International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 10, October 2016)

Analysis of Pseudo N-Gram Model on Telugu Document Classification

N. Swapna[1], B. Padmaja Rani[2]
[1]Research scholar, JNTU College of Engineering, Hyderabad TS, India
[2]Department of CSE, JNTU College of Engineering, Hyderabad, TS, India

***Abstract*-- Pseudo N-gram is the process of finding the root word by stripping the word from the end. Pseudo N-gram model is an alternative to N-gram and word based language models. In this paper we propose a model applicable to categorization of Telugu documents. India is a multilingual country. Telugu is the third most spoken language in India and one of the fifteen most spoken language n the world. It is the official language of the states of Telangana and Andhra Pradesh. Telugu is agglomerative and the canonical structure is described as ((C) C) CV. The structure evolves with a set of syllables known as consonants and vowels. Where vowel consonant (V C) core is the basic unit optionally preceded by one (or) two consonants. Words formed from this set evolved into a large corpus. Every word consists of one (or) several morphemes. Keeping in view of these complexities we propose a pseudo N-gram that provides an alternative to a N-gram and word based language models in achieving document categorization for the Telugu script (or) language.**

***Keywords*-- Morpheme, N-gram, Pseudo N-gram, Syllable, Telugu, Text categorization.**

I. INTRODUCTION

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. It is the authorized language of the Telangana and Andhra Pradesh states in south India. Amit et al [3] 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. Telugu is a rich morphological language that has high word conflation [4]. Fig.1 shows that Telugu language has 13 vowels and 36 consonants. Padmaja Rani et al [8] 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. Goverdhan. A Durga k et al [6] projected a technique with ontology text categorization for Telugu digital-items and retrieval system. G. Bhradwaja kumar et .al [5][7] have generated almost 39 million

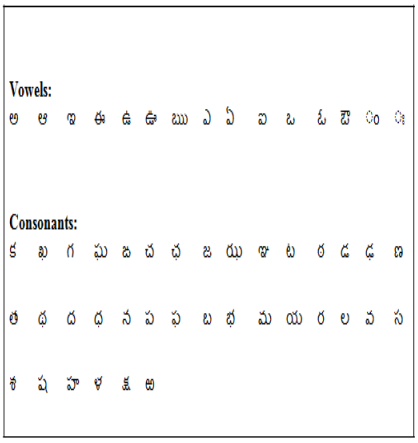

Figure I. Telugu vowels and consonants.

Telugu word text quantities, which is analyzed and developed by them. Various approaches for extraction of root words have been reported foe European and Indian languages like Assam by Utpal Sharma [2]. For the best of our knowledge , this is the first time Pseudo N-gram approach has been employed on extracting Telugu root words. Our methodology for Telugu text categorization by extracting valid root word by stripping the word from end from a given text corpus is similar to technique used for preparing words based on Rule based Telugu morphological analyzer as in [9]. Stripping length will be varied based on word length. Maximum stripping length is 5 and minimum is 2. The paper is structured as follows ; section 2 describe the system overview , section 3 explains Testing and results and at the last, section 4 conclusions is drawn.

II. SYSTEM OVERVIEW

The system design of the proposed approach can be divided into two parts is shown in the Fig.2 The first part is pre-processing and tokenizes the document in order to identify the words and the second part is the pseudo N-gram and categorization.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 10, October 2016)

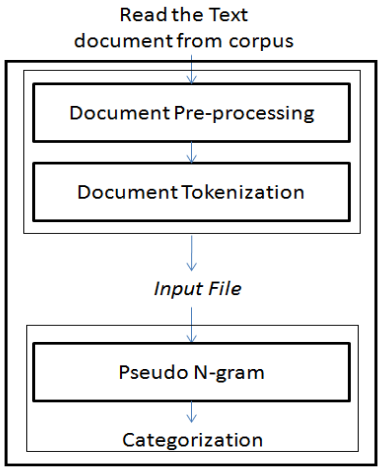

Figure II. System Design

A. Pre-Processing and tokenization

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*.

