Evaluating Performance of Pre-trained Word Embeddings on Assamese, a Low-resource Language

Dhrubajyoti Pathak, Sukumar Nandi, Priyankoo Sarmah
Centre for Linguistic Science and Technology
IIT Guwahati, North Guwahati, India, 781039
drbj153@iitg.ac.in, sukumar@iitg.ac.in, priynakoo@iitg.ac.in

Abstract

Word embeddings and Language models are the building blocks of modern Deep Neural Network-based Natural Language Processing. They are extensively explored in high-resource languages and provide state-of-the-art (SOTA) performance for a wide range of downstream tasks. Nevertheless, these word embeddings are not explored in languages such as Assamese, where resources are limited. Furthermore, there has been limited study into the performance evaluation of these word embeddings for low-resource languages in downstream tasks. In this research, we explore the current state of Assamese pre-trained word embeddings. We evaluate these embeddings' performance on sequence labeling tasks such as Parts-of-speech and Named Entity Recognition. In order to assess the efficiency of the embeddings, experiments are performed utilizing both ensemble and individual word embedding approaches. The ensembling approach that uses three word embeddings outperforms the others. In the paper, the outcomes of the investigations are described. The results of this comparative performance evaluation may assist researchers in choosing an Assamese pre-trained word embedding for subsequent tasks.

Keywords: Word embedding, Assamese text processing, POS tagger, NER tagger, Word embedding evaluation

1. Introduction

Deep Neural Networks (DNN) are a crucial component of modern Natural Language Processing (NLP). Word embedding is vital in a deep learning-based model. Contextualized word embeddings (Devlin et al., 2019; Akbik et al., 2018; Mikolov et al., 2013; Pennington et al., 2014) have led to considerable advances across various NLP tasks. In certain areas of NLP, they perform nearly as well as humans. These word embeddings are language-dependent and were trained on a large unlabeled dataset. Real-valued vector representations of the words within a sentence provide the capability to convey contextual information. It produces high-quality word representations for resource-rich languages. On the other hand, learning high-quality representations is difficult for languages with limited resources. The size of the text corpus for a language with limited resources is insufficient to train word embeddings that capture both semantic and syntactic meaning. The lack of linguistic resources is a significant challenge in studying NLP tasks in low-resource languages.

Assamese is a highly inflectional Indo-Aryan language with a rich morphology. It is one of the scheduled languages of the Indian Constitution and is spoken primarily in Assam, a state in north-east India. Assamese has 15 million native speakers and 7.5 million second-language speakers. Although Assamese literature has a rich literary history, recent advances in NLP are still understudied. Due to the scarcity of language resources, it gets less attention in the NLP research community. In the study, we observed that pre-trained Assamese word embedding models are not explored in any downstream NLP tasks such as Parts of Speech (POS) labeling, Named Entity Recognition (NER), question answering, etc.

This paper aims to assess the existing pre-trained word embedding in two downstream sequence labeling tasks: POS and NER. In this study, we conduct an empirical comparison of word embeddings utilized in Assamese sequence labeling tasks. We explore eleven word embeddings that have been found to achieve SOTA performance in downstream tasks in resource-rich languages. We develop POS and NER labeling models utilizing the Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF) architecture (Huang et al., 2015; Akbik et al., 2019).

Our contributions can be summarised as follows-

1. We explore the available Assamese pre-trained word embeddings.
2. We report an in-depth assessment of word embedding performance in the sequence labeling task.
3. The embedding models are evaluated using three approaches: individual, stacked with two embeddings, and stacked with three embeddings.
4. The best-performing POS and NER models are made available to the research commu-

LREC-COLING 2024, pages 6418–6425

nity [1].

The paper is organized as follows: Section 2 provides an overview of the Assamese language. We provide a brief summary of the several word embeddings that we employed in the evaluation experiment in Section 3. Section 4 provides a brief overview of the dataset used to train various models. The sequence modeling architecture and training configuration are detailed in Sections 5 and 6, respectively. All the analyses and experiments that were carried out to evaluate the various word embeddings are described in Section 7. Lastly, we conclude the paper in section 8..

2. Assamese: A morphologically rich language

Rich in morphological characteristics, Assamese (ISO 639-3) is a low-resource and highly inflectional language. Assamese, an Indo-Aryan language, is mostly spoken in the state of Assam, located in the northeastern region of India. It has around 25 million (Caswell, 2022) native speakers. Examples of a sentence in Assamese-

মাজুলী বিশ্বৰ সৰ্ববৃহৎ নদীদ্বীপ
majuli world largest river island
majuli bisvɔr sɔrbabrɪhɔt nɔdidvip
Majuli is the largest river island in the world

In the Assamese language, words can be classified into two distinct classes according to their morphological characteristics. These are- (a) the indiclinable অব্যয় /ɔbyɔy/, and (b) the inflected সব্যয় /sɔbyɔy/ (Pathak et al., 2023). When used in a sentence, /ɔbyɔy/ does not undergo any morphological alterations, such as আৰু /ɑru/ (and), যদি /jɔdi/ (if).

