Proceedings of the WILDRE-6 Workshop @LREC2020, pages 1–6
Marseille, 20 June 2022

Introducing EM-FT for Manipuri-English Neural Machine Translation

Rudali Huidrom and Yves Lepage
Waseda University
Kitakyushu, Japan
{rudali.huidrom@ruri, yves,lepage@}waseda.jp

Abstract

This paper introduces pretrained word embeddings for Manipuri, a low-resourced Indian language. The pretrained word embeddings based on fastText is capable of handling the highly agglutinative language Manipuri (mni). We then perform machine translation (MT) experiments using neural network (NN) models. In this paper, we confirm the following observations. Firstly, the reported BLEU score of the Transformer architecture with fastText word embedding model EM-FT performs better than without in all the NMT experiments. Secondly, we observe that adding more training data from a different domain of the test data negatively impacts translation accuracy. The resources reported in this paper are made available in the ELRA catalogue to help the low-resourced languages community with MT/NLP tasks.

Keywords: neural machine translation, low resource language, language technology

1. Introduction

Manipuri, a highly agglutinative low-resourced Indian language from the Sino-Tibetan language family, is reported as an extremely low-resourced language for MT/NLP tasks and developing an MT system for the already available size of data is a true challenge (Huidrom and Lepage, 2020). Manipuri is a morphologically-rich in nature. We introduce the creation of a pretrained word embedding model for Manipuri based on fastText that we call **EM-FT** (meaning, a **F**ast**T**ext word embedding model exclusively trained on Manipuri from the EM Corpus (Huidrom et al., 2021)) especially, for use in MT of Manipuri-English language pair.

The objective of this paper is to introduce word embeddings (Mikolov et al., 2013b; Peters et al., 2018) during MT. We trained the word embeddings using the monolingual Manipuri (locally known as Meiteilon) data of 1.88 million sentences from the EM Corpus. In particular, introducing this embeddings help in initialization and/or transfer learning for NLP/MT tasks and in cross-lingual transfer (Kakwani et al., 2020) for learning multilingual embeddings. The quality of the embeddings depends on the size of the monolingual data (Mikolov et al., 2013a; Bojanowski et al., 2017) which is a resource that is not widely available for many languages. In our case, the monolingual data of 1.88 million sentences is one of the largest among the available monolingual corpora of Manipuri.

In this paper, we report experiments for MT using the PMIndia data set and the *EM Corpus*. Furthermore, we report experiments on MT where we introduce our pretrained embeddings (*EM-FT*) during training.

The main contributions of our work are as follows:

- This is the first work to create a pretrained word embedding model trained from comparatively large Manipuri monolingual data and introduce it during MT of Manipuri–English language pair.
- We have shown the impact in the differences of the domains used for test data and training data in our experiments on MT.

The structure of the paper is as follows. Section 2 describes previous work. Section 3 gives details about the data set used. Section 4 presents the methodology. Section 5 describes the experiments, their results and provides an analysis. Section 7 concludes and proposes future directions.

2. Related Work

Word embeddings (Mikolov et al., 2013b; Peters et al., 2018) are a way of representing words in real-valued vector representations (Ortiz Suárez et al., 2019) and it can encode morphology, semantics, syntax etc. to some extent and it provides transfer learning in NLP tasks. Some examples of word embeddings are word2vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). These models are context-free in nature: a given word has a single vector representation independent of context. Some of the existing word embeddings trained for Indian languages are The Polyglot (Al-Rfou' et al., 2013), IndicFT (Kakwani et al., 2020) and, fastText (Bojanowski et al., 2017; Mikolov et al., 2018; Grave et al., 2018) trained on Wikipedia and Common Crawl data. None of these includes Manipuri trained on large corpora or even a limited corpus.

Although NMT models are trained end-to-end with continuous representations that mitigate the sparsity problem in comparison to the rigid SMT architectures (Koehn et al., 2003), NMT (Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) possess a risk for low translation accuracy for low-resourced languages of small amounts (Koehn and Knowles, 2017). However, it is to note that SMT is still superior in training for corpus which are not big enough.

