CDetector-A Vector Space Document Model Based Context Identification Approach for Telugu Documents

Rajeshkumar S Gone[1], Jatinderkumar R Saini[2], Kande Srinivas[3]
[1]*Department of CSE, SVS Group of Institutions, Warangal, India*
[2]*Narmada College of Computer Applications, Bharuch, India*
[3]*Department of CSE, SVS Group of Institutions, Warangal, India*
(E-mail: rajeshgone2007@gmail.com, saini_expert@yahoo.com, sri.kande3679@gmail.com)

***Abstract*—**The ongoing data on the web changes progressively and surges rapidly, which makes extensive trouble in getting to intriguing data. The most effective method to mine hot events, how to break down the relationship of context, how to identify the context of selected document data and how to sort out data basically are testing errands. To address these issues we are proposing a tool called CDetector-A Vector Space Document Model Based Context Identification of Telugu Language Documents. We partitioned the procedure into three stages, in the first stage, we group a lot of Telugu information archives into a large size Telugu corpus (approximately 18,985,337 words available in the corpus), later this corpus is transformed into a well-formatted stop words list. On other hands, we group large amount Telugu data documents into vector space document model information of unique words with its occurrence frequency. We consider this as fine-grained information, on which we apply filtering, retrieving and sorting techniques. In the second stage, we expel Telugu stop words from information and include position parameter alongside frequency, these parameters enabled us to compute coordinating matching score in various domains. In the third stage, we were coordinating these words in various domains, and after that, we utilized position and occurrence parameters to figure out coordinating matching score of given document in particular space domain. At the end, as a prediction, the maximum matching score of the particular domain for given document is considered as context. We also perform detailed experiments on genuine news reports or documents and the outcomes show that the proposed approach is promising for implementation in the real world.

Keywords*—Context Identification; Vector Space Document Model; Corpus; Telugu Context Identification; Natural language processing (NLP); Machine Translation System (MTS),*

I. INTRODUCTION

The objective of designing a new tool for context identification of Telugu language documents is to improve the efficiency of reading and accessing Telugu data, context detection and tracking hot events over rapidly progressing Telugu data on the Internet. Getting information has turned into the essential approach to know what is happening around the world. But the organization of web information is redundant and messy. Identification of contexts of Telugu documents is becoming increasingly necessary for developing intelligent semantic web and information retrieval systems for Telugu data. Advances in computational technologies have allowed the possibility of creating lots of new tools to detect topics and track hot topics from documents. Although the traditional topic detection and tracking methods cannot find hot topics of Telugu documents, they are not context-aware for Telugu data and the granularity of detected topic is broad.

In this paper, we propose a tool called CDetector, a vector space document model based context identification approach for Telugu language documents, and the main contribution of this paper is summarized by three stages:

- We group a lot of Telugu information archives into a large size of Telugu corpus (approximately 18,985,337 words available in the corpus), as per authors knowledge this is the largest available corpus of Telugu language with approximate 19 million words. Later this corpus is transformed into a well-formatted stop words list. This stop words list is used in the second stage. On other hands, we group large amount Telugu data documents into vector space document model information of unique words with its occurrence frequency. We consider this as fine-grained information, on which we apply filtering, retrieving and sorting techniques.
- We use stop words list made in the first stage to match and expel stop words from vector space document model information and include position parameter alongside frequency, these parameters enabled us to compute coordinating matching score in various domains.
- We were coordinating these words in various domains, and after that, we utilized position and occurrence parameters to figure out coordinating matching score of given document in particular space domain. At the end, as a prediction, the maximum matching score of the particular domain for given document is considered as context. We also perform detailed experiments on genuine news reports or documents and the outcomes show that the proposed approach is promising for implementation in the real world.

The rest of this paper is organized as follows: Related work is introduced, which represents how to model vector space document model with occurrence parameter and construct co-occurrence matching score with occurrence and word frequency parameters with coordinating available domain data. Section III describes the system procedure and algorithms for context identification. Then the experiments of Section IV prove that our approach is valid and innovative through several comparative experiments. In Section V, we discuss the conclusion of this paper.

