NITK-IT_NLP@NSURL2019: Transfer Learning based POS Tagger for Under Resourced Bhojpuri and Magahi Language

Anand Kumar M
Department of Information Technology
National Institute of Technology Karnataka (NITK), Surathkal
Mangalore, India
m_anandkumar@nitk.edu.in

Abstract

Part-of-Speech (POS) tagging is the primary step in the language processing task and also known to perform well automatically with a massive amount of training data. But the POS annotated training data are not available for most of the languages. The languages which do not have sufficient resources to build statistical Natural Language Processing (NLP) applications are called low-resource languages. This paper presents the machine learning-based POS tagging for low resource languages Bhojpuri and Maghali. The work is submitted to the Shared task on Low-level NLP Tools for Bhojpuri Language and Magahi Language at NSURL 2019. We develop a traditional feature-based SVM method and transfer learning-based sequence tagger using new BERT embedding, which enables better generalization to unseen words and provides regularization. The results with given minimal amounts of POS annotated data on Bhojpuri and Maghali languages show that our proposed architecture outperformed the results of the other participants and achieved the new state-of-the-art POS tagger.

1 Introduction

Part-of-Speech tagging is one of the essential stages in language processing applications. POS tagger and tagged corpus are necessary for natural language processing (NLP) to support advanced researches such as parsing, language translation, and speech recognition. If languages consist of considerable resources in terms of data, then the less engineering of hand-crafted rules is enough for robust and better performance. At the same time, the existing NLP tools are trained over large annotated corpora using machine learning techniques. But these resources are not available for most of the languages. Usually, the languages that have received relatively less attention from NLP are less popular due to their lack of available resources and are often called low-resource languages. In this work, we present methods for automatically building a POS tagger for low-resource language Bhojpuri and Maghali with minimal need for human annotation. It is difficult for researchers to produce significant resources for low-resource languages without continuous funding.

Bhojpuri is a less-resourced Indo-Aryan language of the Asian continent spoken by the western Bihar and eastern Uttar Pradesh of India and the Terai region of Nepal. Bhojpuri is socio-linguistically considered one of the Hindi dialects[1]. Magahi language is also known as Magadhi, is a language spoken in Bihar, West Bengal and Jharkhand states of India. It is also an under-resourced language and has a vibrant and old tradition of folk songs and stories[2].

There are presently between six and seven thousand languages spoken in the world (Lewis, 2009; Nettle, 1998; Wagner et al., 1999), but research in Natural Language Processing (NLP) focuses on only a small number of language. The number of internet users in a country is proportional to the regional language usage and resources available. The development of NLP applications of low-resource languages helps to increase the Internet usage of the particular region.

Research into language-independent NLP methods is desperately needed because they are appropriate in low-resource settings, and such techniques easily applied to many low-resource languages at once. The under-resourced languages can use unsupervised learning, transfer learning, and joint multilingual or polyglot learning for building NLP applications. Unsupervised feature

[1] https://en.wikipedia.org/wiki/Bhojpuri_language

[2] https://en.wikipedia.org/wiki/Magahi_language

extraction and clustering approaches used in the first learning model to build Statistical NLP applications for less-resourced languages. The variations of transfer learning include cross-lingual transfer learning, zero-shot learning, and one-shot learning (Tsvetkov). Cross-lingual transfer learning converts the resources and models from the resource-rich source language to under-resourced target language. Zero-shot learning trains a model in one domain and conceives that it generalizes in the other domain of under-resourced languages. One-shot learning trains a model in one domain and uses only a few examples from an under-resourced domain to adapt it. Transfer learning, unfortunately, only works well for closely related languages. Joint learning of resource-rich and resource-poor languages tried to provide universal representation for languages.

2 Related Works

For resource-poor languages, Feldman, Hana, and Brew (Feldman et al., 2006; Hana et al., 2004) described a method for creating taggers by combining a POS tagger and morphological analyzer. The POS tagger and morphological analyzer for closely-related source languages are helped to produce the tools for a low-resource target language. The drawback of this approach is that it is unfortunately applicable for closely related languages. Das and Petrov (Das and Petrov, 2011) proposed a new cross-lingual tagging using projected tags, and these tags are regularized using graph-based label propagation. Cross-lingual projection annotation model uses parallel corpora to bootstrap a POS tagging process without significant annotation efforts for a less-resourced language. Word-alignment (Nichols and Hwa, 2005; Yarowsky et al., 2001), and word-embedding (Adams et al., 2017) models used in bilingual and multilingual-based tagging where at least there is one resource-rich language which can help in numerous borrowings. Garrette et al. (Garrette and Baldridge, 2013; Garrette et al., 2013) explored building automatic POS taggers from tag dictionaries which created using human annotators. Unsupervised models have received perhaps the most attention for POS tagging (Johnson, 2007). The main difficulty with this unsupervised model is evaluation, where the induced word clusters and gold POS tag classes (Christodoulopoulos et al., 2010) need to compare quantitatively. SVMs widely applied for Indic language processing tasks like POS tagging, Chunking, and Morphological processing (Dhanalakshmi et al., 2009; Velliangiri et al., 2010).

