RULE BASED PSEUDO N-GRAM MODEL FOR TELUGU SCRIPT

N. Swapna[1], B. Padmaja Rani[2]

[1]Research Scholar, Department of CSE, JNTU College of Engineering, Hyderabad, TS
[2]Department of CSE, JNTU College of Engineering, Hyderabad, TS

Abstract

With the increasingly widespread use of computers and the internet in India, large amount of information in Indian languages are becoming available on the web. Automatic information processing and retrieval is therefore becoming an urgent need in the Indian context. This paper presents a new Rule based Pseudo N-gram for Telugu language. Rule based Pseudo N-gram is an approach, which provides a system that gives set of rules to extracting root words by removing inflections which were unrecognized by Pseudo N-gram. Pseudo N-gram can act as a preliminary stage for Rule based Pseudo N-gram. Pseudo N-gram is process of stripping the word from the end. We composed five rules to describe a Rule based Pseudo N-gram. The rules are written based on the morphology, grammar rules and word derivation structure of 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 an authorized language of the states of Telangana and Andhra Pradesh. Telugu is a rich morphological large that has high word conflation. Keeping in view of these complexities, we propose a Rule based Pseudo N-gram that provides a reasonable alternative to word based models and is also used for text categorization. We have conducted the experiments on randomly selected Telugu documents and we found the accuracy of Rule based Pseudo N-gram is up to 97.8%.

Key words: *Rule Based Pseudo N-Gram, Pseudo N-Gram, Text Categorization, Morphology, Grammar Rule.*

----------***----------

1. INTRODUCTION

Information Retrieval in local language is getting more popularity in developing countries like India. Automatic information processing and retrieval is therefore becoming an urgent need in the Indian context. Moreover, since India is a multilingual country, Telugu is the third most spoken languages in the world. It is the authorized language of the states of Telangana and Andhra Pradesh. There is an also a vast increase in Telugu language text documents with use of computers and the internet. When users need to retrieve some information from the Telugu language document depending on the query, they may get some irrelevant information (or) may lose some important information. Processing of Telugu documents is more difficult, because of the complexity when compared to European and other Indian languages. Telugu is morphologically rich language which contains too many suffixes and inflections. Building efficient information retrieval system for Telugu is a challenging task due to the richness in morphology and high word conflation feature of the language. The canonical structure is described as ((C) C) C V. The Telugu structure evolves with a set of syllables known as vowels and consonants, where consonant vowel (CV) core is the basic unit optionally preceded by one (or) two consonants. They alone form independent syllables. Fig.1 (a) shows that Telugu language has 13 vowels and 36 consonants, Fig.1 (b) shows Telugu gunintamulu. In this paper, we proposed a Rule based Pseudo N-gram for Telugu language to extract the root word, which were unrecognized by pseudo N-gram. In this approach, Pseudo N-gram can act as preliminary stage. 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.Pseudo N-gram model is an alternative to N-gram and word based language models. We composed five rules to describe a Rule based Pseudo N-gram. These rules are used to replace the characters (or) syllables, when the words are not recognized by Pseudo N-gram.

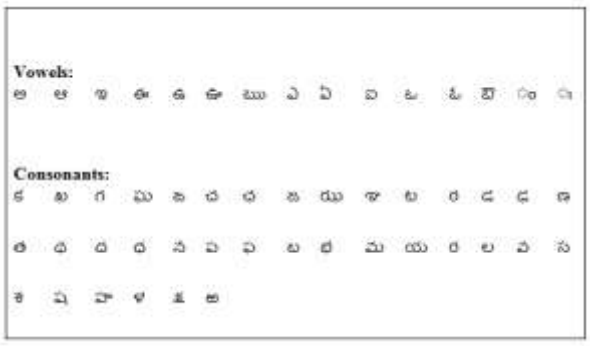

Fig 1 (a) Telugu vowels and consonants

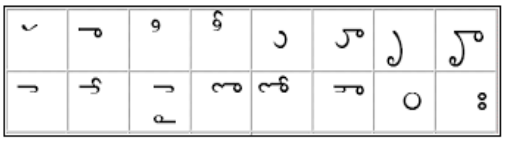

Fig 1 (b) Telugu Gunintamulu

2. RELATED WORK

Telugu is one of the old and traditional languages of India. Singhal Amit et al [4] surveyed that, in India the Telugu native speakers are above 50 millions. Many researchers and linguists built Telugu IR system. Pingali V.V Prasad et al [6] stated that, the Telugu IR work is at the initial level. Because of lagging of the resources for Telugu causes poor