II. LITERATURE REVIEW

As far as, our study of past and contemporary literature for this field is concerned, some work is done for Telugu language using Natural Language Processing (NLP) techniques. For examples, English-to-Telugu machine translation system, Identification of Telugu Devanagari Script [5], Named Entity recognition of Telugu language [6] etc. There is no published literature on NLP for Telugu [8]. As per best knowledge of the author, this is the first attempt to develop a tool to detect context of Telugu documents.

In our research, we noticed that some work is done in context identification. One of them is to identify the context of published papers to give best match result of papers to users who are looking for published research papers for their particular interested area of research [7]. Another one is subjective context identification of text (Natural language text like English) using machine-learning models [8].

The main challenge of our work is to improve the efficiency of reading and accessing Telugu data and identifying the context of given Telugu document. This will enable us to use this tool in Natural Language Processing (NLP), Machine Translation Systems (MTS), Telugu WordNet, Symantec web, classification and Information retrieval systems (IRS) etc.

III. METHODOLOGY

To identify the correct context of given Telugu documents is our main motive of our work. Towards this end, we designed algorithms for tool CDetector. Describe outline of proposed CDetector's algorithms are given below, followed by its description in the form of pseudo code & Data source.

A. Data Collection and Clustering

We first collected large size of Telugu documents from different online articles, blogs and websites [9][10][11][12][13][14][15][16][17], which consist of lot of Telugu Stories, Ramayana, Mahabharata, Political News, Entertainment News, Sports News and Telugu Blogs etc. This formed a large size corpus of Telugu language approx. 228.5 MB data-size and it consists of 15 documents with 18,985,337 words. As a next step, we processed this large corpus to get maximum possible stop-words of Telugu language. As per our research & knowledge after processing available large corpus we got a list of stop-words with 343 most commonly used Telugu words. This stop-words list is used in next steps.

B. Data Processing

We processed given Telugu document data to produce vector space document model information. This information is well formatted and it consists of words occurrence frequency. To produce this information we first split the content into words, then we count the total number of repeats of each word in the same document and finally we got highest to lowest used unique words with its repeat occurrence frequency.

C. Data Extraction and Model Building

We extracted our vector space document model information and removed stop-words from it. At this stage, we got final fine-grained information without stop-words and with occurrence frequency. We added position parameter along with occurrence frequency to enable calculation of matching score with available domains data. We built a model to calculate matching score is:

$$S_m = S_m + ((N_{ds} - p) * o) \quad (1)$$

Where, S_m = Domain Matching Score
N_{ds} = Document Size (Number of Words)
p = word position in vector space document model content
o = occurrence frequency of unique word

D. Mapping Co-relation with Available Domains

According to built model & its equation, we match one by one word from vector space document model with available selected domains data. On each match of words on both sides, we did a calculation of matching score. We repeated this step till all the matches finished, by adding a matching score in itself & got a final grand matching score of the entire domain. In some cases, if no word is matching with any domain it is considered as it belongs to a catch-all domain. With this same procedure, we got different domains matching scores. At the end, we are compared each matching score with each other & considered the maximum matching score's domain is a context of given input.

E. Producing Results

At the final stage, we got the values of the total number of words processed, the total number of unique words, the total number of stop-words, total number of words after removing stop-words, matching scores of each domain and finally identified context domain of given Telugu document. We displayed results using all this available data.

As discussed in above our main process is partitioned into three stages for easy understanding. Here is the graphical representation of our CDetector structure in this figure [FIGURE I] and its algorithm.

