classes so that inappropriate combinations can be checked. Even this simplification has not resulted in practical systems of adequate coverage and accuracy.

Stemming is therefore a promising alternative technology at this point of time. Stemming ignores the details of the complex morphophonemic processes involved and attempts to obtain the root (or the stem) by more direct and superficial manipulations of the surface string. A variety of stemming techniques have been developed. A heuristic stemmer based on the premise that *"the best place to cut a word into a root and a (combined) suffix is the one that globally maximizes the probability of the root as also that of the suffix"* has been built. Telugu is mainly a suffixing language and clustering the word initial n-grams has been used as a means of stemming. Suffix tree approach has been found to give better results. Hidden Markov Model (HMM) based external saMdhi splitter has also been developed. See [Kumar, 2007] for more details. More work is required to develop high performance morphological analyzers and stemmers for Telugu.

5 Spelling Error Detection and Correction

Indic scripts are based on phonetics and the units of writing, namely akshara-s are composed of basic sound units (phonemes) including vowels and consonants. The correspondence between written units and phonemes is quite straight forward although not exactly one to one. Thus there are no alphabets, letters or characters in Indic scripts. The term 'character' is now being applied to units of encoding in digital representations in computers and must be understood as such. There are thus no spellings at all. Nevertheless, a variety of mistakes can get into electronic documents. Both typographical mistakes (hitting the wrong key) and cognitive errors (such as confusions between aspirated and unaspirated forms) are common. We need something like a spell checker to detect and help us correct such mistakes. We shall therefore continue to use terms such as 'spelling error' and 'spell checker'.

Spelling errors can occur anywhere in the root

or in any of the several suffixes or in the saMdhi (juncture) between them. Detecting spelling errors therefore requires a comprehensive lexicon (including loan words, alternative forms, domain specific and technical terms, named entities (such as place names and names of organizations), etc.) as also a high performance morphological analyzer. If a word is not found in the lexicon and it could not be analyzed by the morphological analyzer either, it could then be taken as a spelling error. With the kinds of dictionaries and morphological analyzers currently available, there would be too many 'false alarms'.

Spell checkers usually do not attempt to correct the errors automatically. Instead they offer a list of suggestions for the user to choose from. This again requires a comprehensive dictionary and a good morphological generator.

An alternative to dictionary based spell checking is to use statistical models such as Markov Models to characterize the sequences of aksharas that appear in the words of the given language with varying degrees of probability. A hybrid model, combining large lists of inflected word forms and heuristics and statistics based spelling error detection and correction system has been developed [Murthy, 2001]. Improved versions are expected to be brought out in the near future.

6 POS Tagging and Chunking

Words have many meanings or senses and words often also belong to more than one part of speech (POS) category. A dictionary simply lists all possible POS categories for a word and does not provide any method to choose the correct tag in a given context. 'ceppu' can be a verb as well as a noun but in a given sentence, we can usually disambiguate between these two from the context. A POS tagger attempts to disambiguate POS tag ambiguities. Given a sentence and the possible tags for each of its words, a POS tagger can give the most likely tag for each of the words in the sentence. Rule based as well as statistical (for example HMM) based POS taggers have been developed for English and other positional languages. Indian languages are relatively free word order languages. Also, words are morphologically rich and POS ambiguities in the roots are often disambiguated automatically when the roots are inflected. Thus morphology plays a crucial role and POS taggers need to deal only with the small degree of ambiguities that remain after morphological analysis.

From the point of view of developing syntactic grammars and parsers, we need a good deal of detailed morpho-syntactic information for each word and broad POS categories will not suffice. Larger POS tag sets, however, lead to data sparsity problems as also difficulties in extensions and refinements. Therefore, a large, hierarchical (tree structured) and hence extensible tag set has been designed for Telugu. A dictionary of about 3500 closed class words has also been developed in this tag set and a comprehensive dictionary is also being developed. Once adequate dictionaries and morphological analyzers are in place, POS tagging would become a relatively simpler task for Telugu.