The other category of word, sɔbyɔy, undergoes changes in its morphological structure due to the addition of different affixes to the root word. Affixes significantly influence word construction in Assamese. Assamese employs an extensive number of suffixes that are appended to the end of words, in contrast to English, which mostly uses word order to communicate grammatical meaning. These suffixes convey grammatical features such as case (subject, object, etc.), tense (past, present, future), and plurality. A word's class may change following suffixation in a sentence. Table 2 demonstrates how suffixation alters the class of a word. In the example in Table 2, the word চৰ /sɔr/ meaning "slap" initially serves as a noun. However, when it is suffixed with –আ /ɑ/ (চৰা), it changes its class and becomes a verb, similarly, in the other two examples, upon suffixation, an adjective (ৰঙা /rɔŋa/ *'red'*) transforms into an adverb (ৰঙাকৈ) and an adjective (কোমল /komɔl/ *'tender'*) into a noun (কোমলতা *'tenderness'*).

In the instance of NER, a word may be classified into multiple categories based on how it is used in a sentence (Pathak et al., 2022). For example, মানস /manɔs/ is a boy's name with the label PERSON. The name refers to both a river and a national park in Assam, which is labeled as LOCATION. The name also refers to the sacred lake *Mansarovar* (/manɔs sɔrovɔr/) on Kailash Mountain, which is also a LOCATION. The word মানস can also be used as a NOUN to convey desire, wish, or something prayed for. Additionally, the NER labeling process is further complicated by challenges such as Nested Entities, the Agglutinative nature of the language, and the lack of capitalization for a noun word. The complexity introduced by these linguistic attributes impacts the performance of POS or NER labeling.

3. Word Embeddings

Modern natural language processing relies heavily on word embedding, which encodes words as numeric vectors. These vectors convey the meaning and context of word(s), so related words have similar vector representations. This enables computers to comprehend the associations that exist between words and to perform downstream tasks such as classification, machine translation, sentiment analysis, and question answering more efficiently.

There are two distinct types of word embeddings: contextual and non-contextual. Non-contextual word embedding focuses solely on the words or subwords within a word or phrase in order to capture both the syntactic and semantic meaning. Conversely, contextual word embeddings consider not just the individual word or character but also the context in which it appears. There may

[1] https://anonymous.4open.science/r/eval-asm-embed-854E/
[2] https://github.com/stanfordnlp/GloVe
[3] https://fasttext.cc/docs/en/pretrained-vectors.html
[4] https://github.com/bheinzerling/bpemb
[5] https://www.cfilt.iitb.ac.in/~diptesh/embeddings/
[6] https://huggingface.co/bert-base-multilingual-cased
[7] https://tinyurl.com/XLM-R-Embed
[8] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md
[9] https://indicnlp.ai4bharat.org/indic-bert/
[10] https://huggingface.co/google/muril-base-cased

Table 1: Details of word embeddings for Assamese used in our experiment

Word Embeddings	Trained Corpus
WordEmbeddings (Glove) (Pennington et al., 2014) [2]	Wiki FastTextEmbeddings (Bojanowski et al., 2017) [3] Wiki
Byte Pair (Heinzerling and Strube, 2018) [4]	Wiki
ELMO Embedding (Peters et al., 2018) [5]	Wiki +ILCI Dataset
mBERT Embedding (Devlin et al., 2018) [6]	Wiki
XLM-R Embedding (Conneau et al., 2020) [7]	CommonCrawl
FlairEmbeddings (Akbik et al., 2018) [8]	Website: `jw.org`
IndicBERT (Kakwani et al., 2020) [9]	`Scraping`
MuRIL (Khanuja et al., 2021) [10]	`CommonCrawl` + Wiki

Table 2: Class change after affixation

চৰ /sɔr/ *'slap'* (Noun) + —আ /ɑ/ (suffix) → চৰা /sɔrɑ/ *'slap'* (Verb)
ৰঙা /rɔŋa/ *'red'* (Adjective) + —কৈ /kɔi/ (suffix) → ৰঙাকৈ /rɔŋakɔɪ/ *'in red'* (Adverb)
কোমল /komɔl/ *'tender'* (Adjective) + —অতা /ɔtɑ/ (suffix) → কোমলতা /komɔlɔtɑ/ *'tenderness'* (Noun)

