IJCTA, 10(03), 2017, pp. 229-235

Character based Bidirectional LSTM for Disambiguating Tamil Part of Speech Categories

K.S. Gokul Krishnan* A. Pooja* M. Anand Kumar* and K.P. Soman*

Abstract : Part of speech (POS) tagging is the process of labeling a part of speech tag to each and every word in the corpus. In this pa- per POS tagging for Tamil language is performed by using Bidirectional Long Short Term Memory. A C2W (character to word) model instead of traditional word lookup table for obtaining word embeddings using BLSTM is presented. The C2W model uses characters to form a vector representation of a word. The word embedding from C2W model is used by BLSTM to tag the words in the corpus. This method, when tested with 3723 words produced highest accuracy of 86.45%.
keywords: Part of Speech tagging, Neural Network, BLSTM, C2W model, word embedding.

1. INTRODUCTION

Tamil is a Dravidian language, predominantly spoken in Southern state of Tamil Nadu in India and in Sri Lanka. It is spoken by around 80 million people in several countries around the globe. But still research advances in language processing domain for Tamil are mediocre.

Part-of-speech tagging is an essential aspect in speech recognition, machine translation, natural language parsing, morphological parsing and information retrieval. The part-of-speech in a natural language carries large amount of infor- mation about a word and its neighbor [1]. We can design a system with either grammatical category or features for POS tagging. Tamil is a morphologically rich language, with large grammatical features. So designing a system with gram- matical features becomes complex. Hence we propose a Tamil POS tagger system based only on grammatical categories.

The main problem while performing POS tagging task is selecting a part- of-speech label for a word. Since a word can occur in different part-of-speech categories in different contexts, ambiguity must be resolved. For POS tagging problems, machine learning approaches are regularly applied [5] [6] [7]. In this paper, deep learning approach for POS Tagging in Tamil is used. In our pro- posed method, the POS tagging task is performed using Bidirectional Long Short Term Memory. Our model tested with 3723 words produced tagging accuracy of 86.45%.

2. POS TAGGING IN TAMIL

Tamil language is agglutinative in nature which makes POS tagging a complex process. The key challenge in Tamil POS tagging is disambiguation of POS categories. Nouns in Tamil have inflections in case and number. Verbs in Tamil language have inflections in suffixes of number, gender, tense [10].

Many grammatical rule based and statistical methods have been developed and used for POS tagging [8] [9]. A rule based method requires extensive knowledge about the complex grammatical structures which

* Center for Computational Engineering and Networking Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham Amrita University India, *E-mail : Pooja1311.ashok@gmail.com*

makes it impractical to develop a POS tagging system. A POS tagger for Tamil based on morphology has been developed [3]. A POS tagging method using SVM has also been de- veloped for Malayalam [2]. A BLSTM based POS tagger has been designed for English [4].

There are many tagsets available for Tamil (AUKBC, Vasuranganathan tagset, CIIL Tagset for Tamil, etc.). But these tagsets are very large. It contained both grammatical features and categories. So naturally, complexity is increased during POS tagging which leads to reduced tagging accuracy. A tagset containing only grammatical categories was required. Hence we decided to use Amrita tagset for Tamil which contains 32 tags as described in [10].

A corpus of 33216 words was used for training the model. For validation and testing a corpus of 3742 and 3723 words respectively were used. The data is given to the model in conLL format. The conLL format is a structure in which the first column contains the words followed by a tab or space which in turn is followed by the second column, which contains the tags.

3. BACKGROUND

3.1. Bidirectional Long Short Term Memory (BLSTM)

Long Short Term Memory is an artificial neural network. An LSTM cell has the ability to hold a value for a long amount of time. LSTMs are a solution to the long term dependency problem.

Figure 1 shows the schematic of an LSTM unit. The information flows into the cell at multiple points. The combination of present and past state is given to the cell and also to the three gates, which decides how the input will be handled. The gates determine whether to let new input in, erase the present cell state, and/or let that state impact the networks output at the present time step. Here Bidirectional LSTM is used, because the POS category of a word depends on both the previous and future words in the sentence. To predict a POS category of word in a sentence, it is essential to look at both the left and the right contexts. The updates computed are described below in (1-5) [4].