Example 1:

Telugu document: ఒక వనంలో గజేంద్రం2 అనే మదపుటేనుగు తన భార్యలైన దశలక్షకోటి ఆడ ఏనుగులతో విహరిస్తూ దాహం తీర్చుకోడానికి మహావనం మధ్యనున్న సరోవరానికి బయలుదేరింది. దాహం తీరాక జలవిహారంపై బుద్ధిపుట్టి, తన ఆడ ఏనుగులతో సరోవరాన్ని అల్లకల్లోలం చేస్తూ విజృంభించి జలకేళిలో లీనమై ఉన్న గజేంద్రాన్ని, గొప్ప మొసలి కోరలు గుచ్చి ముందరి కాలొకటి పట్టుకొన్నది.

Pre-processed Telugu document is

ఒక వనంలో గజేంద్రం అనే మదపుటేనుగు తన భార్యలైన దశలక్షకోటి ఆడ ఏనుగులతో విహరిస్తూ దాహం తీర్చుకోడానికి మహావనం మధ్యనున్న సరోవరానికి బయలుదేరింది దాహం తీరాక జలవిహారంపై బుద్ధిపుట్టి తన ఆడ ఏనుగులతో సరోవరాన్ని అల్లకల్లోలం చేస్తూ విజృంభించి జలకేళిలో లీనమై ఉన్న గజేంద్రాన్ని గొప్ప మొసలి కోరలు గుచ్చి ముందరి కాలొకటి పట్టుకొన్నది

Tokenized Document is

ఒక వనంలో గజేంద్రం అనే మదపుటేనుగు తన భార్యలైన

దశలక్షకోటి ఆడ ఏనుగులతో విహరిస్తూ దాహం

తీర్చుకోడానికి మహావనం మధ్యనున్న సరోవరానికి

బయలుదేరింది దాహం తీరాక జలవిహారం పై

బుద్ధిపుట్టి తన ఆడ ఏనుగులతో సరోవరాన్ని

అల్లకల్లోలం చేస్తూ విజృంభించి జలకేళిలో లీనమై

ఉన్న గజేంద్రాన్ని గొప్ప మొసలి కోరలు గుచ్చి

ముందరి కాలొకటి పట్టుకొన్నది

All these tokens are stored in *Input File.*

B. Pseudo N-gram and categorization

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.

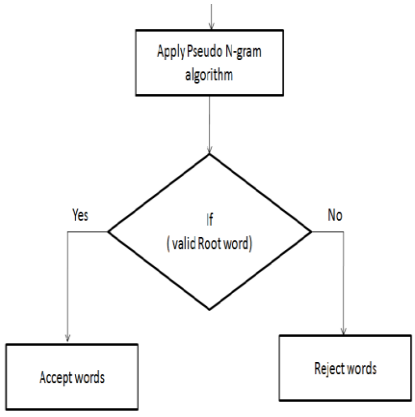

Figure III. Pseudo N-gram

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 10, October 2016)

Document categorization using words is complex task due to the nature of the corpus. To reduce the computational complexity it is necessary to adopt a corpus database with base words. Pseudo N-gram based categorization of words is one of the solution to achievethe above gole. The proposed algorithm is as follows.

Pseudo N-gram Algorithm

1. Start
2. Take the LIST [1000] [12]of words as input.
3. SET i=0
4. WHILE (LIST[i] != NULL) Repeat the steps from 5 to 18
   Otherwise go to step 19
5. Read one WORD at a time from LIST i.e WORD = LIST[i]
6. Find the length of WORD as Word_Len=strlen(WORD)
7. SET STRIPPING_LENGTH=0
8. IF (Word_Len > 2) THEN go to step 9 otherwise go to step18
9. 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
10. SET Count=0
11. WHILE (STRIPPING_LENGTH=>0) Repeat steps from 12 to 17
12. IF(Count< Word_Len- STRIPPING_LENGTH- 1)
    THEN repeat steps 13 and 14
13. TEMP_WORD [Count] = WORD[Count]
14. SET Count= Count+1 then go to Step 12
15. IF (TEMP_WORD == Valid Root WORD) /*Validate with Telugu dictionary */
    THEN Go to Step 16 otherwise Go to step 17
16. WRITE WORD in Accepted File Go to step 18
17. SET STRIPPING_LENGTH= STRIPPING_LENGTH – 1 Go to Step 11
18. SET i = i+1 then go to step 4
19. EXIT

III. Testing And 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 from the word lists was used in extracting valid root by pseudo N-gram algorithm and the results are presented in Table 1,which contains list of words with intial stripping and final root word for evaluation.