Chunking is the first step towards the development of syntactic grammars and parsing systems. Despite a good deal of theoretical work

done in Indian languages in syntax, there are hardly any substantial computational grammars or parsers. Chunkers group contiguous words into simple phrases (also called word groups). Phrases or word groups used in chunking are non-recursive (no np within vp, vp within np etc.) and can be easily captured using Finite State Grammars. Finite State Grammars and Chunking systems for Telugu are being developed. Full parsing could be taken up in the next phase of research and development. External saMdhi needs to be handled before POS tagging and chunking and this is currently being explored from a computational point of view.

7 Named Entity Recognition

Named Entity Recognition involves the identification of named entities such as person names, location names, names of organizations, monetary expressions, dates, numerical expressions etc. which are not normally found in dictionaries. The task has important significance in the Internet search engines and is an important task in many of the Language Engineering applications such as Machine Translation, Question-Answering systems, Indexing for Information Retrieval and Automatic Summarization.

There has been a considerable amount of work on NER in English. Much of the previous work on name finding is based on one of the following approaches: (1) hand-crafted or automatically acquired rules or finite state patterns (2) look up from large name lists or other specialized resources (3) data driven approaches exploiting the statistical properties of the language (statistical models). Not much work has been done in NER in Indian languages in general and Telugu in particular. Here we report our recent work on NER for Telugu [Srikanth and Murthy, 2008]. NER in Telugu (and other Indian languages) is challenging because of the absence of capitalization feature, high ambiguity between named entities and common words (most Sanskrit based names have common meanings - kamala, vinaya, jagannaatha, ravi) and higher degree of variations along several dimensions (telugudees'aM, TiDiPi, te.dee.paa., dees'aM etc., raMgaareDDi is ambiguous between person name and place name, kiraN can be masculine or feminine person name, goodrej can be person name or organization name). In this work we have used part of the LERC-UoH Telugu corpus developed by us.

Named entities are generally nouns and it is therefore useful to build a noun identifier. Nouns can be recognized by eliminating verbs, adjectives and closed class words. Verbs can be recognized by the idiosyncratic suffixes they take as also from the fact that Telugu is a verb final language. We have built a CRF based binary classifier for noun identification. Training data of 13,425 words has been developed manually by annotating each word as noun or not-noun. We use a morphological analyzer, along with a dictionary, to recognize inflected forms of words. Named entities are usually remain unrecognized after this step. Stop words (short, frequently occurring words, mainly the function words) are eliminated. Prefixes, suffixes and contextual cue words are used to recognize nouns. The CRF trained with the basic template which consists of the current word, the feature vector of the current word and the output tag of the previous word as the features, was tested on a test data of 6,223

words and an F-measure of 91.95% was obtained.

Nouns can the be checked for named entities. Several heuristics are used. For example, 'naayuDu' is a person suffix useful for recognizing person names and 'baad' is a place suffix useful for recognizing place names such as haidaraabaad and sikiMdraabaad, and cue words such as 'maMtri' and 'adhyakSuDu' trigger person name contexts . Lists (also called gazetteers) of location names, organization names etc. have been built and used. Regular Expressions have also been used for pattern matching. F-Measures of 66% to 97% have been obtained. Using this heuristic system, a manually tagged corpus of 72,157 words has been developed. Another named entity recognizer is then built using supervised machine learning (CRFs) and F-Measures between 80% and 97% have been obtained. It has also been shown that 'majority tag' concept can give even better results.

8 Automatic Text Categorization

Automatic Classification of Text Documents based on the subject matter or topic is useful in organization, indexing and searching of large document collections as in a Search Engine. Text Categorization helps in word sense disambiguation and machine translation since words tend to be used in particular senses in particular domains. Rule based approaches have not done well and most of the recent work in text categorization is based on statistical and machine learning approaches. Here the computer 'learns' which terms (words or phrases) are used frequently in which categories by analyzing a given collection of labelled training data. Thereafter, any new document can be classified based on the term occurrences using the already learned models. Machine learning approaches are fast and can be easily re-trained and adapted to new languages and domains. In fact high performance systems can be built with very little or no linguistic analysis or manual effort.

