May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

Deep Learning Based Grammar Checker for Kannada

[1]Caryappa B C, [2]Vishwanath R Hulipalled, [3]J B Simha

[1,2,3]School of C and IT, Reva University, Bangalore, India
[1]caryappa.cari@gmail.com, [2]vishwanath.rh@reva.edu.in, [3]jbsimha@reva.edu.in

Article Info
Volume 83
Page Number: 4524-4530
Publication Issue:
May - June 2020

Abstract
Language is the most basic and traditionally natural means of communication in the present day to day conduct. And grammar places a vital role in the success of a language. As Humans have been trained throughout out life with a abundance data that is accumulated, refined over course of time with rules and compression of relevance to enable us to understand and converse between one another. But to incorporate such under-standing to a machine , to be able to evaluate and differentiate contextual information into proper grammatical form hence to validate that the information is in the right form is also equally important in the present day as it well as it a complex chore . The paper addresses this problem and proposes the development of such grammar checking tool for the Dravidian language Kannada. One of the first consideration is that the complexity of the language poses a challenge and opting to use a rule based approach is a easier solution and allows to identify flagged errors efficiently. It requires a linguistic expert to draw out hundreds of sequential rules that is complex to maintain. Here, a model is proposed that uses a deep learning approach to train a LSTM (Long Short Term Memory) neural model trained over a large data set to achieve the required classification, using a context retaining representation of the data achieved through Word2Vec along TensorFlow and Keras packages. The proposed model is capable of efficiently performing Grammatical error detection (GED)

Index Terms**:** *Natural language processing, LSTM, GED, Word2Vec, Deep learning, Neural Network, Word Embedding*

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

1. Introduction

Language is a source for construing correspondence between individuals either verbally or through a composed medium. Communication among people occur usually through natural language, and in the southern part of India is dominant of the Dravidian languages. One of the Dravidian dialects prominent in the state of Karnataka (earlier ”Mysore”) is Kannada, a Dravidian derivative language. As with any natural language trans-communication, it is essential to validate the sentences conversed. Morphemes, phonemes, words, phrases, clauses, sentences, vocabulary and grammar are the ingredient of all natural language. All authentic sentences of a language must keep the principles of that language (sentence structure). Invalid sentences does not achieve to transfer or share knowledge, hence out rightly rejected.

A sentence’s architecture is perceived hierarchically at various levels of abstraction, i.e. surface level (at the word level), POS (part-of-speech) level to abstract level. The sentence formation closely banks on the syntactically admissible structures encoded in the language grammar regulation [1]. The primary sentence structures majorly depend on the positions of Subject, Object, Verb are followed in the grammar of natural languages of the world. All of these are referenced as word order. Natural Language Processing or also refereed to as Language Technology is a class of computer science and AI that is concerned with communication between human and machine language. NLP is leading as one of the prominent research today. NLP process can be sub classified as Natural Language Understanding and Natural Language Generation. Operating on natural

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

languages using computers is known as NLP. The sub-field of Machine Learning, Deep Learning, is currently leading the innovations in NLP. Deep Learning can be summarized as a form of machine learning that enables computers to gain from experience and comprehend the world in terms of a hierarchy of concepts. Just like other Machine Learning areas, a neural network can be designed similarly that can be adapted to the processing of raw text data. One important characteristic that text data will encounter when creating a solution however, is that the text data is high dimensional. Applications of ML have gotten much traction in the last few years in what is known as sequence to sequence space. Long Short-Term Memory units were introduced as an enhancement to Recurrent Neural Networks alone as a shortcoming of the Recurrent Neural networks is that it is only capable of dealing with short-term dependencies [2]. They address this problem by introducing a long- 18 term memory into the network. This is where future predictions are based on the previous output, and the prediction to the previous output was calculated using the one before that one. This technique will help in the grammatical correction since the nature of sentences completely rely on context to be formulated correctly [3]

Rest of the paper is organized as follows: section II discusses on Natural Language Processing, in section III Related works is discussed. Section IV illustrates the proposed methodology and section V the experimental results have been discussed. Finally, the paper is concluded.