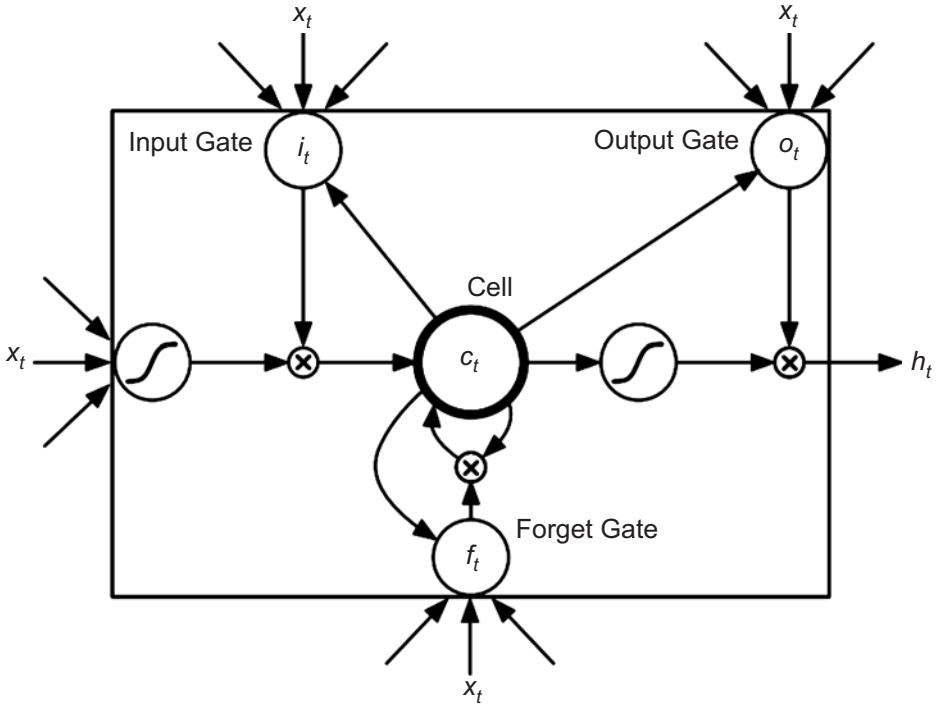

Figure 1: Schematic of an LSTM unit

$$i_t = \sigma(W_{ix} x_t + W_{ih} h_{t-1} + W_{ic} c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{fx} x_t + W_{fh} h_{t-1} + W_{fc} c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{cx} x_t + W_{ch} h_{t-1} + b_c) \quad (3)$$

$$0_t = \sigma(W_{ox} x_t + W_{oh} h_{t-1} + W_{oc} c_t + b_t) \quad (4)$$

$$h_t = 0_t \tanh(c_t) \quad (5)$$

where b is the bias vector. is component wise product function and σ is a component wise logistic sigmoid function.

4. METHODOLOGY

The proposed method for POS tagging using BLSTM is depicted as a block diagram in Fig. 2.

- The input passed into the model is in conLL format.
- These inputs are read by conLL reader. It separates the words and the tags.
- Those words which are generated from conLL reader are passed through C2W model where the words are disambiguated to characters and word embeddings are formed.
- Using the word embeddings in context, the sentences are labelled by a softmax function.