Fully automatic text categorization systems have been built for Telugu and other Indian languages and tested on news articles and other standard corpora such as the DoE-CIIL corpora. No manual or linguistic analysis is required and no dictionaries, morphological analyzers etc. are needed. A variety of techniques have been applied including nearest neighbour classifiers (based on the principle that birds of the same feather flock together - the category of a document is likely to be the same as the category of its nearest neighbours), Bayesian methods (based on a principled combination of prior knowledge and likelihood of specific terms occurring in documents of specific categories) and Support Vector Machines (SVMs) that try to maximize the separation between classes thereby reducing classification errors. Kernel based SVMs using Polynomial and Radial Basis Function kernels have also been tried. Soft margin linear SVMs have performed the best [Raghuveer and Murthy, 2007], [Murthy, 2005], [Murthy, 2006].

Not all words present in a document are useful in classifying the document. Some words, such as function words, occur very frequently in all classes of documents and so are useless. Very rarely occurring terms will also be of little use. The number of words in a language is usually quite large and it is important to de-

termine which words or phrases are most useful for classification. A simple statistical measure called 'Mutual Information' has been found to be very effective in this process. Mutual information between a term and a category quantifies what we can say about the document if we know that the given term occurs and what we can say about the possibility of the term occurring if the category is known. Fully automatic, language independent, high performance text categorization systems have been built for Telugu and other Indian languages using these ideas. An F-Measure of 80.38% has been obtained for the DoE-CIIL Telugu corpus where the six major categories are not fully disjoint. On news articles in four distinct categories, an F-Measure of 96.39% has been obtained. See [Raghuveer and Murthy, 2007] for more details.

9 Referential Entity Resolution

To understand a given coherent text, we not only need to understand the words, and structure and meaning of sentences, but also inter-sentence relations. Anaphoric references such as pronominals and anaphors are well explored in linguistics and binding theory principles (principal A, B and C) are well known. Computationally, the task is harder since we need to find out not what items can/cannot corefer but actually find what refers to what and what is the semantic relation between the two. Definite noun phrases need to be considered too. Given the sentences 'I purchased a pen yesterday. The cap was broken' we understand that 'the cap' is related to, and is a part of, 'the pen'. Referentially dependent entities are abbreviated in terms of information content and resolving these references will help us to get complete information. In fact even ellipsis is a kind of referential entity.

Very little work has been done in resolving referential entities in a computational framework in Indian languages [Murthy, Sobha and Muthukumari, 2007]. We have started some work on referential entity resolution for Telugu. A 1500 sentence manually annotated corpus has been developed where not only the actual antecedents but also all possible antecedents are indicated for the sake of initial study. Basic statistics have been obtained and it has been found that about 60% of the references can be resolved within the current to previous four sentences if we apply gender and number agreement constraints. However, gender is not always known. More detailed analysis and design of a hybrid reference resolving system is under way.

10 Language Identification

There are many languages spoken in India and there are also many scripts in which these languages are written. Given the very nature of multi-lingualism in India and the language policies we have, it is natural that multi-lingual documents are very common. Also, the correspondence between language and script is not one to one - a given language may be written in several different scripts and a given script can be used to write several languages. For example, Devanagari script is used to write Sanskrit, Hindi, Marathi, etc. Sindhi is written in Devanagari, Gujarati, as also in a Perso-Arabic Script. Given this scenario, language identification assumes an important role. Since many multi-lingual documents (such as forms) have short pieces of texts, it is important to

be able to recognize language from small text samples.

An automatic tool that can perform language identification from small text samples with high accuracy has been developed. No dictionaries or other lexical resources are required. N-grams of akshara-s that occur in word initial, medial and final positions are used as features and a Multiple Linear Regression (MLR) classifier is built from training data for pair-wise classification between 9 major Indian languages. A script grammar, depicted as a Finite State Grammar, is used to segment texts into akshara-s. Over 93% accuracy is obtained for texts as short as 5 akshara-s and over 99% accuracy is seen if the text is about 15 akshara-s. If 20 to 25 akshara-s are available, nearly 100% accuracy is possible. Note that the texts need not be complete words. The adequacy of the model and the significance of individual features have been analyzed. If the languages are somewhat distinct, bi-grams are sufficient but if the languages are very similar to each other, trigram features are required to obtain comparable accuracies. Language identification between Indo-Aryan and Dravidian languages, as also within Indo-Aryan pairs and within Dravidian pairs have been separately checked. Clear separation of these language families is seen - Dravidian languages are similar among themselves, Indo-Aryan languages are similar among themselves and these two families are distinct from one another, thereby providing quantitative evidence to these well known facts. Tamil script uses a relatively small number of units and as expected, Tamil is most easily distinguished from any other language. It has also been shown, with quantitative evidence, that akshara-s are appropriate units in Indic scripts, not characters or bytes. See [Murthy and Kumar, 2006] and [Murthy, 2006] for full details.

