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

LREC-COLING 2024
Torino, Italy

Dhrubajyoti Pathak, Prof. Sukumar Nandi & Prof. Priyankoo Sarmah

Centre for Linguistic Science and Technology

Indian Institute of Technology Guwahati

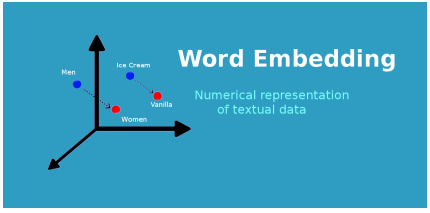
20-25 May, 2024

Contributions of the paper

- Exploration of pre-trained word embeddings for low-resource language Assamese.
- We report an in-depth assessment of word embedding performance in the sequence labeling task.
- The embedding models are evaluated using three approaches: individual, stacked with two embeddings, and stacked with three embeddings.
- The best-performing POS and NER models are made available to the research community

Word embeddings

Word embedding in NLP

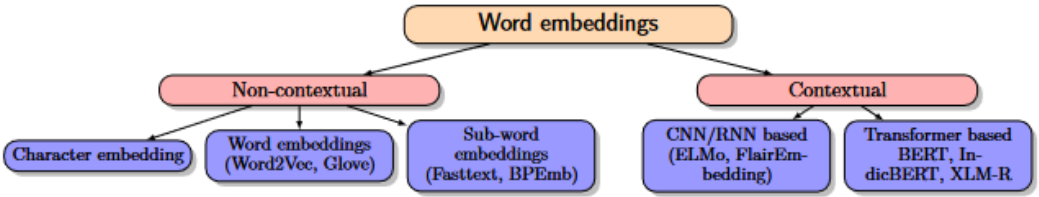


Word embedding

- Word embeddings are a numerical representation of textual data.
- Help machines figure out what words mean.
- Word embedding is vital in a deep learning-based model in Natural Language Processing (NLP).
- Word embeddings are language-dependent and were trained on a large unlabeled dataset.

Applications of word embeddings

Types of word embeddings



Word embeddings and low-resource language

- To produce high-quality word representations for a languages requires a large corpus.
- Learning high-quality representations is difficult for low-resource languages (LRL)
- LRLs often have limited amounts of raw text data as well as annotated data.
- Due to limited resources, LRLs attract minimal interest from researchers. Even rarer are those who work in the area of performance evaluation using deep learning models.

Assamese language

Assamese Language

- Assamese or Asamiya, pronounced, /ɔxɔmijɑ/ is a morphologically rich, Indo-Aryan language spoken in Assam, a state of northeast India.
- A scheduled languages of India, spoken by more than 15 million [Cen11] people.
- Highly inflectional and morphologically rich language.
- Although Assamese has a very old and rich literary history, technology development in NLP is still in a nascent stage.
- Due to the scarcity of language resources, it gets less attention in the NLP research community.

Assamese Language

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

Objective

- Explore pre-trained word embeddings for Assamese that have been found to achieve SOTA performance in downstream tasks in resource-rich languages.
- Evaluation of the existing pre-trained word embedding for Assamese
- Evaluation to be conducted in two downstream sequence labeling tasks: POS and NER.

Word Embeddings that cover Assamese I

- Classic Word Embeddings
  - **Glove** [PSM14]: Distinct word has precisely one pre-computed embedding. Trained on Wikipedia, CommonCrawl and Twitter corpus.
  - **Word2Vec** [MCCD13]: Trained on Google News corpus comprises about 6B tokens.
- Sub-word Embeddings
  - **FastText embeddings** [BGJM17]: Trained on Common Crawl and Wikipedia corpus covers 157 languages.
  - **Byte-Pair Embeddings** [HS18] : Trained on Wikipedia for 275 language,

Word Embeddings that cover Assamese II

- Contextualized word embeddings
  - **Multilingual Bidirectional Encoder Representations from Transformers (BERT)** [DCLT18] : Pre-trained model trained on 104 languages including Assamese.
  - **Multilingual Representations for Indian Languages (MuRIL)** [KBM+21]: Based on BERT base architecture pre-trained on 17 Indian languages, including Assamese.
  - **XLM-RoBERT (XLM-R)** [CKG+20]: Multi-lingual pre-trained language model, trained on CommonCrawl data.
  - **IndicBERT** [KKG+20]: ALBERT-based language models for 11 languages trained on IndicCorp dataset.

Word Embeddings that cover Assamese III

- Contextualized word embeddings
  - **Embeddings from Language Models (ELMo)** [PNI+18]: Trained on Wikipedia (1.9B) and monolingual news crawl data from WMT 2008-2012 (3.6B).
  - **Flair Embedding** [ABV18]: A Contextual string embeddings. The multilingual Flair embedding covers 343 languages, trained on JW300 corpus.

