Global Journal of Computer Science and Technology: H Information & Technology
Volume 16 Issue 4 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Telugu Text Categorization using Language Models

By Swapna Narala, B. Padmaja Rani & K. Ramakrishna
JNTU College of Engineering

Abstract- Document categorization has become an emerging technique in the field of research due to the abundance of documents available in digital form. In this paper we propose language dependent and independent models applicable to categorization of Telugu documents. India is a multilingual country; a provision is made for each of the Indian states to choose their own authorized language for communicating at the state level for legitimate purpose. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. Hence, the Classification of text documents based on languages is crucial. 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. A variant of k-nearest neighbors algorithm used for categorization process. The results obtained by the Comparisons of language dependent and independent models.

Keywords: *text categorization, language dependent and independent models, k-nearest neighbors.*

GJCST-H Classification: *D.2.11,D.2.12*

Strictly as per the compliance and regulations of:

Telugu Text Categorization using Language Models

Swapna Narala [α], B. Padmaja Rani [σ] & K. Ramakrishna [ρ]

Abstract- Document categorization has become an emerging technique in the field of research due to the abundance of documents available in digital form. In this paper we propose language dependent and independent models applicable to categorization of Telugu documents. India is a multilingual country; a provision is made for each of the Indian states to choose their own authorized language for communicating at the state level for legitimate purpose. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. Hence, the Classification of text documents based on languages is crucial. 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. A variant of k-nearest neighbors algorithm used for categorization process. The results obtained by the Comparisons of language dependent and independent models.

Keywords: text categorization, language dependent and independent models, k-nearest neighbors.

I. Introduction

Now a day's huge amount of information is being posted on to the web. In order to get useful information from the web, the information available has to be categorized. Text Categorization is the task of automatically categorizing a set of unlabeled text documents to their corresponding categories from a predefined category set [2]. These categories can be viewed as a set of documents and test document can be treated as a query to the system. The measures to evaluate the information retrieval systems are often applicable to measure effectiveness text categorization systems [1]. Text categorization has many applications [2], like information retrieval system, search engine, text filtering, word sense disambiguation, language identification, POS tagging and machine translation etc.

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 [6] surveyed that in India the Telugu native speakers are above 50 million. It was positioned between13 to 17 largest spoken languages all over the world. Telugu is a rich morphological language that has high word conflation [7]. Various approaches for text categorization have been done on Indian languages. Most of the works have been reported on Telugu language. M Narayana Swamy et al have used KNN, NB and decision tree classifier [4]. They have experiment on Kannada, Tamil and Telugu corpus statistics is illustrated by Zipf's law. Analysis of N-gram model on text classification was proposed in the work of [5]. Goverdhan. A Durga k et al [3] projected a technique with ontology text categorization for Telugu digital-items and retrieval system. For the best of our knowledge, this is the first time our proposed language models have been applied for Telugu text categorization. The paper is structured as follows; section 2 describes the system overview, section 3 explains Testing and results and at the last, a section 4 conclusion is drawn.

II. System Overview

The system design of the proposed approach can be shown in the Figure.1. First read a text document 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*. Language dependent and independent models 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, Rule based suffix removal stemming, N-gramming, pseudo N-gramming and Rule based Pseudo N-gramming. Finally, apply the text categorization.

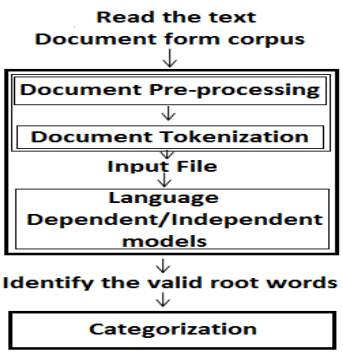

Figure 1: Proposed Approach

Author α: Research scholar, JNTU College of Engineering, Hyderabad TS, India. e-mail: swapnanaralas@gmail.com
Author σ: Department of CSE, JNTU College of Engineering, Hyderabad, TS, India. e-mail: padmaja_jntuh@yahoo.co.in
Author ρ: Department of CSE, CMR College of Engineering, Hyderabad, TS, India. e-mail: krkrishna.cse@gmail.com

a) Proposed language models

Our proposed language models are categorized in three ways are shown in figure 2. These models take raw words from *Input File* as input and identify the root word.