11 Speech Recognition and Synthesis

A variety of Telugu speech data including read speech, word lists, etc. have been recorded and analyzed. A speaker independent continuous speech recognition system has been built using the HTK toolkit. HMM models at triphone level have been built for the purpose. The same technology has also been applied to isolated word recognition in the context of spoken password recognition. A speaker recognition system has also been built and tested. Isolated word recognition systems have also been built using Dynamic Time Warping algorithm for two phase template matching with clustering and applied to bus reservation and telephone dialing applications. A variety of speech analysis and processing tools have been acquired and research and development at various levels initiated. A large vocabulary text-to-speech synthesis system has also been developed using di-phone concatenation method. See [Sreekanth and Murthy, 2005] for more details.

Telugu script has more or less one to one correspondence with the phoneme set and this makes it possible to phonetically transcribe written texts using a rule based approach. The 39 Million word LERC-UoH text corpus of Telugu has been transcribed fully at both phoneme and phone levels. Extensive statistical analysis is being carried out including word initial / medial / final distributions of phoneme / phone n-grams, coverage analysis, word length analysis etc. These studies will pave the way for developing phonetically rich / balanced but minimal

data sets for further research and development in speech technologies for Telugu.

12 Optical Character Recognition

A large number of Telugu text documents are available in printed form but not in electronic form. Making texts available in electronic forms is important as this provides greater flexibility in editing, searching, sorting, etc. as also for automatic processing by the computer. Typing in is slow, tedious and error-prone and Optical Character Recognition (OCR) provides us with an alternative technology. An OCR system takes a scanned image of a text page and recognizes the characters in the text. The output will be electronic text similar to typed in text and can be edited and processed the same way. A font independent OCR system has been developed for Telugu with a recognition accuracies ranging from 90% to 97% based on the quality of input images. In the preprocessing stage, the input images are binarized, skew if any is detected and corrected, text and graphics portions are separated and multi-column text is handled. Text lines are then detected and word are detected within lines. Words are further decomposed into connected components and each connected component is recognized by comparing with stored templates after size normalization using Fringe Distance. The recognized components are composed into akshara-s (and hence words and lines) in the post-processing stage to produce ISCII/UNICODE output text. Further improvements are on to make this more robust and reliable. The whole system is available as a library of modules for maximum flexibility and convenience in research and development. See [Negi, Murthy and Bhagvati, 2006] for details.

13 Other Technologies, Tools and Resources for Telugu

An advanced multi-lingual word processor celled *AKSHARA* has been designed, developed and released for free public use. AKSHARA handles 9 major scripts, including Telugu, and provides free fonts for each of these scripts. AKSHARA supports sending and receiving emails, creating web pages, basic statistical analysis, transliteration between scripts, etc. A spell checker is included for Telugu. Unlike other commercial softwares for Indian language processing, AKSHARA is fully compliant with ISCII, the national standard, as also Unicode.

A History-Society-Culture portal depicting a wide variety of topics related to the culture and traditions of the Telugu people has been developed.

A variety of tools for inter-conversion between different encoding schemes, tools for Roman Transliteration, dictionary formatting and web-enabling, etc. have been developed. Visit www.LanguageTechnologies.ac.in for more details.

14 Conclusions

In this paper we have described our efforts at Department of Computer and Information Sciences, University of Hyderabad, over the last many years in building a variety of resources, tools and technologies for Telugu. Over the years, some of these activities have

been supported by the Department of Information Technology, Government of India (by establishing a Resource Centre for Indian Language Technology Solutions (Telugu)) and the University Grants Commission. A large number of Masters and PhD students have made significant contributions. Currently, work is under way to develop wide coverage dictionaries, morphological analyzers, chunkers and shallow parsers and in referential entity resolution, apart from speech recognition for Telugu. Specialized search engines and other intelligent information retrieval technologies are being addressed.