Data set	Language pair	sentence pairs	words / sent.	word types
PMIndia	Manipuri	7,419	15	22,289
	English		19	18,502
EM Corpus	Manipuri	124,975	21	74,516
	English		26	64,501

Table 1: Statistics on the data set used.

Our work consists of creating a pretrained word embeddings for Manipuri, which is one of a kind available for this low-resourced Indian language. We analyse how introducing our word embeddings to the NMT model during its training and further adding sentences from EM Corpus to the training set from base data set impact the translation quality and its results. In this paper, we propose to introduce word embeddings for Manipuri, EM-FT and study the impact of it.

3. Dataset

In this section, we discuss the nature of the low-resourced language, Manipuri, and the data set in use for our experiments, namely the PMIndia data set and the EM Corpus.

3.1. Manipuri (Meiteilon)

Manipuri, also known as Meiteilon, is an Indian language from the Sino-Tibetan language family. It follows the SOV (Subject-Object-Verb) syntax structure. Manipuri is predominately spoken in the Indian state Manipur with about two million native speakers. Manipuri is classified as 'vulnerable language' by UNESCO (Moseley and Nicolas, 2010), it is one of the two Indian languages listed in the 8th Schedule of the Indian Constitution as endangered.

Manipuri has two writing systems: Eastern Nagari Script (also known as the Bengali Script) and Meitei Mayek. We use Manipuri written in Eastern Nagari Script for all of our works. Again, Manipuri is a low-resourced language that has not been explored much in computational linguistics. One of the reasons being the limited amount of available resources. In this paper, we aim to bridge this gap by sharing our resources publicly.

3.2. PMIndia dataset

The PMIndia data set[1] (Haddow and Kirefu, 2020) is used as our base data set and with EM Corpus for MT. This data set (monolingual and parallel corpora) contains the official documents from the Prime Minister Office of the Government of India. There are 13 Indian languages and English in it. We use the parallel corpora of the Manipuri–English language pair for our experiments. Statistics about the data are described in Table 1. There are 7,419 sentences in parallel for the Manipuri–English language pair with the average number of words per sentence in Manipuri and English as 15 and 19. The available number of sentences in the monolingual corpora for Manipuri reported as 41,699 sentences.

3.3. EM Corpus

The Ema-lon Manipuri Corpus (translation: our mother tongue Manipuri Corpus), abbreviated as the EM Corpus[2] (Huidrom et al., 2021) is a comparable corpus created by collecting news articles daily from a newspaper website known as "The Sangai Express," which is available in both languages. An average of 14,000 sentences is crawled for this language pair daily. The reported data is being collected from August 2020 to March 2021. The domain of the EM Corpus (Rudali Huidrom and Yves Lepage, 2021) includes general articles, news on state, national and international affairs, sports and entertainment news, and the editorial.

The monolingual and parallel corpora contain 3.33 million and 124,975 sentences in total for both languages, along with the number of words per sentence in Manipuri and English to be 21 and 26. It is to note that the number of word types in each language reflects the number of sentences and the structure of the language: it is natural that the more the sentence pairs, the higher the number of word types as reported in Table 1.

4. Methodology

Our first series of experiments introduces the creation of a pretrained word embedding model exclusively for Manipuri. In particular, we chose fastText as it is capable of integrating subword information using character n-gram embeddings during training (Kakwani et al., 2020). Some of the previously published results on pretrained word embeddings suggests that fastText performs better than word-level algorithms like GloVe, word2vec for morphologically rich languages (Mikolov et al., 2013b; Pennington et al., 2014). We trained skipgram models on the monolingual data of EM corpus of 1.88 million sentences of Manipuri. We introduced the pretrained word embeddings during the training of the NMT models.

Our second series of experiments introduces the performance of MT task by iteratively increasing the amount of training data. We use 5,000 sentences from our base data set, PMIndia data set in the NMT models in the first iteration. It is the baseline of our experiment. The training data other than the base data set is created by adding 10,000 sentences each sampled from the EM corpus to its previous iteration. All our models are validated and tested on 1,000 sentences each from the PMIndia dataset. The training data from the base data set share the same domain with the test and validation data set.

