Experiments in Telugu Language using Language Dependent and Independent Models

[1]Swapna Narala, [2]B. Padmaja Rani, [3]K. Ramakrishna
[1]JNTU College of Engineering, Hyderabad TS, India
[2,3]Dept. of CSE, JNTU College of Engineering, Hyderabad, TS, India

Abstract
Developing better methods for finding root words is important for improving the processing of Indian languages. In this paper we discuss the language-dependent and independent approaches for Telugu, an Indian language. Automatic information processing and retrieval in local languages, is therefore becoming an urgent need in the Indian context. Moreover, since India is a multilingual country, Telugu is the third most spoken language in India and one of the fifteen most spoken languages in the world. It is the official language of the states of Telangana and Andhra Pradesh. There is an also a vast increase in Telugu languages text documents. Because of the complexity of Telugu language, we propose three methods for finding the root words for a given document. They are pseudo N-gramming is a language independent and other are vibhaktulu based stemming and suffix removal stemming ,which are language dependent models. Rule based pseudo N-gramming is a hybrid model. In order to reap the benefits of more than one type of approach, we also consider the effectiveness of the combination of both types of approaches. We focus on Telugu document retrieval.

Keywords
N-gramming, Pseudo N-gramming, Rule based pseudo N-gramming, Telugu, Vibhaktulu based stemming

I. Introduction

Researchers in Information Retrieval (IR) have demonstrated with a substantial variety of approaches to document retrieval for Indian languages. India is a rich country with 22 listed and 100 Non-listed Languages. Indian country is a multilingual with more than 700 speaking languages. However, most of the Indian languages have no script in order to recognize and write digitally. Institutes like Central Institute of Indian Languages (CIIL), Technology Development in Indian Languages (TDIL), are working under the Ministry of Science and Technology, The govt. of India is encouraging to develop the user friendly software, search engines, digital libraries, etc., in local languages. The studies on Indian local languages all over the world are on the rising side, but particularly it is less for Telugu language.

Telugu is one of the old and traditional languages of India, and it is categorized as one of the Dravidian language family unit with its own high-class script. Telugu is the authorized language of the Andhra Pradesh (AP) and Telangana states in south India. Singhal Amit et al [10] surveyed that in India the Telugu native speakers are above 50 million. It was positioned between 13 to 17 largest spoken languages all over the world. On Telugu IR most of the studies stated that the work is at the initial level [7]. Because of lagging of the resources for Telugu causes poor growth in IR and in its applications [4].Many researchers and linguists built Telugu retrieval system, but they didn't followed unique/ benchmark test collection to compare the performance of Telugu IR systems. It has been observed that most of the experiments are being conducted to analyze the behavior of the Telugu language in retrieval process. The main reason is that, the Telugu is a rich morphological language that has high word conflation [2], where the word logic disambiguation can't be, resolved easily [2]. From the web statistics, it is found that mostly used languages in India in the order of their usage are Telugu and Sanskrit. Padmaja Rani et al [8] proposed the search engine for Telugu documents retrieval, which is experimented using Syllable-N-gram model. In this approach a good transliteration system had been adopted, where the system considers the English keyword, and displays the result word in Telugu. Here authors have showed that the n-gram model with length-3 will enhance the search capacity. To get further better recall and accuracy, Kolikipogu Ramakrishna, and K Bhanuprakash et al [9] proposed the irregular key term look, for which, it uses the phrases to search in document collection. G Bharadwaja Kumar et.al [1][3] have generated almost 39 Million Telugu word text quantities, which is analyzed and developed by them. In 2006 P.Prasad et al [5-6] researched a cross verbal communication query based summarization scheme for Telugu - English language pair. In this paper, we investigate the effectiveness of language-dependent and language-independent approaches to document retrieval. Language independent, approaches that do not depend on knowledge of the language at hand. The best known example of language-independent approaches are N-gramming and Pseudo N-gramming techniques. Where language dependent approaches are vibhaktulu based stemming, suffix removal stemming and rule based pseudo N-gramming. The paper is structured as follows: Section II describe the language dependent, Independent approaches and hybrid model, section III explains Testing and results and at the last, section IV conclusions is drawn.

II. Language Dependent and Independent Approaches

Our proposed approach can be divided into two parts is shown in the Fig. 1 The first part is pre-processing and tokenizes the document in order to identify the words and the second part is, to extract valid root words from a tokenized document using language dependent and independent models.