growth in information retrieval and in it's applications[5]. Telugu is a rich morphological language that has high word confloation where the word logic disambiguation can't be, resolved easily[7]. Padmaja Rani et al [9] proposed the search engine for Telugu document retrieval, which is experimented using syllable N-gram model. In this approach a good transliteration system had been adapted and they have showed that the N.-gram model with length-3 will enhance the search capacity. Vishnu Vardhan et al [2] proposed the overlapping trigram technique for Telugu text categorization. A Durga k et al [8] projected a technique with ontology text categorization for Telugu digital-items and retrieval system. Sunitha, kalyani et al [3] suggested a approach to improve the performance of Telugu rule based Morphological analyzer. Pseudo N-gram is new method that reduces words with the same root to a common form, usually by removing derivational and inflectional suffixes from each word [1].It is an alternative to stemming, N-graming and other word based models. For the best of our knowledge , Pseudo N-gram is the introductory approach has been hired on extracting Telugu root words.

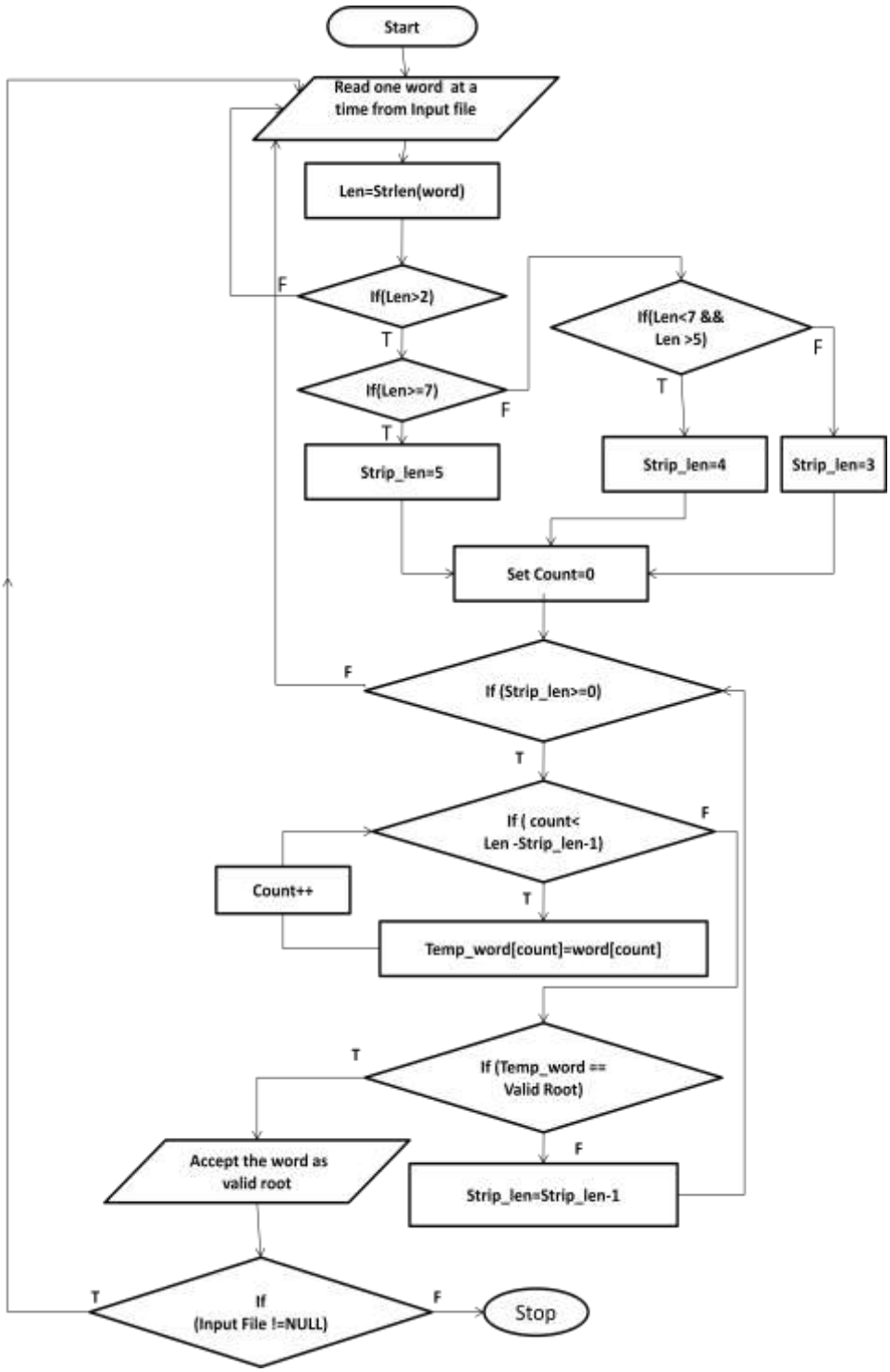

Fig.2 Flowchart of Pseudo N-gram

3. PSEUDO N-GRAM

Pseudo N-gram is a procedure that reduces words by stripping derivational and inflectional suffixes from each word to get valid root[1]. It takes raw words from *Input File* as input. Read one word at a time from file. First find the length of word and fix the stripping length. 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. 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 and perform categorization is shown in fig.3 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. Fig.2 shows the flowchart of Pseudo N-gram.

4. PROPOSED SYSTEM

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 algorithm, which is shown in Fig .3 for some possible suffixes that frequently occur in the Telugu Language. The rules are used to replace characters. These rules are shown in Table 1.