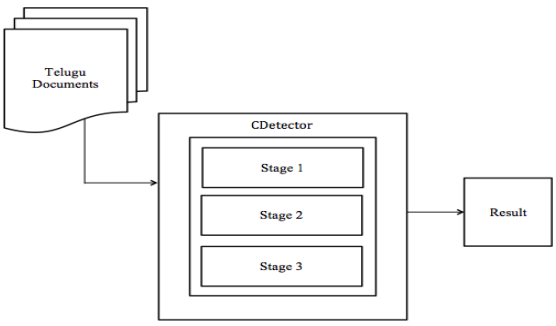


FIGURE I. STRUCTURE OF CDETECTOR

We designed two algorithms called *FilterStart* and *FilterEnd*, *FilterStart* algorithm covers the procedure of stage one and *FilterEnd* algorithm to cover the procedure of stage two and stage three.

Algorithm 1: Stage1: *FilterStart*

Input: Telugu Document Content

Output: Vector Space Document Model with Occurrence Frequency (with total number of processed words, total number of unique words, total number of stop words, total number of words used in calculation)

```
inputFileContent ← read document content from textarea;
if inputFileContent is empty then
    return please enter content first;
br ← read available telugu stop-words list from file sw.txt;
stopWordsFileContent ← merge each line of sw.txt content
        from br to single paragraph;
array words1 ← split inputFileContent words;
array words2 ← split stopWordsFileContent stop-words;
totalWordCount =0, stopWordsCount=0;
for w1 in words1 do
    if w1 is not empty then
        n ← count occurrence frequency of w1;
            map ← (w1, n);
        for w2 in words2 do
            if w2 is not empty then
                if w1 equals w2 then
                    remove w1 from map;
                    stopWordsCount++;
    totalWordCount ++;
list ← sort map;
uniqueWordsCount=0;
for entry in list do
    filter1.txt ← write entry.key &
    entry.value (words, frequency);
    uniqueWordsCount++;
```

Algorithm 2: Stage 2 & Stage 3: *FilterEnd*

Input: Vector Space Document Model with Occurrence Frequency

Output: Identified Domain Matching Score

```
domainScore ← domainMatch("sports.txt");
repeat domainMatch function N times and store domainScore
in respected variables;
show results based on domainScore's comparison &
totalWordCount, stopWordsCount, uniqueWordsCount
function domainMatch(paramfile)
    br1 ← read vector space document model words
        list from file filter1.txt;
    br2 ← read vector space document model words
        list from parameter file paramfile;
    domainContent ← merge each line of paramfile content
        from br2 to single paragraph;
    array words ← split domainContent words;
    matchScore = 0;
    while line ← readline from br1 is not equals to null do
        array columns ← split line;
        word1 ← columns[0];
        occurrence ← parseInt(columns[1]);
        for w in words do
            if w is not empty then
                if w equals word1 then
                    matchScore ← matchScore + ((DocSize –
position) * occurrence);
        position++;
    return matchScore;
```

IV. RESULTS & FINDING

We conducted different experiments on Telugu articles took from different websites [9][10][11][12][13][14][15][16][17]. They consist of Telugu Stories, Ramayana, Mahabharata, Political News, Entertainment News, Sports News and Telugu Blogs etc. Out of 100+ trials in 93.6% cases CDetector identifies correct result, 4.1% cases identified as catch-all context domain & rest 2.3% identified wrong results. Finally, we build the ground truth of top 20 trials of Telugu documents and producing confusion matrix, precision, and recall, F-measure based on it.

TABLE I. CONFUSION MATRIX TABLE

n=20		Predicted				
		Animal & Birds	*Food & Health*	*Politics*	*Sports*	*Catch-All*
Actual	*Animal &* Birds	3	0	0	0	0
	Food & Health	0	2	0	0	0
	Politics	0	0	6	0	1
	Sports	0	0	1	6	1
	Catch-All	0	0	0	0	1

Above confusion matrix table [TABLE I] is as follows, we have 20 documents for 20 trials, so n=20, we are passed 3 "Animal & Birds" context domain documents to our tool and it is identified correctly in the same context, next we passed 2 "Food & Health" & 1 "Catch-All" context domain documents