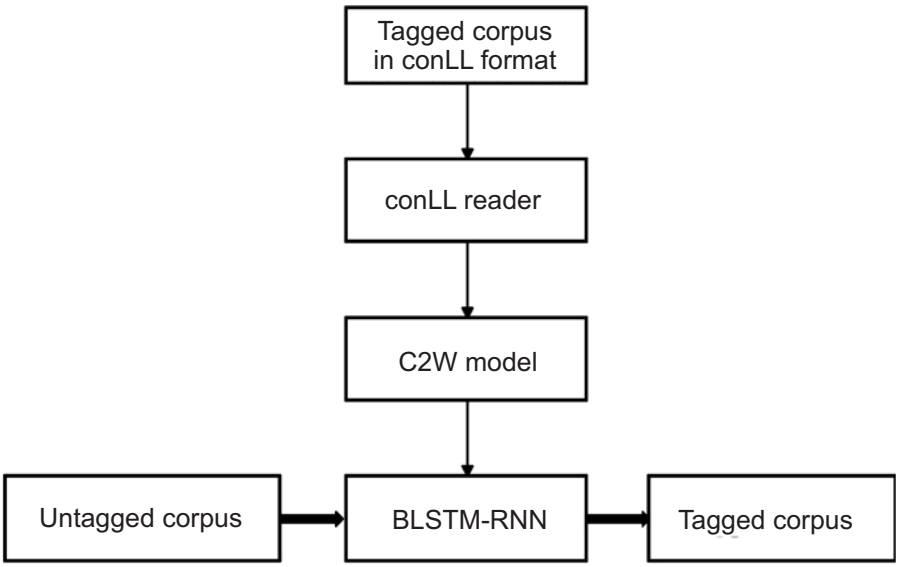

Figure 2: Block diagram of the proposed method

An illustration of our proposed method for POS tagging is shown in Fig. 3.

An illustration of compositional character to word model (C2W) model is shown in Fig. (4). The output vector of our C2W model is same as the vector from word look up table.

The C2W model is a compositional character to word model. Our aim is to obtain a vector for representing word w, which is the input. The vector is of dimension d. It is obtained by following steps.

1. First our word w is split into characters c_1 , c_2 ...c_m , where m is the length of the word w.
2. Every character c_i is assigned a vector the index in vocabulary.
3. A projection layer P_c is created. $P_c \in R^{dc} \times |C|$, where d_c is the dimension of each character, which is the number of parameters. C is the character set.
4. An input character is projected as $e_{ci} = P_c \cdot 1_{ci}$ where 1_{ci} is the one hot vector with one on the index of vocabulary.
5. The character vectors e_{ci} where $i = 1..m$, is passed to BLSTM unit.
6. In forward LSTM, the forward state sequence s_0^f ... s_m^f is computed.
7. In backward LSTM, the backward state sequence s_m^b ... s_0^b is computed.
8. Both the forward and backward state sequences are combined to form the representation of the word w. The word representation is shown in (6), where D^f, D^b are factors of combination.

$$C_w = Ds_m + Ds_0 + b_d \quad (6)$$

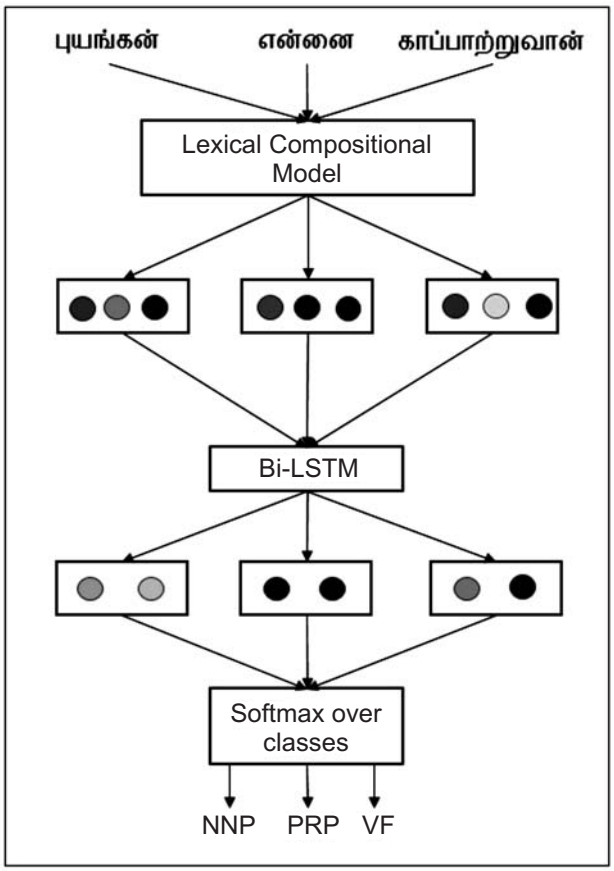

Figure 3: Illustration of our model for POS tagging

5. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed model is executed for multiple variation of parameters such as learning rate, word dimension, character dimension, character state dimension, context state dimension, context model and update. The variation of parameters is tabulated in Table 1. The proposed is executed for 864 times with multiple variation of parameters.