First a text document is read line by line from corpus and each line is pre-processed by elimination of non-Telugu characters, numerals and special characters like colons, semicolons and quotes. Then a pre-processed document is tokenized and extracts the raw words. Words in Telugu text are separated by spaces and are extracted with spaces as delimiter from the document and place all raw words in Input File.

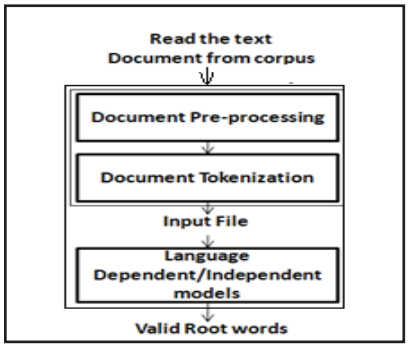

Fig. 1: Proposed Approach

Language dependent and independent model are takes raw words from Input File as input. Read one word at a time from file. Finally find the root word by applying various models like vibhaktulu based stemming, suffix removal stemming, N-gramming and Pseudo N-gramming. Then the accuracy of these models is calculated and compared.

A. Vibhaktulu Based Stemming

Dialect / Vibhaktulu based stemming is the process of finding the root word by removing the last one or more syllables / characters from the word, which are matched with Telugu vibhaktulu. It is a language dependent model. Telugu vibhaktulu are shown in Table 1.

Table 1: Telugu Vibhaktulu

Telugu	Significance	Usual Suffixes
Prathama Vibhakti (ప్రథమ విభక్తి)	Subject of sentence	డు,ము,వు,లు
Dviteeya Vibhakti (ద్వితీయ విభక్తి)	Object of action	నిస్ ,నుస్, లస్, కూర్చి, గురించి
Truteeya Vibhakti (తృతీయ విభక్తి)	Means by which action is done (Instrumental), Association, or means by which action is done (Social)	చేతస్, చేస్, తొడస్, తోస్
Chaturthi Vibhakti (చతుర్థి విభక్తి)	Object to whom action is performed, Object for whom action is performed	కొఱకుస్ , కై
Panchami Vibhakti పంచమి విభక్తి)	Motion from an animate/inanimate object	వలనస్, కంటెస్ , పట్టి
Shashthi Vibhakti (షష్ఠి విభక్తి)	Possessive	కిస్ ,కుస్, యొక్క , లోస్, లోపలస్
Saptami Vibhakti (సప్తమి విభక్తి)	Place in which, On the person of (animate) in the presence of	అందుస్, నస్

Suffix removal stemming:

Suffix removal stemming is the language dependent model. It is a process of finding the root word from the word by removing the matched suffix with suffix list.

The suffix list is shown Table 2. Maximum suffix length will be 2(two) and minimum is one.

Table 2: Suffix List

కి కు ,కె ,కై ,గా ,గాను, చే, తో,ను లు ల ,లోన, పైన ,లతో, నికి,గాని, నుండి , పై, నున్న, కంటూ, మైన , న్నాయి , డం, డు , కంటూ, స్తాయి, లకే, లకు, లే...etc

C. N-gramming

N-gramming is a language independent model. It is a sequence of characters or words extracted from a text. N-grams are divided into two categories: (1) character based and (2) word based. A character N-gram is a set of n consecutive characters extracted from a word. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram", size 3 is a "trigram" and so on. The bi-grams formed in this way from the Telugu word ప్రయోజనాలు are *ప్ర, ప్రయో,యోజ,జనా,నాలు, లు*. The word ప్రయోజనాలు with 5 letters results in 6 bi-grams. The tri-grams are *ప్రయో, ప్రయోజ, యోజనా, , జనాలు, నాలు*.