2. Natural Language Processing

Natural Language processing is an interdisciplinary branch of etymological (linguistic) and computer science studied under the Artificial Intelligence (AI) that produced an associated zone called 'Computational Linguistics' which focuses on processing of natural languages on computational devices. Natural language processing (NLP) is a theory-inclined stretch of computational techniques for the automated analysis and representation of human language. NLP research has expanded from the era of punch cards as well as batch processing, in which the investigation and analysis of a sentence could take up to 7 minutes, to the era of in which millions of webpages can be processed under a second. NLP empowers computers to perform a broad scope of natural language related tasks at all levels, reaching from parsing and part-of-speech (POS) tagging, to machine translation and dialogue systems.

NLP has been a field of research for decades that attempts to make computers understand and speak human languages as well as humans do. It has spawned multiple platforms of hu-man to computer interaction such as Amazon's Alexa, Apple's Siri, spell checkers, sentence summarizers, and many more. For years, researchers attempted to create a language modeling technique that attempted get computers to communicate with humans at a level that humans feel they are communicating with another human[12]. The work done by scientists in the early 1950s is considered the era of machine translation, where being able to treat text and language in general as information allowed the possibility that language might be manipulated on the new digital computers that were then being constructed.

3. Related Work

A present trend in Neural machine translation (NMT) has shown a significant concern in incorporating linguistic knowledge. So far the existing previous works have chosen either to customize and engineer NMT encoder to include syntactic information into the transition model, or to generalize and condense the embedding layer to encode added linguistic features. The past approaches are mainly centered on injecting the syntactic architecture of the source sentence into the encoding process, which results in a complicated model that is deficient of flexibility to consolidate more types knowledge [1]. The following expands word embeddings by considering newer morphological information as feature to advance the word portrayal. Thus it does not specifically adjust the engagement from word embedding from additional linguistic knowledge [2]. To deal with corresponding constraint, "a knowledge-aware NMT" approach that models added extra linguistic features in similarity to the word feature proposed. The root of the idea projected is that modeling a linguistic features at the word level (knowledge block) employing a "Recurrent Neural Network (RNN)". The word features are further encoded again using an RNN in sentence level. A knowledge aware gate and attention gate are deployed in decoding to dynamically govern the proportion of information adding to generation of target words from various sources. Comprehensive experiments present that this method is capable of improved accounting for importance of additional linguistics.

One amongst the primary requisites in resolving various problems related to Natural Language Processing (NLP), data mining, etc. in a deep learning based model is adequate input representation. Absence of sufficient appropriate representation for the input opens the problem of data sparsity, and it presents an particular challenge to address the underlying issue. r , i.e. lacking of sufficiently large corpus that is desired to train a Word embedding system. This Work proposes an efficient method to enhance the word embeddings in less-resourced language by utilizing the influence of bilingual word embeddings learned from various corpus. Training and valuating a deep learning Long Short Term Memory (LSTM) based architecture and demonstrate the success of proposed approach for two levels of sentiment analysis (i.e., Aspect term extraction and Sentiment classification). With further aid from by handcrafted features in neural network for prediction. Applying the model in two experimental setups: multilingual and cross-lingual. Experimental outcomes show the essence of the proposed approach against the innovation strategies.

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

The paper "UTTAM: An Efficient Spelling Correction System for Hindi Language Based on Supervised Learning" proposes a system called "UTTAM,"[3] for amending spelling mistakes in Hindi language implementing a supervised ma-chine learning. In contrary to other dialects, Hindi is made up of a characters, words with inflection and complex characters, morphologically similar character sets. The proposed method examines the human behavior i.e., the type and frequency of spelling errors committed by people in hindi text[3]. Establish-ing on the frequency of spelling errors, the heterogeneous data is accumilated in matrices. This information in matrices is used to generate and develop the suitable candidate words for an input word. After generating the candidate words, the Viterbi algorithm is implemented on these candidate words to perform the error correction. The Viterbi algorithm identifies the best sequence of candidate words to process the input sentence. For Hindi, it is a novel attempt for real world error correction. For non-word errors, the results depict that "UTTAM" performs efficiently than the present SpellGuru and Saksham systems.