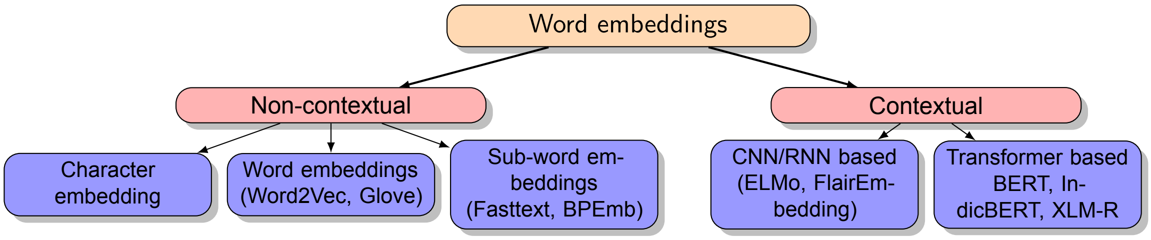

Figure 1: Categories of Word embeddings

be cases when the vector representation of a word changes based on its position in the sentence.

In our experiment, we employ pre-trained Assamese word embeddings that are publicly available. Table 1 lists all of the pre-trained word embeddings for Assamese as well as the size of their training corpus. The multilingual corpus is used to train these Assamese word embeddings. In the majority of instances, the corpus is taken from the Wikipedia dump. In our analysis, we observe that, in comparison to other Indian languages, the majority of the Assamese word embeddings that have already been trained have been developed using a very small set of corpus.

4. Dataset

Training a DL-based sequence model needs a large, annotated dataset. Conversely, annotated datasets for low-resource languages are scarce due to the tedious and time-consuming method of creating a dataset that is suitable for training. Additionally, the verification of the annotated dataset involves a substantial investment of linguistic resources and the involvement of language experts in the field. During our study of the relevant literature, we came across that there is just one POS annotated dataset that is available to the general public.

We obtained the manually labeled dataset (ILCI-II, 2020) from the Technology Development for Indian Languages (TDIL), Government of India. The dataset is available at ILCI-II (2020). The BIS-tagged Assamese dataset comprises original Assamese writings from several disciplines, including agriculture, art and culture, business, education, entertainment, geography, history, literature, philosophy, public administration, religion, and sports. The BIS tagset has been named the official standard for annotating data in Indian languages. The word count in the POS datasets is 404k words, and there are 35k sentences. The dataset has a total of forty-one (41) tags, which are categorized into eleven (11) top-level categories. Table 3 provides details about the tagset.

The AsNER dataset is employed for NER training (Pathak et al., 2022). With five entity classes, the AsNER dataset comprises approximately 99k tokens. The details on the entity classes are presented in Table 4. The annotated dataset was organized using the column format specified by CoNLL-2003 (Sang and De Meulder, 2003). Each line in the column contains a single word accompanied by the corresponding POS or NER tag, which is separated by a tab space. In order to facilitate training, the dataset was initially sampled at

Table 3: POS tagset (ILCI-II, 2020)

Sl .No	Category	Type	Tag
1	Noun	Abstract	N_ANN
		Common	N_CNN
		Noun (Location)	N_NST
		Material	N_MNN
		Proper	N_NNP
		Verbal Noun	N_VNN
		Noun (unclassified)	N_NN
2	Pronoun	Indefinite	PR_PRI
		Personal	PR_PRP
		Reciprocal	PR_PRC
		Reflexive	PR_PRF
		Relative	PR_PRL
		Wh-words	PR_PRQ
3	Adjective	Adjectival Adverb	J_JJ
		Verbal	J_VJJ
		Proper	J_PJJ
4	Demonstrative	Deictic	DM_DMD
		Indefinite	DM_DMI
		Relative	DM_DMR
		Wh-words	DM_DMQ
5	Verb	Auxiliary	V_VAUX
		Main	V_VM
		Transitive	V_VBT
		In-transitive	V_VBI
6	Adverb		RB
7	Conjunction	Conjunction	CC_CCD
		Co-ordinator	CC_CCS
8	Particles	Classifier	RP_RPD
		Interjection	RP_INJ
		Intensifier	RP_INTF
		Negation	RP_NEG
		Particles (unclassified)	SUF
9	Quantifiers	General	QT_QTF
		Cardinals	QT_QTC
		Ordinals	QT_QTO
10	Post Position		PSP
11	Residuals	Foreign word	RD_RDF
		Echowords	RD_ECH
		Punctuation	RD_PUNC
		Symbol	RD_SYM
		Unknown	RD_UNK