Table 1
Parameter variations

Parameters	*Values / Methods*
Word features	Words, words with prefix, suffix and prefix-suffix characters, characters with prefix, suffix and prefix-suffix
Context model	window, BLSTM
Update method	regular, momentum
Learning rate	0.2, 0.3
Word dimension	50,100
Character dimension	50,100
Character state dimension	50,100,150
Context state dimension	50,100,150

The highest test accuracy and its corresponding OOV accuracy for different features like word and character which further classified as word and character with prefix, suffix and prefix suffix are tabulated in Table 2. From Table 2 it is evident that, when the features vary, the accuracies also change accordingly. From Table 2, it is also evident that character features produce better results than word features.

Table 2
Comparison of word features

Features	*Test*		*Validation*	
	Accuracy (%)	*OOV Accuracy (%)*	*Accuracy (%)*	*OOV Accuracy (%)*
Word	76.86	44.08	78.79	47.70
Word with prefix	84.46	61.23	86.45	68.16
Word with suffix	84.75	64.61	86.45	67.24
Word with prefix, suffix	85.26	64.86	85.99	66.97
Character	81.93	72.21	83.28	71.37
Character with prefix	86.45	66.89	87.36	70.82
Character with suffix	83.08	59.54	86.10	68.53
Character with prefix, suffix	86.29	66.72	89.09	72.20

Table 3
Comparison of context models

Context Model	*Test*		*Validation*	
	Accuracy (%)	*OOV Accuracy (%)*	*Accuracy (%)*	*OOV Accuracy (%)*
Window	80.87	63.59	82.79	64.22
BLSTM	86.45	66.89	87.36	70.82

The context model in our system is BLSTM. Our BLSTM model is compared with another context model *i.e*. window. Our BLSTM model produces higher accuracy than window based model. The highest accuracies of both models are reported in Table 3.

The update methods also produce variation in accuracy. The update methods used where momentum and regular. The momentum update produces higher accuracies compared to regular update. The highest accuracies of both update methods are listed in Table 4.

Table 4
Comparison of update methods

Update Method	*Test*		*Validation*	
	Accuracy (%)	*OOV Accuracy (%)*	*Accuracy (%)*	*OOV Accuracy (%)*
Regular	41.53	41.55	43.25	40.45
Momentum	86.45	66.89	87.36	70.82

6. CONCLUSION

A method for POS tagging of Tamil language using BLSTM is proposed. This is the first attempt in applying BLSTM based model for POS tagging of an Indian language. This is also the first attempt in applying character based method for an Indian language. A compositional character to word (C2W model) is used instead of traditional word lookup table and word embeddings. The C2W models use characters to create word embedding instead of words. The C2W model produces better results compared to word look up table. Character based models produce higher accuracy than word based models. Character based models when executed with prefix and prefix-suffix produce excellent results. POS tagging of unlabeled Tamil language corpus using BLSTM can be considered as a future work.