[1] https://data.statmt.org/PMIndia/

[2] http://catalog.elra.info/en-us/repository/browse/ELRA-W0316/

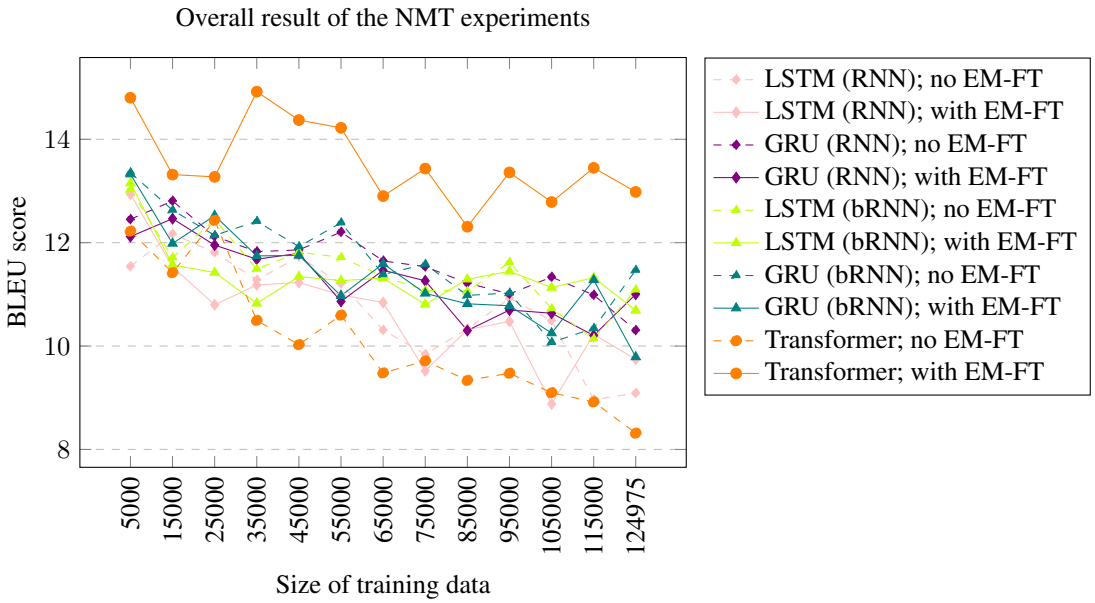


Figure 1: shows the result from the NMT experiments. All the experiments with s EM-FT are represented by a solid line otherwise, a dashed line. The results are color-coded and marked by different shapes for NMT architecture with different RNN type and/or different encoder & decoder respectively.

5. Experiments and results

5.1. EM-FT: fastText word embeddings

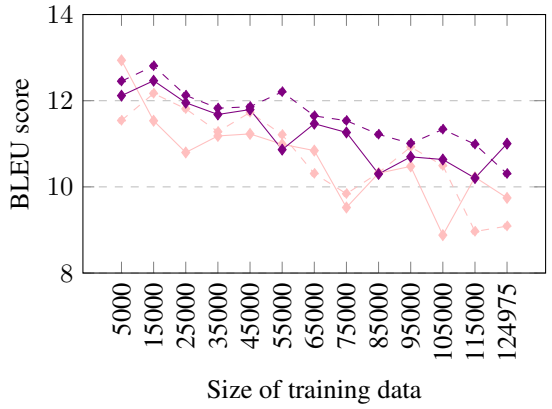
We train fastText (Bojanowski et al., 2017) word embeddings for Manipuri. In context to similar works, one of the most notable examples is when (Mikolov et al., 2018) first trained fastText in English using Common Crawl and for other 157 languages (Grave et al., 2018). However, there are no word embeddings reported for Manipur so far until our work. We train 300-dimensional word embeddings on our monolingual corpus obtained from the EM corpus. In particular, we chose fastText as it is capable of integrating subword information using character n-gram embeddings during training, and fastText is said to perform well among other word embeddings for morphologically rich languages (Kakwani et al., 2020). We trained skipgram models on our monolingual data of 1.88 million sentences of Manipuri for five epochs with a maximum and minimum character length as 2 and 5, size of the context window as 5 and 5 negative examples for each instance.