D. Pseudo N-gramming

Pseudo N-gram is the process of finding the root word by stripping the word from the end. Stripping length will be varied based on word length. Maximum stripping length is 5 and minimum is 2, and then apply Pseudo N-gram algorithm for each word, which is shown in the Fig. 2. For each step, strip the word from end and check, if it is valid root or not. If it is valid root, then extract root, accept. If it is not a valid root, decrease the stripping length by one and check. This process is repeated until stripping length not equals to zero. It is also language independent model.

```
Step 1.  Start
Step 2.  Take the LIST [1000] [12]of words as input.
Step 3.  SET i=0
Step 4.  WHILE ( LIST[ i ] != NULL ) Repeat the steps from 5 to 19
         Otherwise go to      step20
Step 5.  Read one WORD at a time from LIST i.e WORD = LIST[i]
Step 6.  Find the length of WORD as  Word_Len=strlen(WORD)
Step 7.  SET STRIPPING_LENGTH=0
Step 8.  IF (Word_Len <= 2 )  THEN go to step 9 otherwise go to step10
Step 9.  Printf WORD in UNI_Bigram_LIST goto step 19
Step 10. IF ( Word_Len  >=7 )
         THEN
               STRIPPING_LENGTH=5
         ELSE
               IF ( Word_Len < 7 OR Word_Len >=5 )
                     THEN
                           STRIPPING_LENGTH=4
                     ELSE
                           STRIPPING_LENGTH=3
         END IF
Step 11. SET Count=0
Step 12. WHILE  (STRIPPING_LENGTH=>0) Repeat steps from 13 to 18
Step 13. IF ( Count < Word_Len - STRIPPING_LENGTH- 1)
         THEN repeat steps   14 and 15
Step 14. TEMP_WORD [Count] = WORD[Count]
Step 15. SET Count= Count+1 then go to Step 13
Step 16. IF (TEMP_WORD == Valid Root WORD)          /*  Check Manually  */
         THEN Go to Step 17 otherwise Go to step 18
Step 17. Print WORD in valid root file Go to step 19
Step 18. SET STRIPPING_LENGTH= STRIPPING_LENGTH – 1 Go to Step 12
Step 19. SET i = i+1 then go to step 4
Step 20. EXIT
```

Fig. 2: Algorithm of Pseudo N-gramming

E. Hybrid Approach

This approach is combination both language independent and dependent models. Pseudo N-gram is a base method for this processing to remove suffixes from words. The result of Pseudo N-gram of some words normally contains inflections. The inflections in the stem word cannot be removed using simple Pseudo N-gram. We have designed rule based Pseudo N-gram of some possible suffixes that frequently occur in the Telugu Language which are shown in Table 3. The rules are used to replace characters.

Table 3: Rules for Rule based Pseudo N-gram

S.No	List of characters/syllable sound found as suffix	Replacement characters	List of Words are not recognized by Pseudo N-gram	List of words recognized by Rule based Pseudo N-gram
1	అ,ఆ	అం, ఉ, ఇ	చెప్పడానికి ప్రమాదాన్ని కరటాల వర్షానికి గంపెడాశ పెళ్ళయిన	చెప్పడం ప్రమాదం కరటం వర్షం గంపెడు పెళ్ళి
2	ఇ ,ఇం	ఉ, అ, అం	నిమిదించడం దివుడి హింసించి	నిమిదం దివుడు హింస
3	ఉ , ఉ + లు , ఓ +లు	ఇ, అం	ఆక్కుయుల్ని ఎడారులు సన్నజాజుల చీకట్లో	ఆక్కుతి ఎడారి సన్నజాజి చీకటి
4	ఎ ,ఏ, ఎం	ఇ ,ఉ ,అ, అం	చోటిక్కడ మురిపెంగ పనిమిలేదు ఎక్కడక్కడో	చోటు మురిపం పని ఎక్కడ
5	ఒ ,ఓ ,ఔ	ఇ ,ఉ	కాలోకటి తాడోపేడో	కాలు తాడు
6	అం	ఉ	పగలంతా కళ్ళం తా	పగలు కళ్ళు

III. Testing and Results

The experiments were performed on Telugu Corpus by language dependent and independent model. This work has been implemented on sample selection of 1,550 documents. A sequenece of words from the input file was used in extracting valid root and the results are presented. Accuracy of each model is calculated and compared. Language independent models have more accuracy than dependent models. Hybrid model has more accuracy than dependent models and independent models shown in fig. 3.