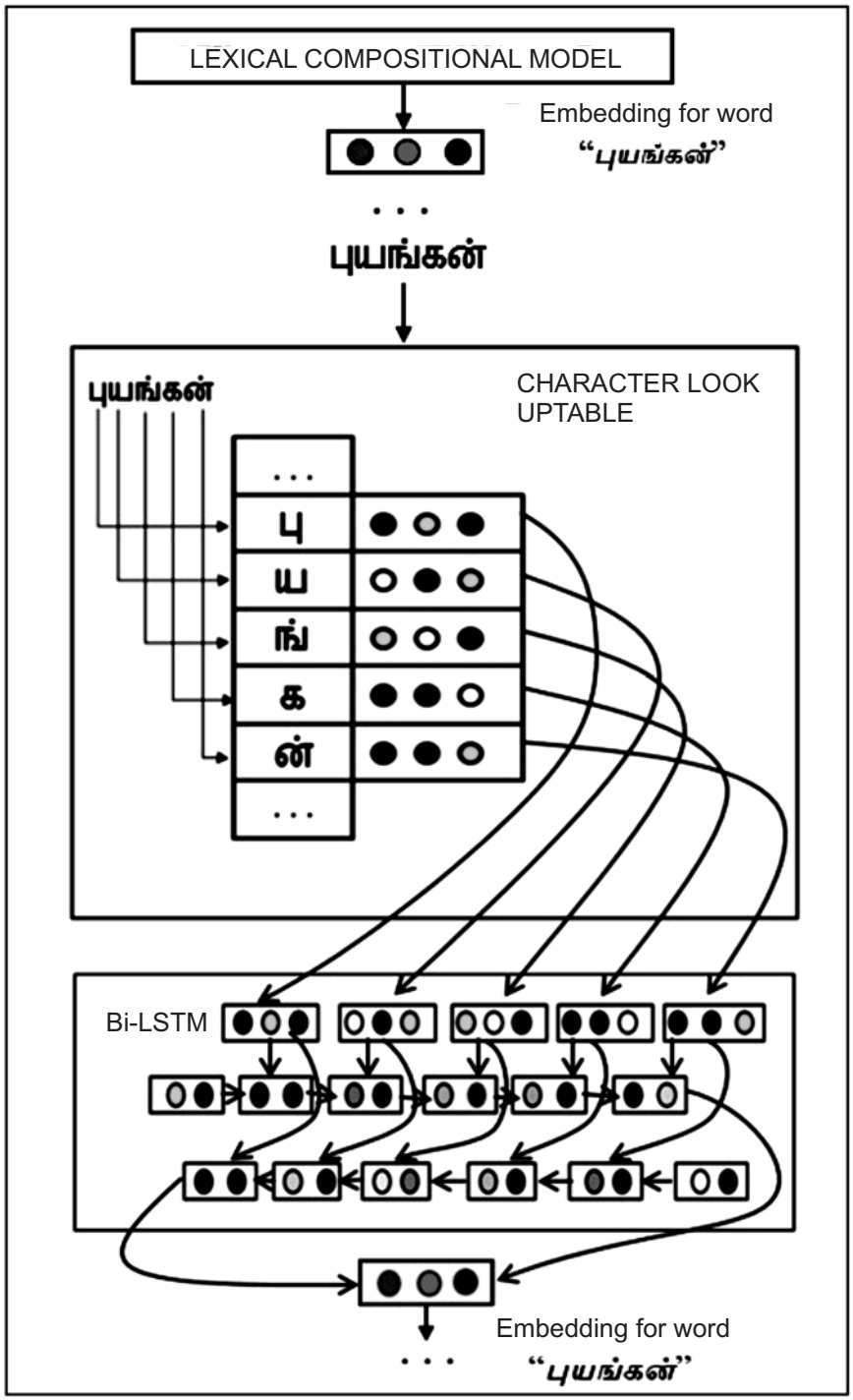

Figure 4: Illustration of our C2W model

7. REFERENCES

1. Hasan, Fahim Muhammad, Naushad UzZaman, and Mumit Khan. ”Comparison of different POS Tagging Techniques (N-Gram, HMM and Brills tagger) for Bangla.” Advances and Innovations in Systems, Computing Sciences and Software Engineer- ing. Springer Netherlands, 2007. 121-126.
2. Antony, Santhanu Mohan, and Soman. ”SVM Based Part of Speech Tagger for Malayalam.” Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference on. IEEE, 2010.
3. Lakshmana Pandian,and Geetha. ”Morpheme based Language Model for Tami.l Part-of-Speech Tagging.” polibits 38 (2008): 19-25.

4. Ling, Wang, et al. ”Finding function in form: Compositional character models for open vocabulary word representation.” arXiv preprint arXiv:1508.02096 (2015).
5. Wang, Peilu, et al. ”Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network.” arXiv preprint arXiv:1510.06168 (2015).
6. Lee, Yoong Keok, Aria Haghighi and Regina Barzilay. ”Simple Type-Level Unsuper- vised POS Tagging.” in Proceedings of the EMNLP 2010: Conference on Empirical Methods in Natural Language Processing, Oct. 9-11, 2010, MIT, Massachusetts, USA.
7. Sgaard, Anders. ”Semisupervised condensed nearest neighbor for part-of-speech tag- ging.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011.
8. Anand Kumar, et al. ”Factored Statistical Machine Translation System for English to Tamil Language.” Pertanika Journal of Social Sciences & Humanities 22.4 (2014).
9. Brill, Eric. “A simple rule-based part of speech tagger.” Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.
10. Dhanalakshmi, et al. “Tamil Part-of-Speech tagger based on SVM- Tool.”Proceedings of the COLIPS International Conference on natural language processing (IALP), Chiang Mai, Thailand. 2008.