POS information in an annotated corpus of a language is generally an essential requirement for processing of natural language in computational linguistics and computer science. The context and utilization of the relevant words inferred from the POS data has proved to be helpful in NLP application[4]. Languages such as Spanish, English and French infer the POS information from an available untagged text to assist future Natural Language Processing tasks. A huge tagged corpora ,human tagged, is used to train a supervised learning model for Parts-of-speech tagging to obtain finest performance. The insufficiency of tagged corpus for languages like Sanskrit , Kannada and another resource poor languages serve as impediment for these conventional supervised learning algorithm as it generally require rich datasets[4].

The methodology proposes an approach to tag POS to words that are fed as input, based on deep learning. The proposed approach, that is unsupervised , as Sanskrit lacks a large annotated corpora unlike some other languages and uses the untagged Sanskrit corpus generated by JNU. Then performs dimensionality reduction, implementing auto encoder, to encode and compress the vectors depiction which are compatible for clustering in the vector space. After speculating with various dimensions of the compressed depiction and current one which promises better clustering performance. To retain the semantic data the embeddings are processed through a bi-directional Long short term memory (BLSTM) autoencoder. And assign nominate a POS tag to the obtained cluster and reference the model against a tagger corpus to generate the result [4].

Natural language can described as the process of inter-changeability of information between people. Grammar is elementary in language and it contains laid out regulation. Words are rudimentary units of grammar and these units combine together to construct sentences [5]. These sentences are formulated by using the constraints set according to grammar. Grammar is a set of rules and these rules are used to formulate sentences. There are many grammatical errors occurring during the writing process. This paper discussed on a survey upon grammar checkers for various languages of India that are modeled using deep learning techniques. Grammar evaluation is a fundamental process for the task of writing. Grammar consists of many statute including past, present and future. There are numerous grammar checker for various languages which strive to improve accuracy for minimizing error. Below Table 1 shows few of the NLP Works carried out in Kannada language.

A. Grammar Checker

Grammatical error correction (GEC) is a challenging labor due to the inconsistency of the form of errors and the syntactic and semantic dependencies of the errors on the context of the input word. An ample amount of grammatical rectification systems utilize classification and rule-based approach for amending explicit errors [6]. However, these systems use multiple linguistic cues as features. The standard linguistic reasoning tools like parts-of-speech (POS) taggers and parsers are regularly trained on well-formed text and perform crudely on ungrammatical text [7]. This injects further errors and limits the performance of rule-based and classification technique to Grammar error correction (GEC).

B. Rule Based Grammar Checking

The traditional methodology of language structure checking is to physically plan syntax rule. These accurate and detailed rules are designed by linguistic experts [16]. A Kannada POS tagged Kannada text corpora is validated against a characterized set of rules and a coordinating guideline is applied to identify any error [7]. The method seems quite simple as it is easy to modify, include and eliminate a rule, nonetheless, drafting out the rule requires a extensive knowledge of the language being worked with. Rule based system can explain accurately the flagged errors thus making the system reliant for the purpose computer aided language learning [19]. It can be challenging to maintain hundreds of grammatical rules. Below Table 2 shows Grammar Checker for Indian languages[17,18].

Table I: NLP Works in Kannada

Reasearch Paper	Method Used	Conclusion
"Kernel Based Part Of Speech Tagger For Kannada"	NLP, POS Tagger, SVM	This research presents a Part-Of-Speech tagger for Kannada language which is developed using SVM

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

		kernel model. The research conductes a linguistic study to resolve the internal linguistic structure of a Kannada sentence and using that come up with a suitable tagset.
”Building a Kannada POS Tagger Using Machine Learning and Neural Network Models”	Part of Speech (POS), NLP, SVM	The best F1-score of 0.92 was obtained in the CRF model on using a window feature of [-2,+2] and in the neural network model where character embedding is used along with pre-trained word embeddings. The accuracy of the neural networks can be improved by training it on a larger dataset.
”A novel approach to Sandhi splitting at Character level for Kannada Language”	Natural Language Processing, Sandhi Splitter, Conditional Random Fields	The proposed technique in this work delivers a new perception of existence of a character level sequence in the text where the Sandhi can be split.
”Development of Prototype Morphological Analyzer for the South Indian Language of Kannada”	Finite State Machine, Kannada Content Management, Natural Language Processing.	The study delivers a morphological analyzer that can be implimented inn spell checking, POS tagging and stemming simultaneously. This works as an efficient medium for the pre-processing activities of Kannada document digitization and content management.