BERT stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), which is devised to pre-train deep bidirectional representations from an unlabeled corpus by combining both left and right context in all layer. It has achieved significant progress in transfer learning for natural language understanding using the transformer architecture. The Bhojpuri POS tagged data (Singh and Jha, 2015) has been developed by using BIS guidelines. POS tagger, monolingual corpus, and Morphological Analyser are also available for Magahi language (Kumar et al., 2016). The Magahi corpora were created from blogs and stories and annotated using BIS tagset (Kumar et al., 2014).

3 POS tagging for Bhojpuri and Magahi

In the NSURL shared task, we have developed two different methods for POS tagging the Bhojpuri and Magahi languages. This section explains the data set description and the detailed methodology developed for the shared task.

3.1 Data set description

Table.1 shows the statistics of the Bhojpuri and Magahi POS data set given by the task organizers. Both language sentences were POS tagged using the Bureau of Indian Standards (BIS) annotation scheme, which is a common standard of annotation for Indian languages. Compared with the Bhojpuri tagset, Magahi consists of more tags. Bhojpuri words tagged with Fine-grained tags and Magahi words annotated with course-grained tags. The Bhojpuri language contains a more average number of words per sentence compared with the Magahi language.

3.2 Methodology

The organizers give sequence labeled POS training data in the word per line fashion. Test data provided in the same format without the POS labels. We have used two different methods to develop the POS tagger for Bhojpuri and Magahi. Figure 1 shows the methodology of the proposed model. The first method based on the common features with the Support Vector Machine (SVM) classi-

Table 1: Data set Description

Data set Description	Bhojpuri		Magahi	
	Train	Test	Train	Test
Number of sentences	4500	532	4575	604
Number of tokens	94686	10582	61431	8204
Avg Sentence length	21.04	19.89	13.43	13.58
POS-Tag set size	33		18	

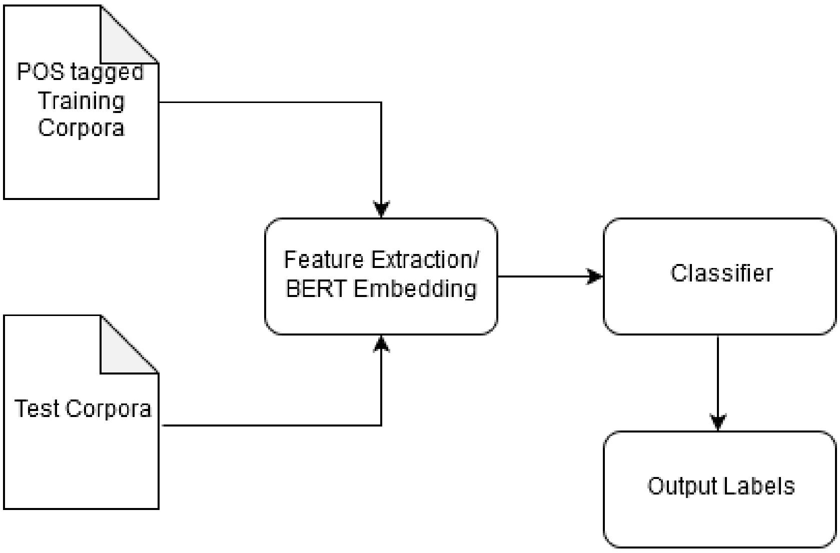
POS tagged Training Corpora
Test Corpora
Feature Extraction/ BERT Embedding
Classifier
Output Labels

Figure 1: Methodology.

fier. This model experiment with combinations of the character bigrams, trigrams, 4-grams, 5-grams, and the full word as features. We have also considered the previous two words and the next two words as additional features. These proposed features can extract the salient features from the text. We have used the SVMLight tool (Giménez and Márquez, 2004; Joachims, 1999) for classifying and tuned the C parameter values based on cross-validation.

In the transfer-learning based method, we have used the BERT multilingual pre-trained embedding in the initial layer. The BERT embeddings consider each token and the sentence from the data set and assign the contextual representation for each token. The logits layer used in the last neuron layer of neural network for the classification task. The parameters settings for BERT given as follows, 12-layer, 768-hidden, 12-heads, 110M parameters, batch_size=8, Adam Optimizer and, Learning rate = 0.0001 with final Cross-Entropy Loss.

3.3 Experiments and Results

The parameters of the learning models are fixed using standard validation techniques. For tuning the SVM parameters, we have used 10-fold cross-validation. In the case of transfer learning randomly selected 10 percent of the training data are considered as validation data and the accuracies are reported in Table 2. Table 3 and 4 show the accuracies of the developed POS tagger achieved on the shared task. We have submitted our runs in the team name of ”NITK-IT_NLP” and ”SUB1” represents the conventional feature-based SVM classifier and ”SUB2” refers to the transfer learning model.