4.1 Rules for Rule based Pseudo N-gram

Table 1. Rules for Rule based Pseudo N-gram

S.No	List of characters/syllable sound found as suffix	Replacement characters	After Pseudo N-gram	After Rule based Pseudo N-gram
1	అ,ఆ	అం, ఉ, ఇ	చె ప్పడానికి ప్రమాదాన్ని కెరటాల వర్షానికి గంపెడాశ పెళ్ళయిన ఎగరాలంటే	చెప్పడం ప్రమాదం కెరటం వర్షం గంపెడు పెళ్ళి ఎగురు
2	ఇ ,ఇం	ఉ, అ, అం	నిషేదించడం దేవుడి హింసించే	నిషేదం దేవుడు హింస
3	ఉ , ఉ + లు , ఓ +లు	ఇ, అం	ఆకృతుల్ని ఎడారులు సన్నజాజుల చీకట్లో	ఆకృతి ఎడారి సన్నజాజి చీకటి
4	ఎ ,ఏ, ఎం	ఇ ,ఉ ,అ, అం	చోటెక్కడ మురిపెంగ పనేమిలేదు ఎక్కడెక్కడో	చోటు మురిపం పని ఎక్కడ
5	అం	ఉ	పగలంతా క ళ్ళం తా	పగలు కళ్ళు

4.2 Rule based Pseudo N-gram Algorithm

```
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 step12
Step 9.   IF (TEMP_WORD == Valid Root WORD)  /*  Check Manually  */
          THEN Go to Step 10 otherwise Go to step 11
Step 10.  Printf WORD in UNI_Bigram_LIST goto step 20
Step 11.  Check the rules for last syllable of a word with the replacement
          character using  rule based Pseudo N-gram and replace the characters or
          syllables then Go to Step 9.
Step 12.  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 13.  SET Count=0
Step 14.  WHILE  (STRIPPING_LENGTH=>0) Repeat steps from 15 to 21
Step 15.  IF ( Count < Word_Len - STRIPPING_LENGTH- 1)
          THEN repeat steps  16 and 17
Step 16.  TEMP_WORD [Count] = WORD[Count]
Step 17.  SET Count= Count+1 then go to Step 15
Step 18.  IF (TEMP_WORD == Valid Root WORD)  /*  Check Manually  */
          THEN Go to Step 19 otherwise Go to step 20
Step 19.  Print WORD in valid root file Go to step 21
Step 20.  Check the rules for last syllable of a word with the replacement
          character using  rule based Pseudo N-gram and replace the characters or
          syllables then Go to Step 16.
Step 21.  SET STRIPPING_LENGTH=STRIPPING_LENGTH – 1 Goto Step 14
Step 22.  SET i = i+1 then go to step 4
Step 23.  EXIT
```