Table II: Grammar Checker for Indian Languages

Language	Technique	Results
Bangla	Statistical	The Model performs better for Bangla language in correspondence with English
Punjabi	Rule Based	Indicates precise message of the error and provide suggestive corrections. Works well in case of simple, compound and complex sentences as input.
Hindi	Rule Based	Generates assuring outcomes for simple sentences
Urdu	Rule Based	Checks structural and grammatical oversight in sentences. Suggests error correction.

4. Methodology

Machine learning is presently the most trending domain in every field and also more profitable to use in grammar check-ing. Most promising results can be established implementing a supervised learning. These strategies take advantage of an annotated corpora which is used to execute statistical analysis on the text to naturally identify and address grammar/syntax errors. Although in contrast to rule based frameworks, it is challenging to explain and establish the errors, with this system. Machine learning framework does not require complex understanding of the grammar as it is completely reliant on the corpus it is trained on. Unavailability of a good and well established large annotated corpora hampering the application of technique for grammar checking.

A. Preprocessing

Usual preprocessing steps are used to take some text data in its raw form and transform it into text data that will be more useful for neural network processing. Using python as our data scripting language, the model was trained by using TensorFlow, the dataset was increased by getting some snippets of text and created the incorrect/correct pairs by sample be the target[6].

Obtaining a randomized sample from our data.

Randomly mutate some of the characteristics of the sample of data.

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

Having the untouched original and presumably correct syntax.

The random mutations introduced were limited to targeting articles of the sentence, removing the part following a verb contraction, and substituting words that have multiple mean-ings but the same phonetic sound or pronunciation[10]. This was done 3 times to increase the data and could potentially be done more times, but considering time constraints, three was chosen. The model trained can be classified as a sequence-to-sequence model.

B. Word2Vec Encoding

Number of current Natural Language Processing (NLP) frameworks and techniques approach with words as granular units - there is no conception of similarity between words, as these are expressed as indices in a vocabulary. This approach has several positive aspects - effortlessness, robustness and the experience that simple models trained on huge amounts of data outperform compound framework trained on lesser data[11]. But the draw back with most encoding techniques is that they do not account for the context of the word encoded, here where Word2Vec relates. W2V (Word2Vec) is a collection of associated or relevant models that are used to generate word embeddings [8]. This constitutes a two-layered neural network trained to reassemble the morphological contexts of the word. Word2Vec generated a vector space representation based of the large textual corpus fed to the model, usually of a high dimension of several hundred, where each unique word in the corpus being assigned a respective vector in a the space. Word vectors are depicted in vector space with the outcome that the words with common context are represented closer to each other in the vector space. Word2vec was a model developed by a team at google driven by Tomas Mikolov [9]. Word2Vec models can be structured either with hierarchical SoftMax or negative sampling or a combination of both. To minimize calculation the hierarchical SoftMax adopts Huffman tree that approximates log-likelihood [13].

The negative sampling approach, on the other hand tackles the maximization issue by curbing the log-probability of the sample of negative instances. Hierarchical SoftMax excels for contrary words while negative sampling outmatches when considering recurrent words and better with low dimensional vectors[14]. With expansion in training epochs, hierarchical SoftMax presents irrelevant. The amount of words that are integrated as context of the input word is dependent on the size set to the context window. The generally endorsed value is 10 for skip-gram and 5 for CBOW[14]. Figure 1 indicates the process flow for building the Word to vector model.