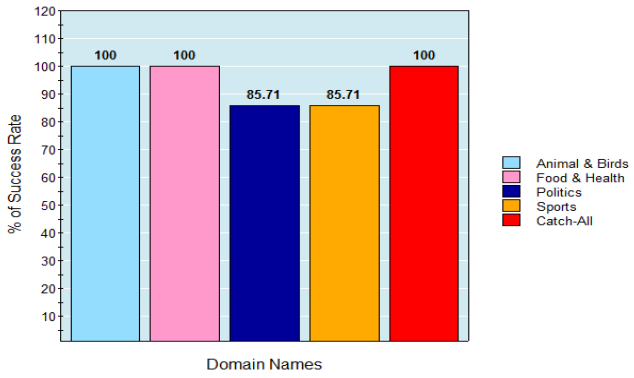
to our tool and again we got correct identification result, But if we focus other cases. In the first case, we passed 7 "Politics" context domain documents to our tool & it is identified 6 domains in politics & another 1 in the catch-all domain, which is wrong. In the second case, we passed 7 "Sports" context domain documents to our tool, and it is identified as 6 domains in sports but 1 in politics, which is wrong.

This is the special case and we went through this issue because of recent news of removing Anil Kumble as cricket head coach because of different kinds of politics. In this type of ambiguity, it is actually sports context but our tool is getting false because this document contains a combination of a lot of political Telugu words.

TABLE II. PRECISION, RECALL AND F-MEASURE TABLE

	Precision	Recall	F-measure
Animal & Birds	100%	100%	100%
Food & Health	100%	100%	100%
Politics	85.71%	30%	44.44%
Sports	85.71%	30%	44.44%
Catch-All	100%	100%	100%

This precision, recall and the F-measure table [TABLE-II] is calculated based on 20 documents only. For example, for "Politics" precision is 6/7=85.71%, recall is 6/20=30% and F-measure is harmonic mean of precision and recall is 44.44% [21][22][23][24][25][26].



V. CONCLUSION

In this paper, we focus on how to improve the efficiency of reading and accessing Telugu data, context detection and tracking hot events over rapidly progressing Telugu data on the Internet. To address this issues we propose two algorithms FilterStart and FilterEnd, the second algorithm is dependent on the first algorithm. This same approach can be used to any level of context identification. For example, we can even identify which sports context is present in given Telugu sports document. We feel that there is a need of CDetector to overcome all these issues and the outcomes show that the proposed approach is promising for implementation in the real world.

In real world application, our tool will be usable in some areas like digitalization in government or private data warehouses, improve machine translation system (MTS) (Telugu to English), classification, Symantec Web, Information Retrieval System (IRS), Telugu WordNet etc. Our work will focus on mining sentiment, emotion detection from Telugu documents in the future.

REFERENCES

[1] Meng Zhao, Chen Zhang, Siyu Lu, Hui Zhang, "STeller:An Approach for Context-Aware Story Detection Using Different Similarity Metrics and Dense Subgraph Mining", IEEE 2016.

[2] Jatinderkumar R Saini, Apurva A Desai, "A Supervised Machine Learning Approach with Re-training for Un- structured Document Classification in UBE", INFOCOMP, 2010.

[3] Jatinderkumar R Saini, "First Classified Annotated Bibliography of NLP Tasks in the Burmese Language of Myanmar", INFOCOMP, 2016.

[4] Durga Prasad Palanati, Ramakrishna Kolikipogu, "Decision List Algorithm for Word Sense Disambiguation for TELUGU Natural Language Processing", IJECCE, 2013.

[5] M C Padma, P A Vijaya, "Identification of Telugu, Devanagari and English Scripts using Discriminating Features", IJCSIT, 2009.

[6] Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, "A prototype for context identification of scientific papers via agent-based text mining", IEEE, 2017.

[7] Ovidiu Fortu, Dan Moldovan, "Identification of Textual Contexts", Springer, 2015.