Training corpus used in pre-trained word embedding for Assamese

Word Embeddings	**Trained Corpus**
WordEmbeddings (Glove)	Wiki
FastTextEmbeddings	Wiki
Byte Pair	Wiki
ELMO Embedding	Wiki +ILCI Dataset
mBERT Embedding	Wiki
XLM-R Embedding	CommonCrawl
FlairEmbeddings	Website: `jw.org`
IndicBERT	`Scraping`
MuRIL	`CommonCrawl` + Wiki

Evaluation experiment

Modeling POS and NER labeling employing word embeddings

Dataset

- POS labeling: 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).
- Forty-one (41) tags and eleven (11) top-level categories
- NER labeling: AsNER dataset is employed for NER training [PNS22].
- With five entity classes, the dataset comprises 99k tokens.

Dataset statistics

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

Training Details

- Hidden size of 512 (POS) and 1024 (NER)
- Maximum sequence length of 128
- Mini batch size of 32
- Trained for 100 epochs on Nvidia Tesla P100 GPU, with 128Gb RAM.
- 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.

Word embedding performance in POS labeling 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**

Word embedding performance in POS labeling in stacked 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

Word embedding performance in POS labeling in stacked 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

Experiment Outcome

Word embedding in individual approach

- Three separate iterations of the tests are carried out for both the POS and NER labeling.
- Contextual embeddings exhibit significantly higher F1 scores compared to the non-contextual ones.
- The highest mean F1-score of 0.8156 and 0.7894, respectively, for POS and NER by MuRIL embedding.

Word embedding in ensemble approach

- 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).
- Performance gets 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+ Character Embedding is 0.8387, which is higher than the top F1-score of 0.8236 in the individual approach for POS labeling.
- In NER labeling, 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.

Analysis

Analysis I

- Contextual word embeddings outperform non-contextual embeddings in both sequence labeling tasks.
- MuRIL demonstrates superior performance in sequence labeling for the Assamese language compared to all other word embeddings. MuRIL training corpus is the largest among all training corpora. This indicates that the size of the corpus is an important factor to consider when training word embedding models.
- The ensembling of pre-trained Assamese word embeddings has been found to improve their performance in sequence labeling tasks.
- The ensembling approach increases the performance of sequence tagging even when used in languages with limited resources.

Analysis II

- Non-contextual embeddings, particularly Character Embeddings, perform significantly better in the ensemble approach.
- Overfitting: The performance of some combinations of word embedding in the Stacked method drops when compared to the performance in the individual method.
- Sometimes, the more embeddings we use, the greater the chance that the model learns something that is too specific and does not generalize well.

Conclusion

Conclusion

- The paper presents an extensive evaluation of the performance of Assamese pre-trained word embedding in the context of sequence labeling tasks.
- Two approaches were employed during the training process: the individual approach and the ensemble approach.
- Observe a performance enhancement when employing the ensemble method, combining one embedding with others.
- First study that has been conducted to investigate the efficiency of pre-trained Assamese word embeddings in sequence labeling tasks.

References I

[ABV18] Alan Akbik, Duncan Blythe, and Roland Vollgraf, *Contextual string embeddings for sequence labeling*, Proceedings of the 27th international conference on computational linguistics, Association for Computational Linguistics, 2018, pp. 1638–1649.

[BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, *Enriching word vectors with subword information*, Transactions of the Association for Computational Linguistics **5** (2017), 135–146.

[Cen11] Census, *ABSTRACT OF SPEAKERS' STRENGTH OF LANGUAGES AND MOTHER TONGUES - 2011*, Accessed April, 2011.

[CKG+20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, *Unsupervised cross-lingual representation learning at scale*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 8440–8451.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).

[HS18] Benjamin Heinzerling and Michael Strube, *BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan) (Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, eds.), European Language Resources Association (ELRA), May 7-12, 2018 2018 (english).

References II

[KBM+21] 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, *Muril: Multilingual representations for indian languages*, 2021.

[KKG+20] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar, *inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 4948–4961.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781 (2013).

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, *Deep contextualized word representations*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 2227–2237.

[PNS22] Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah, *AsNER - annotated dataset and baseline for Assamese named entity recognition*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (Marseille, France) (Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, eds.), European Language Resources Association, June 2022, pp. 6571–6577.

References III

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning, *GloVe: Global vectors for word representation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar), Association for Computational Linguistics, October 2014, pp. 1532–1543.

Thank You!