Example 2: List[20][10]={ పాలసముద్రంలో, ఏనుగులతో, మదపుటేనుగు, భార్యలైన, జలవిహారంపై, సరోవరానికి, నిలబడ్డాయిగాని, తోచలేదు, తప్పించాలో, అల్లకల్లోలం, బుద్ధిపుట్టి, చెల్లాచెదరుగా }

TABLE 1.
Results of Pseudo N-gram algorithm

List of Words before pseudo N-gram	Intial Word length	Intial Stripping Length	Final stripping length to make a Valid Word	Stripped Suffix	Valid Root word
పాలసముద్రంలో	6	4	1	లో	పాలసముద్రం
ఏనుగులతో	5	4	2	లతో	ఏనుగు
మదపుటేనుగు	6	4	0	---	మదపుటేనుగు
భార్యలైన	4	3	2	లైన	భార్య
జలవిహారంపై	6	4	1	పై	జలవిహారం
సరోవరానికి	6	4	0	----	Not a Valid Root
నిలబడ్డాయిగాని	7	5	2	గాని	నిలబడ్డాయి
తోచలేదు	4	3	0	---	తోచలేదు
తప్పించాలో	4	3	1	లో	తప్పించా
అల్లకల్లోలం	5	4	0	---	అల్లకల్లోలం
బుద్ధిపుట్టి	4	3	2	పుట్టి	బుద్ధి
చెల్లాచెదరుగా	6	4	1	గా	చెల్లాచెదరు

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 10, October 2016)

When the words list was given to pseudo N-gram there were some set of words that could not be recognized as valid root. In the above table, the word సరోవరానికి is not a valid root. After applying the pseudo N-gram algorithm, the possible words are సరో, సరోవ, సరోవరా, సరోవరాని, and సరోవరానికి for each stripping వరానికి, రానికి, నికి, and కి with the stripping length like 4,3,2,1 and 0. Valid root for సరోవరానికి is సరోవరం. In proposed algorithm extraction of valid root using pseudo N-gram based categorization are addressed. Disadvantage of pseudo N-gram is, it generates some words that are not valid.

IV. CONCLUSION

Extraction of valid root is already proven method of Information Retrieval in Indian languages. Pseudo N-gram technique is a new language independent and they are well suited for different complex Indian languages like Hindi and Kannada. In this paper, accuracy of pseudo N-gram is more than N-gram model. The maximum accuracy observed is 85% for pseudo N-gram. There is no such report of Extracting Telugu language valid root words using pseudo N-gram. As part of our research work in Telugu categorization, we propose to extend it for recognize all words as valid roots.

Acknowledgements

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 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.

REFERENCES

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

[2] Uppal sharma," unsupervised learning of morphology of a highly inflectional language" Ph.D thesis submitted to Department of Computer Science and Information Technology",Tezpur University, Napaam,Assam,India,2006.

[3] A D Manning, P. Raghavan, and H. Scutze., An introduction to information retrieval, Cambridge: Cambridge university press, Vol. 1, 2009, PP:6.

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

[5] Bharadwaja Kumar, G., Kavi Narayana Murthy, and B. B. Chaudhuri. , Statistical analyses of Telugu text corpora, International journal of Dravidian linguistics (IJDL), vol-36, issue- 2, 2007, PP:71-99.

[6] 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

[7] Murthy, Kavi Narayana, and G. Bharadwaja Kumar., Language identification from small text samples, Journal of Quantitative Linguistics, vol-13, issue-1, 2006, PP:57-80.

[8] 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.

[9] 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).