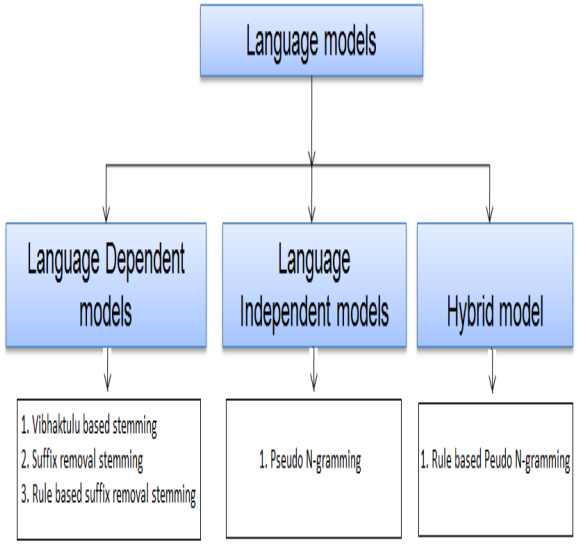

Figure 2: Proposed language models

b) Vibhaktulu based stemming

Vibhaktulu based stemming is a language dependent model. It is the process of finding the root word by removing the last one or more syllables from the word, which are matched with Telugu vibhaktulu. It is observed that, processing the complete set of input words, only 19 to 20% of words with the last syllables are matched to Telugu vibhaktulu.

c) Suffix removal stemming

Suffix removal stemming is the process of finding the root word from the word by removing the matched suffix with suffix list which is shown in figure 3. By observing the Telugu data set, it is found that maximum suffix length will be 2(two) and minimum is one. Suffix removal stemming method giving better performance than vibhaktulu based stemming algorithm. It's accuracy is 58-59%.

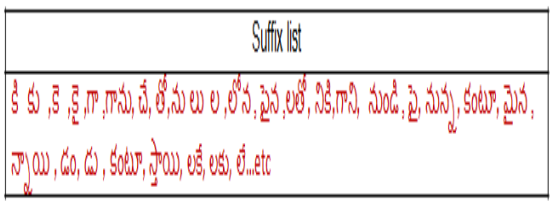

Figure 3: Suffix list

d) Rule based suffix removal Stemming

Suffix removal stemming is a base method for Rule based Suffix removal stemming algorithm. The result of suffix removal stemming words may normally contain inflections. The inflections in the stem word cannot be removed using simple suffix removal. We have designed rule based suffix removal of some possible inflections that frequently occur in the Telugu Language. The rules are used to replace characters are presented in Table 1. By these rules the electiveness of the proposed Rule based Suffix removal stemming algorithm is increased. Accuracy of Rule based suffix removal is 69-70%.

Table 1: Rules for Replacement Syllables

S.No	List of characters/syllable sound found as suffix	Replacement characters
1	అ,ఆ	అం, ఉ, ఇ
2	ఇ ,ఐం	ఉ, అ, అం
3	ఉ , ఉ + లు , ఓ +లు	ఇ, అం
4	ఎ ,ఏ, ఎం	ఇ ,ఉ ,అ, అం
5	ఒ ఓ ఔ	ఇ ,ఉ
6	అం	ఉ

e) 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 taken depending on the word length. Maximum stripping length is 5 and minimum is 2. Example of Pseudo N-gramming is shown in figure 4. It is a language independent.

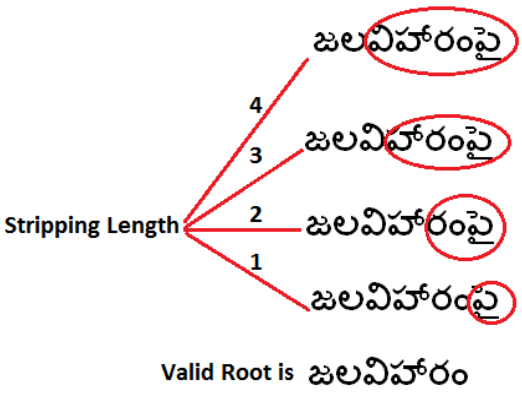

Figure 4: Pseudo N-gramming

A sequenece of words from the *Input File* was used in identifying the valid root by pseudo N-gram algorithm and the results are presented in Table 2,which contains list of words with intial & final stripping length and final valid root word.

Table 2: Result of Pseudo N-gramming

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

f) Rule Based Pseudo N-Gramming

It is a hybrid model. 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 which contain set of rules used to replace characters. These rules used for words normally contain more inflections that frequently occur in the Telugu Language. List of rules with sample example are shown in Table 3.

Table 3: List of rules for Rule based pseudo N-gramming

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 కెరటాల వర్షానికి గంపెడాశ పెళ్ళయిన	చెప్పడం ప్రమాదం కెరటం వర్షం గంపెడు పెళ్ళి
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	చోటక్కడ మురిపెంగ పనిమిలదు ఎక్కడక్కడో	చోటు మురిపం పని ఎక్కడ
5	ఒ ఓ ఔ	ఇ ,ఉ	కాలోకటి తాడోపెడో	కాలు తాడు
6	అం	ఉ	పగలంతా కళ్ళం తా	పగలు కళ్ళు

g) K-NN Classifier

The k-NN classifier is a similarity-based learning method that has been shown to be very effective for a variety of problem domains including text categorization [9, 10]. Given a test document, the k-NN method finds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category. The similarity score of each and every neighbor document to the test document is used as the weight of the classes of the neighbor document.

III. Testing and Results

The proposed models are evaluated on Telugu Corpus, collected from online newspapers and Wikipedia. This work has been implemented on sample selection of 1,500 documents of seven categories are presented in Table 4.