The presentation here is brief and sketchy due to limitations of space but the reader will find complete details in the references cited. All these papers can be read or downloaded from the website (www.LanguageTechnologies.ac.in) too, where the reader will also find information about our work in other languages including Kannada, English and Myanmar.

Technology for Telugu is in its infancy and there is a lot more that can be and should be done, both in terms of scale and in terms of breadth and depth of analysis and application.

References

Atul Negi, Kavi Narayana Murthy, Chakravarthy Bhagvati, "Foundational Issues of Document Engineering in Indian Scripts and a Case Study in Telugu", Vivek, Vol 16, No. 2, 2006, pp 2-7

Beareau of Indian Standards,Indian Script Code for Information Interchange - {ISCII}", "IS 13194: 1991"

G Bharadwaja Kumar, Kavi Narayana Murthy, B B Chaudhuri, "Statistical Analysis of Telugu Text Corpora'' - IJDL, Vol 36, No 2, June 2007, pp - 71-99

M Santhosh Kumar and Kavi Narayana Murthy, ``Automatic Construction of Telugu Thesaurus from Available Lexical Resources'', Creation of Lexical Resources for Indian Language Computing and Processing LRIL-2007, C-DAC, Mumbai, 26-28 March 2007

M Santhosh Kumar and Kavi Narayana Murthy, ``Corpus Based Statistical Approach for Stemming Telugu'', Creation of Lexical Resources for Indian Language Computing and Processing LRIL-2007, C-DAC, Mumbai, 26-28 March 2007

Kavi Narayana Murthy, "A Network and Process Model for Morphological Analysis/Generation", ICOSAL-2 - The Second International Conference on South Asian Languages, 9-11 January 1999, Punjabi University, Patiala, India

Kavi Narayana Murthy, "Issues in the Design of a Spell Checker for Morphologically Rich Languages", 3rd International Conference on South Asian Languages, ICOSAL-3, 4th to 6th January 2001, University of Hyderabad

Kavi Narayana Murthy, "Language Engineering in a Multi-Lingual Environment", in "Semantic Web", A R D Prasad (Ed), DRTC, Indian Statistical Institute (Bangalore), 2003, pp M-1-10

Kavi Narayana Murthy, "On Automatic Construction of a Thesaurus", proceedings of ICSLT-O-COCOSDA 2004 International Conference, Vol-1,

pp 191-194, 17-19 November 2004, New Delhi

Kavi Narayana Murthy, "Automatic Categorization of Telugu News Articles", International Symposium on Linguistics, Quantification and Computation, CALTS, University of Hyderabad, 9-11 March 2005, pp 119-127

Kavi Narayana Murthy, ‘‘Natural Language Processing - An Information Access Perspective’’, Sarada Ranganathan Endowment in Library Science, Bangalore, 2006

Kavi Narayana Murthy, G Bharadwaja Kumar, "Language Identification from Small Text Samples", Journal of Quantitative Linguistics, Vol 13, No. 1, 2006 pp. 57-80

Kavi Narayana Murthy, Sobha, L and Muthukumari, B "Pronominal Resolution In Tamil using Machine Learning", in Christer Johansson (Ed), Proceedings of the First International Workshop on Anaphora Resolution (WAR-I), 28-30 September 2005, BREDT, Bergen, Norway, Cambridge Scholar Publishing, 2007, pp 39-50

K Raghuveer and Kavi Narayana Murthy, ‘‘Text Categorization in Indian Languages using Machine Learning Approaches’’, Proceedings of IICAI-2007, December 2007, Pune, pp 1864-1883

Surabhi Sreekanth, Kavi Narayana Murthy, ‘‘Text-Dependent Speaker Recognition System for Telugu’’, Osmania Papers in Linguistics, Vol. 31, 2005, pp 84-99

P Srikanth and Kavi Narayana Murthy, "Named Entity Recognition for Telugu’’, Proceedings of the IJCNLP-2008 Workshop on NERSSEAL (Named Entity Recognition in South and South East Asian Languages) 11-12 January 2008, Hyderabad, India