From the accuracy tables, it is clear that the SVM based method worked perfectly for Bhojpuri, and the transfer learning model worked well for the Magahi language. Interestingly, the accuracies are indirectly proportional to the tagset size of the language (Usually, the accuracy is comparably less for the language with fine-grained tagset i.e. Bhojpuri language). If we compare the accuracies of both languages, the method which gives good accuracy for one language is provides less accu-

Table 2: Validation Accuracy

Methods	Bhojpuri	Magahi
WordFeat+CharFeat+SVM (10 Fold)	94.38	80.11
TransferLearning-BERT(Random10Per)	90.1	89.0

racy in another language. It is right in vice-versa also. The reason is the number of tags in the tagset and an average number of words in a sentence.

Table 3: Accuracy of Magahi Language

Rank	Team / Run	F1 Score
1	NITK-IT_NLP_SUB2	0.79
2	the_illiterati_mag_1	0.77
2	the_illiterati_mag_2	0.77
3	the_illiterati_mag_3	0.74
4	NITK-IT_NLP_SUB1	0.73

4 Conclusion and Future Scope

Most of the research in Computational Linguistics and NLP focuses on languages that have a massive amount of text corpora. State-of-the-art NLP models also require large amounts of training data from which it can learn parameters and better co-efficient for the machine learning model. Under-resourced languages or less-resourced languages are languages which are lacking large digital text and insufficient handcrafted linguistic resources for building statistical NLP applications. Here we have presented the two POS tagging approaches developed and submitted for the Shared task on Low-level NLP Tools for Bhojpuri Language and Magahi Language at NSURL 2019. The sequence labeling formulation methods acted as a benchmark for fully supervised POS tagging. The proposed SVM based and transfer learning-based models outperform the other submissions by the participants and achieved the new state-of-the-art. It proves the need for transfer learning to the under-resourced languages. Detailed error analysis and tag specific accuracy are the other directions of future research. The research efforts exist that explore which type of linguistic features in the language and other rich-resourced languages contribute to accurate part-of-speech tagging for the low resourced languages under investigation.

Table 4: Accuracy of Bhojpuri Language

Rank	Team / Run	F1 Score
1	NITK-IT_NLP_SUB1	0.95
1	the_illiterati_bho_3	0.95
2	the_illiterati_bho_1	0.93
3	the_illiterati_bho_2	0.92
4	NITK-IT_NLP_SUB2	0.89

Acknowledgment

We want to thank the shared task organizers for organizing the workshop for under-resourced languages.

References

O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn. 2017. Cross-lingual word embeddings for low- resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, pages 937–947.

C. Christodoulopoulos, S. Goldwater, and M. Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584. Association for Computational Linguistics.

D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. *in ACL, pp*, pages 600–609.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. preprint, arXiv.

V. Dhanalakshmi, P. Padmavathy, Anand Kumar, Soman M., Rajendran K. P., and S. Chunker for Tamil. 2009. Chunker for tamil. In *ARTCom 2009 - International Conference on Advances in Recent Technologies in Communication and Computing*.

A. Feldman, J. Hana, and C. Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. *in Proceedings of LREC, pp*, pages 549–554.

D. Garrette and J. Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. *in Proceedings of NAACL-HLT, pp*, pages 138–147.

D. Garrette, J. Mielens, and J. Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. *ACL*, 1:583–592.

Jesús Giménez and Lluís Márquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

J. Hana, A. Feldman, and C. Brew. 2004. A resource-light approach to russian morphology: Tagging russian using czech resources. *in EMNLP, pp*, pages 222–229.

T. Joachims. 1999. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

M. Johnson. 2007. Why doesn't em find good hmm pos-taggers? *in EMNLP-CoNLL, pp*, pages 296–305.

R. Kumar, B. Lahiri, and D. Alok. 2014. Developing lrs for non-scheduled Indian languages. In J. Mariani, editor, *Vetulani Z.* Human Language Technology Challenges for Computer Science and Linguistics. LTC 2011. Lecture Notes in Computer Science, vol 8387. Springer, Cham.

R. Kumar, Atul Kr. Ojha, B. Lahiri, and D. Alok. 2016. Developing resources and tools for some lesser-known languages of india. *Regional ICON(regICON)*, 2016.

M. P. Lewis. 2009. *Ethnologue: Languages of the world sixteenth edition*. ethnologue. com, Dallas, Tex SIL International. Online version.

D. Nettle. 1998. Explaining global patterns of language diversity. *Journal of anthropological archaeology*, 17(4):354–374.

C. Nichols and R. Hwa. 2005. Word alignment and cross-lingual resource acquisition. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 69–72. ACL.

S. Singh and G. N. (2015 Jha. 2015. Statistical tagger for bhojpuri (employing support vector machine). *In Advances in Computing, Communications and Informatics(ICACCI)International Conference*, pages 1524–1529.

Yulia Tsvetkov. Opportunities and challenges in working with low-resource languages. *Slides Part-1*.

D. Velliangiri, M. Anand Kumar, R. U. Rekha, K. P. Soman, and S. Rajendran. 2010. Grammar teaching tools for tamil language. *2010 International Conference on Technology for Education*, 4.

D. A. Wagner, R. L. Venezky, and B. V. Street. 1999. *Literacy: An international handbook*. Westview Press Boulder.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *01, Association for Computational Linguistics, Stroudsburg, PA, USA, p*, pages 1–8, HLT. Proceedings of the First International Conference on Human Language Technology Research.