Fig.2 Algorithm of Rule based Pseudo N-gram

After Pseudo N-gram, if the word ends with " ది, డి, లోనుడె " or has inflection " ఇంఇఅండఎ" are replaced by the character " దం డు,లు తి, డ " with inflection "అం ఉ, ఉ,ఇ, అ ". We use above rules to get the root for inflected word.

a) List of all the possible words which are not identified as root word using pseudo N-Gram.
b) Apply Pseudo N-gram and check the rules for inflected words to be replaced by another character.
c) The replacements characters to be made after removal of suffix character so that valid root can be formed.

5. EXPERIMENTAL RESULTS

The experiments were conducted on Telugu Corpus, collected from online newspapers and wikipedia.This work has been implemented on sample selection of 1,550 documents. A sequenece of words which are not recognized by pseudo N-gram word list was used in extracting valid root by Rule based pseudo N-gram algorithm and the results are presented in Table 2,which contains list of words after pseudo N-gram and Rule based pseudo N-gram with suitable rules for evaluation.

Table 2. Results of Rule based Pseudo N-gram algorithm

List of words before pseudo N-gram	List of words, which are not recognized as valid root after pseudo N-gram	Rule for replacement character	Valid root words after Rule based pseudo N-gram
పర్వతానికి	పర్వతా	Rule 1	పర్వ తం
ప్రయత్నాలు	ప్రయత్నా	Rule 1	ప్రయత్నం
దంతాలతో	దంతాల	Rule 1	దంతాలు
తీరాక	తీరా	Rule 1	తీరు
వాడిగోళ్ళతో	వాడిగోళ్ళ	Rule 1	వాడిగోళ్ళు
శోకాలు	శోకా	Rule 1	శోకం
కారేట్లుగా	కారే	Rule 4	కారు
విజృంభించి	విజృంభిం	Rule 2	విజృంభం
ఫలవృక్షాలతో	ఫలవృక్షా	Rule 1	ఫలవృక్షం
బయులుదేరింది	బయులుదేరిం	Rule 2	బయులుదేరు
జలాశ్రయాలకు	జలాశ్రయా	Rule 1	జలాశ్రయం
శరీరమంతా	శరీర	Rule 1	శరీరం
ఫీంకారానికి	ఫీంకారా	Rule 1	ఫీంకారం
రక్కేది	రక్కే	Rule 4	రక్కు

6. CONCLUSION

Rule based pseudo N-gram technique 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 Rule based pseudo N-gram is more than pseudo N-gram model. In my knowledge, there is no such report of Extracting Telugu language valid root words using pseudo N-gram and Rule based pseudo N-gram. The maximum accuracy observed is 98% for Rule based pseudo N-gram. As part of our research work in Telugu categorization, we propose to extend it for recognize all words as valid roots.

REFERENCES

[1] Porter M.F. " An algorithm for suffix stripping",pogram,14(3),1980,130-137

[2] B.vishnu varthan, L Pratap Reddy,A Vinay babu "A Model for overlapping trigram Technique for telugu script", Journal of Theoretical and applied information Technology, (JAJIT 2007).

[3] Kalyani, N. and Sunitha, K.V.N. (2009): A Novel approach to Improve rule based Telugu Morphological Analyzer, World Congress on Nature & Biologically Inspired Computing (NaBIC 2009).

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

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

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

[7] G.U.Rao, 2008, Functional Specifications of Morphology CLATS, Hyderabad Central University, Version 1.3.1, 2008, PP:1-32.

[8] Mrs.A.Kanaka Durga, Dr.A.Govardhan, Ontology Based Text Categorization Telugu Documents, International Journal of Scientific & Engineering ResearchVolume 2, Issue 9, September- 2011 , PP: 1-4

[9] Rani, Dr B. Padmaja, and 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, 2010, PP:3712-3720.

BIOGRAPHIES

N. Swapna 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 10 ublications 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.