5.2. Machine Translation

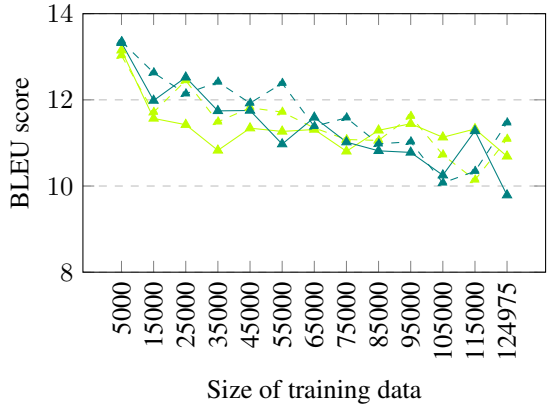
To provide validation in the corpus, we perform NMT on the PMIndia data set as a base data set and later on, PMIndia data set + EM Corpus. As mentioned in 4, as the first series of experiments, MT is performed by further adding 10,000 sentences each sampled from the EM corpus to the base data set for every iteration and are validated and tested on 1,000 sentences of validation and test data set from the PMIndia data set. It is to note that the training data other than the base data set belong to a different domain. In the second series of experiments, we introduce our Manipuri word embeddings EM-FT during the training in the NMT models. We use Joint Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to address the problem of rare words by using sub-word segmentation. We apply BPE on all of our selected data set with 30,000 merge operations to obtain a vocabulary representation of the Manipuri–English language pair. For all of our NMT experiments, we use the OpenNMT-py toolkit (Klein et al., 2017). We preprocess the training and validation data set for the Manipuri–English language pair after applying BPE. We train our model on a 2-layered RNN model with a bidirectional RNN as encoder and a simple RNN as a decoder, and on simple RNN as encoder and decoder. In addition, we train our model on a 2-layered Transformer architecture (Vaswani et al., 2017). We later introduce fastText word embeddings for Manipuri during the training of NMT models. We also train MT models on SMT architecture (Koehn et al., 2003) to validate our data set. We measure the translation accuracy of all our experiments using BLEU with confidence at 95 % (Koehn, 2004).

- **Model Configuration.** Firstly, we choose the default seq2seq architecture with attention mechanism (Luong et al., 2015) provided by OpenNMT-py toolkit (Klein et al., 2017) where both the encoders and decoders are LSTM cells (Hochreiter and Schmidhuber, 1997). Most of the hyperparameters are the default ones provided by the toolkit. Secondly, we choose the Transformer architecture (Vaswani et al., 2017). (Lakew et al., 2018) reports that the transformer architecture usually outperforms the recurrent ones in all their

(a) Expt. results for RNN architecture (1)



(b) Expt. results for RNN architecture (2)



(c) Expt. results for Transformer architecture

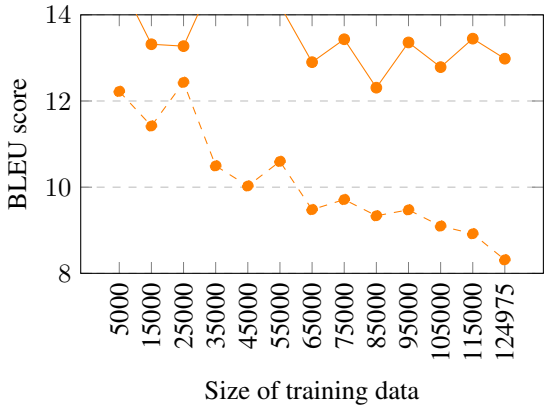


Figure 2: This figure is extracted from Figure 1. The results in (a) are from the NMT experiments for RNN architecture (diamond) with RNN Type as LSTM (pink) and GRU (violet) of encoder type and decoder type as RNN The results in (b) are from the NMT experiments for RNN architecture (triangle) with RNN Type as LSTM (lime) and GRU (teal) of encoder type and decoder type as bRNN & RNN. The results in (c) are from the Transformer architecture. Experiments with s EM-FT are represented by a solid line otherwise, a dashed line.

systems.