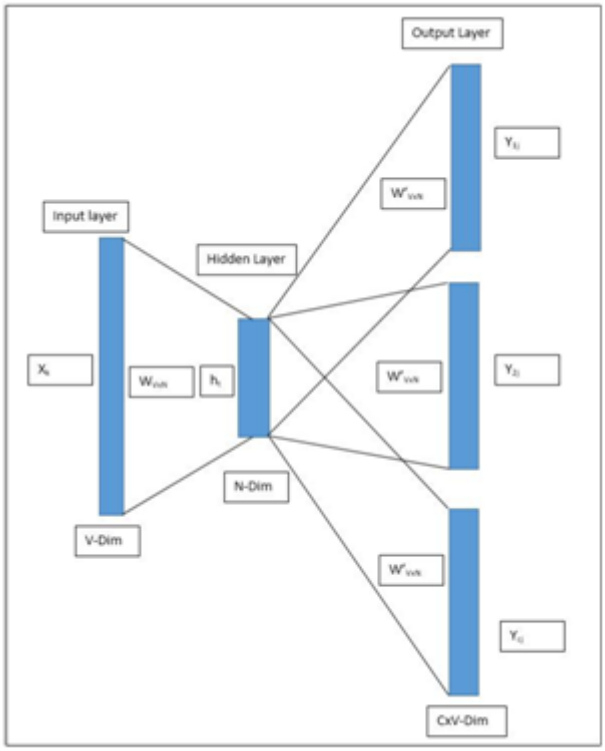

Figure 1: Word2Vec model

Word2Vec(W2V) is not a sole algorithm but a concoction of two functions CBOW(Continuous bag of words) and Skip-Gram model. The pair are shallow neural networks which map word(s) to the objective or target variable which is also a word(s). Either of these technique learn weights which measure as word vector portrayals[15].

5. Experimental Results

The model uses the textual data that is scraped from a children's tale so as to keep the words simple and not obscure to avoid complex ambiguities, when understanding the outcomes of the model. The Observation by using the W2V embedding to embed the trained data is that the model completes training over the corpus resulting with a vector representation in vector space. Figure 2 depicts the vector similarity of the words based on the trained corpus.

The embedding of words is a representation of these words as vectors in a vector space where similar or contextual words reside closer to each other in the vector space. Vector space is assemblage of vectors. As words are represented as vectors, like vectors scalar operations such as addition and multiplication can also be performed on them. The model preforms well to the corpus in context.

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

```
In [100]: w2v.vec_sim("ಓದಿಗೆ",3)

ಓದಿಗೆ 1.0
ನಾನೇನು 0.2118803271539788
ಎಂದರು. 0.20702234192532573

In [101]: w2v.vec_sim("ಆಗದ",3)

ಆಗದ 1.0
ಶಾಲೆಗೆ 0.2789790679220233
ನಾವೇ 0.2319413296673729
```

Figure 2: Word vector similarity index

As evaluating such attempt would need comparison against a ground truth which is lacking is a resource-poor language such as Kannada. But the model presents adequate results in context to the data it is modeled over. The Figure 3 shows the word embedding represented in a vector space. Which when fed to a LSTM neural network that retains the structure of the sentence while holding the context of the words in place. The model does not fair well with existing grammar checkers in other prominent languages such as English, but this is due to the Kannada being an in-fluctuated language and also a resource poor language. With no existing grammar checker present for the Kannada language, this model fair against itself.

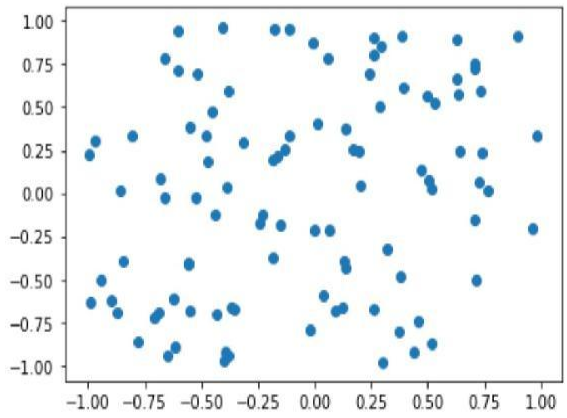

Figure 3: Vector space representation of the word embedding

6. Conclusion

This paper proposed a deep-learning model combining the Word2Vec technique for sophisticated embedding representation of the Kannada language that is then trained with a Neural network with LSTM layer that retains the context to the word increasing the reliability and accuracy of the model. One latency observed is the availability of a large annotated corpus for Kannada language. Further improvement to the paper can be to devise stemming and lemmatization to the data to generate more quality representation of the words. Further training the model on much larger dataset to refine the model and make it publicly available for further research.