Table 4: Details of Entity classes

S. No	Entity Name	Tag
1	Location (regions name, street name, natural locations name, etc.)	LOC
2	Person (names of people, animals, fictional characters, etc.)	PER
3	Organisation	ORG
4	Miscellaneous (includes a broad category such as nationalities, languages, events name, etc.)	MISC
5	Numbers (numbers, money, percentage, and quantity)	NUM
6	Others (not fall in any of the above categories)	O

Table 5: Dataset statistics

Dataset	POS	NER
Train	320599	81422
Dev	39865	8292
Test	40125	8909

random and subsequently divided into three parts: 80% for the training phase, 10% for the development phase, and 10% for the test phase. The dataset's statistics are presented in Table 5. The NER dataset is available at Pathak et al. (2022).

5. Sequence Labelling Architecture

We employ a state-of-the-art neural sequence labeling model that utilizes the classic BiLSTM-CRF architecture (Huang et al., 2015), FLAIR (Akbik et al., 2019) to train the sequence labeling model. The BiLSTM-CRF architecture has been shown to achieve state-of-the-art performance in various downstream tasks such as NER, POS labeling, and chunking, especially for resource-rich languages such as English, German, Spanish, and Dutch respectively (Akbik et al., 2018; Peters et al., 2018). Hence, we employ this framework to conduct experiments pertaining to word embeddings in tasks involving sequence labeling.

6. Training Setup

Nvidia Tesla P100 403 GPU (3,584 Cuda Cores) is utilized in the training of the models. Throughout the training, we followed the hyperparameters recommended by (Reimers and Gurevych, 2017). The hyperparameters are summarized in Table 6. The early stopping technique is implemented when the accuracy of the validation data does not improve. The learning rate annealing technique is employed as well to increase performance while reducing training time. The POS labeling model requires an average of five hours for training and testing, whereas the NER labeling model needs just three hours for training and testing.

7. Experiment result and Analysis

In this section, the results of the experiments that were conducted using ten distinct pre-trained Assamese word embeddings on sequence labeling tasks are presented. In the individual approach, each embedding is used independently of one another. The training process consists of three sets of runs with identical hyper-parameters. The F1-score for POS and NER labeling is reported in Table 7. Three separate iterations of the tests are carried out for both the POS and NER labeling. Subsequently, the mean value of the F1-scores is calculated and listed for each embedding. All contextual embeddings exhibit significantly higher F1-scores compared to the non-contextual ones. With an average F1-score of 0.8156 and 0.7894, respectively, MuRIL embedding in POS and NER labeling performs better than the others.

In the subsequent experiments, we employed the ensemble approach, which enables the concatenation (stacking) of several embeddings to embed the words in a training sentence. The word embedding that performed best (MuRIL) in the individual approach is chosen for use in the ensemble approach (two embeddings). Table 8 summarizes MuRIL's performance when combined with various word embeddings. The labeling F1-score is substantially enhanced when MuRIL is concatenated with other embeddings. The performance of non-contextual embeddings is significantly improved when used in combination with MuRIL. The F1-score obtained from MuRIL embedding with Character Embedding is 0.8387, which is higher than the top F1-score of 0.8236 achieved by the individual approach for POS labeling. In the NER labeling ensembling approach, the XLM-R with MuRIL embedding achieves the highest score of 0.7935, which is nearly similar to the best F1-score (0.793) in the individual method.

To further investigate the efficacy of the ensembling approach, we employed three-word embeddings in the third set of experiments. According to (Akbik et al., 2018), this ensembling approach of three word embeddings performs best for English, with a score of 0.9309. On the basis of the higher efficiency in the ensembling method for two word embeddings, the configuration (MuRIL + Character Embedding) is selected for the ensembling of three word embeddings. Table 9 summarises the results of combining the performance of three-word embeddings. The F1-scores achieved in our experiment are 0.8407 and 0.9098, which are the highest in both POS and NER.