[8] Approaches to Natural Language Processing, http://shodhganga.inflibnet.ac.in/bitstream/10603/25599/7/07_chapter2.pdf, 2017.

[9] Shakti peethas, Spiritually rich and energetic places, http://www.shaktipeethas.org/, 2008-2017.

[10] Indian Festivals Calendar 2014, Hindu Festivals 2014, Biographies of Indians, Stories of Birbal, Panchatantra, Arabian Nights, Telugu Novels, Hindu Scriptures and Epics, Hindu Pilgrim Centers, http://www.bharatadesam.com/, 2011-2012.

[11] పిట్ట కథలు, బుర్ర కథలు, ఇంకా మరెన్నో... | Telugu blog with stories for children and grown-ups alike - these are not original stories, rather, a compilation of folk tales and moral stories I've read since childhood, https://kathalu.wordpress.com, 2015-2016.

[12] నీతి కథలు | మన భారతీయ సంస్కృతికి సవినయ నివాళి, https://neetikathalu.wordpress.com, 2006-2008.

[13] Prajasakti - Telugu Daily News Paper Latest News,Entertainment,Politics,Sports,Etc .., http://www.prajasakti.com/, 2015-2016.

[14] Vaartha - తెలుగు జాతీయ దిన పత్రిక, http://www.vaartha.com/, 2015-2016.

[15] Homepage | Andhrabhoomi - Telugu News Paper Portal | Daily Newspaper in Telugu | Telugu News Headlines | Andhrabhoomi, http://www.andhrabhoomi.net/, 2015-2016.

[16] visalaandhra - visalaandhra, http://www.visalaandhra.com/, 2015-2016.

[17] Telugu News | Online Telugu News | Latest Telugu News | News in Telugu - Oneindia Telugu, http://telugu.oneindia.com/, 2015-2016.

[18] Algorithm, https://en.wikipedia.org/wiki/Algorithm, 2017.

[19] Confusion matrix, https://en.wikipedia.org/wiki/Confusion_matrix, 2017.

[20] Simple guide to confusion matrix terminology, http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/, 2014.

[21] Metacademy, https://metacademy.org/graphs/concepts/f_measure#focus=precision_recall, 2017.

[22] F1 score, https://en.wikipedia.org/wiki/F1_score, 2017.

[23] Harmonic Mean, http://www.alcula.com/calculators/statistics/harmonic-mean/, 2017.

[24] Chetan Kalyan, Min Young Kim, "Detecting emotional scenes using Semantic Analysis on Subtitles", CS224N, Stanford University, 2009.

Rajeshkumar S. Gone has graduated (B.C.A) in computer applications from Veer Narmad South Gujarat University, Surat, Gujarat, India. He has completed his first masters (M.C.A) in computer applications from the same university, the perusing second masters (M.Tech) in computer science and engineering from Jawaharlal Nehru Technological University, Hyderabad, Telangana, India. He is Microsoft certified specialist and Microsoft certified professional. He has 5+ years of technical experience.

Dr. Jatinderkumar R. Saini has graduated (B.Sc.) in Computer Science from Veer Narmad South Gujarat Technological University, Surat, Gujarat, India. He has completed his masters (M.C.A) in computer applications and Ph.D. from the same university. He has 12+ years of academic experience. He is presently working as

Professor & I/C Director at Narmada College of Computer Applications (NCCA), Bharuch, Gujarat, India. His research expertise in Natural Language Processing, Web Mining, Document Analysis, Text Mining, Machine Learning. file in the form of brief detail about his/her academic achievements and research area.

Kande Srinivas has graduated (B.E) in computer science and engineering from Osmania University, Hyderabad, Telangana, India.

He has completed his masters (M.Tech) in computer science and engineering from Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India. He has 16+ years of academic experience. He is presently working as Associate Professor & Head of Department at computer science and engineering Department of SVS Group of Institutions, Warangal, Telangana, India.