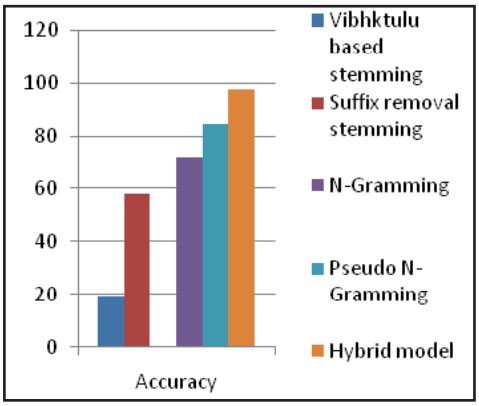

Fig. 3: Accuracy of Language Models

IV. Conclusion

Hybrid model is also well suited for different complex Indian languages like Hindi, Malayalam and Kannada. Our proposed approach will minimize inflections of words; it will become easy for retrieving desired information. In this paper, accuracy of hybrid model is more than language dependent and independent models. In my knowledge, there is no such report of Extracting Telugu language valid root words using Hybrid model. The maximum accuracy observed is 98% for hybrid model. We propose to extend it for Telugu categorization.

References

[1] Bharadwaja Kumar, G., Kavi Narayana Murthy, B. B. Chaudhuri, "Statistical analyses of Telugu text corpora", International journal of Dravidian linguistics (IJDL), Vol. 36, Issue 2, pp. 71-99, 2007.

[2] G.U.Rao,"Functional Specifications of Morphology CLATS", Hyderabad Central University, Version 1.3.1, 2008, pp. 1-32, 2008.

[3] Murthy, Kavi Narayana, G. Bharadwaja Kumar,"Language identification from small text samples", Journal of Quantitative Linguistics, Vol. 13, Issue 1, pp. 57-80, 2006.

[4] N.Murthy, P Srikanth,"Named entity recognition for Telugu", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp. 41–50, 2008.

[5] Prasad Pingali, Jagadeesh Jagarlamudi, Vasudeva Varma, "WEBKHOJ: Indian language IR from multiple character encodings", Proceedings of the 15th international conference on World Wide Web, pp. 801–809, 2006.

[6] Pingali, Prasad, Vasudeva Varma,"Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006", Working Notes of Cross Language Evaluation Forum, Vol. 2, 2006.

[7] Pingili V.V Prasad,"Recall Oriented Approaches for Improved Indian Language Information Access", Ph.D Thesis, 2009.

[8] Rani, Dr B. Padmaja, Dr A. Vinay Babu,"Novel Implementation of Search Engine for Telugu Documents with Syllable N-Gram Model", International Journal of Engineering Science and Technology, Vol. 2, Issue 8, pp. 3712-3720, 2010.

[9] Ramakrishna Kolikipogu, K.Bhanuprakash, K.Neeraja,"Telugu Items Search by key Phrase Analysis Information Retrieval in Telugu Language", IJSAA, Vol. 2, Issue ICRASE12, pp. 34- 37, 2012.

[10] Singhal, Amit,"Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering", Vol. 24, Issue 4, pp. 35–43, 2001.

Swapna Narala received B.Tech. in Computer Science and Information Technology from VREC-Nizamabad, JNT University. M.Tech. in Computer Science & Engineering from JNTU Anantapur and she is Pursuing Ph.D. in the area of Information Retrieval Systems from the Department of Computer Science and Engineering, JNT University Hyderabad. She has 12 years of teaching experience in various engineering colleges. Currently she is working as Associate Professor and Training & Placement Officer in the department of Computer Science Engineering ,Vijay Rural Engineering College, Nizamabad, India. To her credit Mrs.Swapna Narala has 12 publications in various National / International Conference and Journals. She is also a Member of Various Technical Bodies including IEEE, ISTE etc. Her area of interest includes Information Retrieval, Text Mining, Web Mining, Machine Learning, Information Security etc.

B. Padmaja Rani received B.Tech Electronics Engineering from Osmania University, M.Tech in Computer Science from JNT University Hyderabad, India and she has been awarded Ph.D. in Computer Science from JNT University, Hyderabad, India. At present she is working as Professor in the Department of Computer Science and Engineering, JNTUH College of Engineering, JNTU University Hyderabad. She is having 20 years of experience in Industry and Academia. At present she is a Professor of Computer Science and Engineering Department in JNTUH College of Engineering, JNT University, Hyderabad. Her area of Research includes Information Retrieval, Data Mining, Machine Translation, Computer Networks, Software Engineering etc. She is guiding 6 Research Scholars in the area of Information Retrieval and Computer Networks. To her credit she is having more than 60 publications in reputed International Journals and Conferences. She is a member of various advisory committees and Technical Bodies. She is also a Member of Various Technical Associations including ISTE, CSI, IEEE etc.