References

[1] Qiang Li , Derek F. Wong Lidia S. Chao Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang, “Linguistic Knowledge-Aware Neural Machine Translation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume: 26 , Issue: 12 , Dec. 2018.

[2] Md Shad Akhtar ,Palaash Sawant, Sukanta Sen, Asif Ekbal, And Pushpak Bhattacharyya “Improving Word Embedding Coverage in Less-Resourced Languages Through Multi Linguality and Cross-Linguality: A Case Study with Aspect-Based Sentiment Analysis” ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 18, No. 2, Article 15. Publication date: December 2018.

[3] Amita Jain, Minni Jain,Goonjan Jain, Devendra K. Tayal “UTTAM”: An Efficient Spelling Correction System for Hindi Language Based on Supervised Learning, ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 18, No. 1, Article 8. Publication date: November 2018.

[4] Prakhar Srivastava, Kushal Chauhan, Deepanshu Aggarwal, Anupam Shukla , Joydip Dhar, Vrashabh Prasad Jain “Deep Learning Based Unsupervised POS Tagging for Sanskrit” ACAI ’18, December 21–23, 2018.

[5] Neethu S Kumar, Supriya L P “Survey on Grammar Checking and Correction using Deep Learning for Indian Languages”, International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 11 — Nov 2018

[6] Nivedita S. Bhirud1 R.P. Bhavsar2 B.V. Pawar3 , “GRAMMAR CHECKERS FOR NATURAL LANGUAGES: A REVIEW ” , Inter-national Journal on Natural Language Computing (IJNLC) Vol. 6, No.4, August 2017

[7] Muhammad Kamran Malik, “Urdu Named Entity Recognition and Classification System Using Artificial Neural Network” ACM Trans-actions on Asian and Low-Resource Language Information Processing (TALLIP) TALLIP Homepage archive Volume 17 Issue 1, November 2017.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 7 Sep 2013.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Composi-tionality, 16 Oct 2013.

[10] A. Zhila, W.T. Yih, C. Meek, G. Zweig, T. Mikolov. Combining Heterogeneous Models for

May-June 2020
ISSN: 0193-4120 Page No. 4524-4530

Measuring Relational Similarity. NAACL HLT 2013.

[11] J. Turian, L. Ratinov, Y. Bengio. Word Representations: A Simple and General Method for Semi-Supervised Learning. In: Proc. Association for Computational Linguistics, 2010.

[12] H. Schwenk. Continuous space language models. Computer Speech and Language, vol. 21, 2007.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Ac-cepted to NIPS 2013.

[14] T. Mikolov. Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.

[15] T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013.

[16] Daniel Naber. "A Rule-Based Style And Grammar Checker". Diplomar-beit. Technische Fakultat¨ Bielefeld, 2003.

[17] Misha Mittal, Dinesh Kumar, Sanjeev Kumar Sharma, "Grammar Checker for Asia Languages: A Survey", International Journal of Com-puter Applications Information Technology Vol. 9, Issue I, 2016

[18] Alam, M. J., Uzzaman, N., Khan, M. "N-gram based Statistical Gram-mar Checker for Bangla and English." Ninth International Conference on Computer and Information Technology (ICCIT) 2006

[19] Lata Bopche, Gauri Dhopavkar, and ManaliKshirsagar, "Grammar Checking System Using Rule Based Morphological Process for an Indian Language", Global Trends in Information Systems and Software Applications, 4th International Conference, ObCom 2011 Vellore, TN, India, December 2011.