The following analysis can be drawn from the experiment of different word embeddings on sequence labeling-

- Contextual word embeddings outperform non-contextual embeddings in both sequence labeling tasks.
- MuRIL demonstrates superior performance in sequence labeling for the Assamese language when compared to all other word embeddings. It is important to mention that the MuRIL training corpus is the largest (Ref. 1) among all training corpora. This indicates that the size of the corpus is an important factor to consider when training word embedding models.
- The combined application of pre-trained Assamese word embeddings has been found to improve their performance in sequence labeling tasks. In other words, the stacking approach increases the performance of sequence tagging even when used in languages

Table 6: Hyper-parameters

Size of Hidden layer	RNN layer	Word dropout	Mini-batch size	learning rate	Epochs	Sequence length
512 (POS) and 1024 (NER)	1	0.05	32	0.01	100	128

Table 7: Sequence labeling performance in individual method

Embeddings	POS				NER			
	Run 1	Run 2	Run 3	**Mean**	Run 1	Run 2	Run 3	**Mean**
Character Embeddings	0.5563	0.5603	0.5337	0.5501	0.6001	0.5986	0.5805	0.5931
Glove	0.563	0.5502	0.5878	0.567	0.6788	0.5429	0.6051	0.6089
IndicBert	0.6453	0.7896	0.7566	0.7307	0.6583	0.6434	0.6607	0.6541
FastTextEmbeddings	0.7936	0.7851	0.7894	0.7893	0.6794	0.6782	0.6701	0.6759
mBERT	0.7880	0.7997	**0.8164**	0.8014	0.7737	**0.7902**	0.7792	0.7810
XLM-R	0.8129	0.8069	0.7899	0.8032	0.6942	0.6331	0.6812	0.6695
ELMO	0.8109	0.7521	0.7733	0.7788	0.7181	0.7223	07043	0.7149
Byte Pair	0.814	0.7765	0.7896	0.7934	0.7588	0.762	0.7451	0.7553
FlairEmbeddings	0.8172	**0.8144**	0.8021	0.8112	0.6828	0.7195	0.7112	0.7045
MuRIL	**0.8236**	0.8099	0.8132	**0.8156**	**0.793**	0.7843	**0.791**	**0.7894**

Table 8: Sequence labeling performance in ensemble method (Two embeddings)

Stacked Embeddings	**POS**	**NER**
MuRIL + Glove	0.8295	0.7772
MuRIL + FastTextEmbeddings	0.8000	0.5061
MuRIL + Byte Pair	0.8203	0.7756
MuRIL + Character Embeddings	**0.8387**	0.7788
MuRIL + mBERT	0.8237	0.7647
MuRIL + ELMO	0.8338	0.7537
MuRIL + XLM-R	0.8274	**0.7935**
MuRIL + IndicBert	0.8312	0.7681
MuRIL + FlairEmbeddings	0.8294	0.7772

Table 9: Sequence labeling performance of word embeddings in ensemble method (Three embeddings)

Stacked Embeddings	**POS**	**NER**
MuRIL + Character Embedding + Glove	0.8288	0.8259
MuRIL + Character Embedding + Fasttext	0.8306	0.8402
MuRIL + Character Embedding + Byte Pair	0.8317	**0.9098**
MuRIL + Character Embedding + mBERT	0.8274	0.8513
MuRIL + Character Embedding + ELMO	0.8292	0.6456
MuRIL + Character Embedding + XLM-R	0.8284	0.8794
MuRIL + Character Embedding + FlairEmbeddings	**0.8407**	0.8091

with limited resources. Non-contextual embeddings, particularly Character Embeddings, perform significantly better in the ensemble approach.

- It has been observed that the performance of some combinations of word embedding in the Stacked method drops when compared to the performance in the individual method. This is due to “overfitting”. Sometimes, the more embeddings we use, the greater the chance that the model learns something that is too specific and does not generalize well.

8. Conclusion

The paper presents an extensive evaluation of the performance of Assamese pre-trained word embedding in the context of sequence labeling tasks. We focused on recent embeddings that have achieved SOTA performance in downstream tasks. There were two approaches that were employed during the training process: the individual approach and the ensemble approach. We observe a performance enhancement when employing the ensemble method, in which one embedding is combined with others. According to our best knowledge, this is the first study that has been conducted to investigate the efficiency of pre-trained Assamese word embeddings in sequence labeling tasks. We believe that this experiment will assist researchers in selecting word embeddings for sequence labeling tasks in low-resource languages such as Assamese.

9. Bibliographical References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Isaac Caswell. 2022. Google translate learns 24 new languages. Last accessed 19 October 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

JNU ILCI-II. 2020. Assamese monolingual text corpus ilci-ii.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. AsNER - annotated dataset and baseline for Assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577, Marseille, France. European Language Resources Association.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2023. Part-of-speech tagger for assamese using ensembling approach. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(10).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings* *of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

10. Language Resource References

ILCI- II, JNU. 2020. *Assamese Monolingual Text Corpus ILCI-II*. Technology Development for Indian Languages (TDIL), Government of India.

Pathak, Dhrubajyoti and Nandi, Sukumar and Sarmah, Priyankoo. 2022. *AsNER - Assamese Named Entity Dataset*. European Language Resources Association.