Table 4: Categories of Telugu Documents

	క్రీడలు	పాటలు	కథలు	సాహిత్యము	వార్తలు	రాజకీయాలు	నడులు
No. of. Doc	244	110	120	247	100	268	80
Words	44,233	10,640	88,427	26,255	87,552	99,964	27,061

To evaluating the performance of the proposed system using KNN classification, we use the typical evaluation metrics that come from information retrieval – precision (P), recall (R), and F1 measure:

$$P = TP /(TP + FP) \quad \text{.....(1)}$$

$$R = TP/(TP + FN) \quad \text{..... (2)}$$

$$F1 = (2*P*R) /(P + R) \quad \text{.....(3)}$$

Where TP is True Positives, TN is True Negatives, FN is False Negatives and FP is False Positive [8]. We have projected the performance of the proposed language models result with KNN classifier shown in Table 5.

Table 5: Performance of Language models

Language Model	Recall	Precision	F1 Measure
Vibhaktulu based stemming	75.22	86.4	81.94
Suffix removal stemming	76.42	87.5	81.58
Rule based suffix removal stemming	77.11	78.01	77.56
N-Gramming	78.02	84.30	81.03
Pseudo N-Gramming	79.96	82.77	81.34
Rule based Pseudo N-Gramming	82.81	87.77	85.22

Both Recall and Precision are to be high for an efficient performance. F1 measure reflects the overall accuracy. Recall and Precision graph is shown in figure 5(a) and 5(b). From the results we observed that, Rule based Pseudo N-gram model has high precision and recall. So it is efficient model for text categorization.

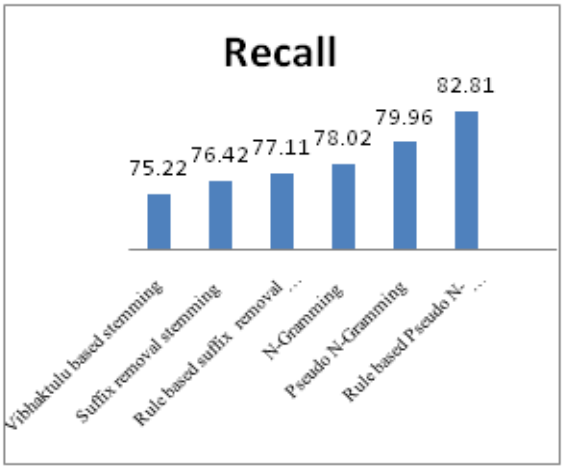

Figure 5(a): Recall Graph

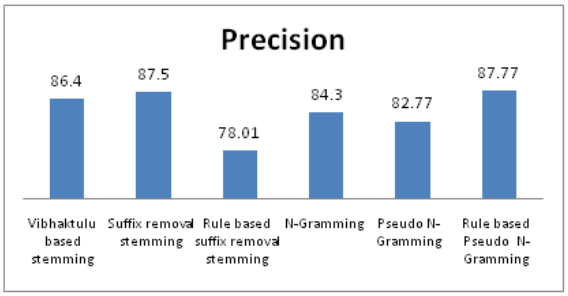

Figure 5 (b): Precision Graph

IV. Conclusion

In this paper, we proposed various language dependent and independent models. Among these models the performance of Rule based pseudo N-gramming is more. So it is well suited for Telugu Text categorization. As part of our research work in Telugu categorization, it is also suitable for other complex Indian languages like Hindi, Malayalam and Kannada.

References Références Referencias

1. T. K. Landauer, P. W. Foltz, D. Laham, "An Introduction to Latent Semantic Analysis", Discourse Processes, 1998, pp. 259-284.
2. Murthy K.N., Automatic Categorization of Telugu News Articles, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad,2003.
3. Mrs. A. Kanaka Durga, Dr. A. Govardhan, Ontology Based Text Categorization Telugu Documents, International Journal of Scientific & Engineering Research Volume 2, Issue 9, September- 2011 , PP: 1-4
4. Indian Language Text Representation and Categorization Using Supervised Learning Algorithm M Narayana Swamy1 ,M. Hanumanthappa
5. Vishnu Vardhan B.Analysis of N-gram model on Telugu Document classification thesis ,2008.

6. A D Manning, P. Raghavan, and H. Scutze., An introduction to information retrieval, Cambridge: Cambridge university press, Vol. 1, 2009, PP:6.
7. U.Rao, 2008, Functional Specifications of Morphology CLATS, Hyderabad Central University, Version 1.3.1, 2008, PP:1-32.
8. Joachims, Thorsten: Learning to classify text using support vector machines: Meth- ods, theory and algorithms. Kluwer Academic Publishers, (2002).
9. Yang, Y. and Liu, X. (1999). A Re-examination of Text Categorization Methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, pages 42-49.
10. Mitchell, T.M. (1996). Machine Learning. McGraw Hill, New York, NY.

This page is intentionally left blank