- **Training Settings.** The hyper-parameters are uniform throughout all the models in our experiments. Firstly, we train the NMT model on a two-layered RNN model having a layer size of 64 for embeddings and 500 for inner layers. The LSTM is trained on encoder type as bidirectional RNN and decoder as a simple RNN, and on simple RNN as encoder and decoder. We also use the general typed global attention mechanism onto the RNN models. The models are trained with 10,000 training steps with checkpoints at every 5,000 steps and a drop-out (Srivastava et al., 2014) rate of 0.3 (Gal and Ghahramani, 2016) across all the NMT experiments. For optimization of the model during training, we use the Adam (Kingma and Ba, 2015) optimizer with a learning rate of 0.001. The number of steps set before dropping the learning rate is 50,000. The decay frequency is the number of steps at which the learning rate starts to drop at each training step taken is 10,000. Secondly, for the Transformer, we used the recommended setting by the OpenNMT-py toolkit, except for the learning rate of 0.001.

6. Results and Analysis

In this particular experimental setting for validating and testing against training data from different domain, it is observed that Transformer with our word embeddings EM-FT outperforms the rest. See results in Figure 1.

As we progress with the adding more training data, we observe a decrease in the BLEU score which is expected. It is to be noted that the decrease is not linear in nature. The data that we add other than the base data set are obtained from the news crawls which are not standardised translated data. Although, the sentences are aligned, the parallel sentences are not exact translations of one another, instead comparable. Our validation and test data are from the PMIndia (Haddow and Kirefu, 2020) dataset whose domain is the official documents from the Prime Minister Office of India. In most of the experiments, the sudden increase of the BLEU score could be the result of seeing similar sentences crawled from the news articles related to the Prime Minister Office while training.

The NMT system reported in this paper did not perform well for Transformer architecture without EM-FT. It is followed by RNN architecture of RNN type as LSTM with encoder & decoder as RNN and with encoder & decoder as bRNN & RNN. Another interesting observation is that RNN architecture of RNN type as GRU with encoder & decoder as RNN as well as with encoder & decoder as bRNN & RNN are similar in nature irrespective of the presence of EM-FT. It is expected of RNN architecture with encoder & decoder as bRNN and RNN to perform better than the encoder & decoder as RNN as observed in the case of RNN architecture of LSTM with or without EM-FT.

In most of the experiments on RNN architecture, there

is a sudden change in BLEU score between 45,000 sentences and 65,000 sentences of training data.

7. Conclusion

This work provided an insight into creation of pretrained word embeddings for Manipuri–English language pair and to use it in NMT task. Firstly, we studied the creation of the pretrained word embeddings using fastText, EM-FT for Manipuri. Secondly, we performed MT on these data, given the condition that it is tested and validated on a data set of a completely different domain. All in all, we confirm the following observations.

First, it is observed in Figure 1 that the Transformer architecture with the EM-FT fastText word embeddings model performs better than without in all the NMT experiments.

Second, the impact of differences in domains used for test data and training data is clearly demonstrated in our experiments. Adding more data from a different domain negatively impacts the translation accuracy.

In the future, we would like to inspect the possibility of increasing the size of data from the same domain by using data-augmentation techniques, so as to increase the translation accuracy of NMT with a BERT (Devlin et al., 2018) model for Manipuri.

8. Bibliographical References

Al-Rfou', R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan et al., editors, *Proceedings of Machine Learning Research*, volume 48, pages 1050–1059, New York, New York, USA, 20–22 Jun. PMLR.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Haddow, B. and Kirefu, F. (2020). Pmindia - a collection of parallel corpora of languages of india. *ArXiv*, abs/2001.09907.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780, November.

Huidrom, R. and Lepage, Y. (2020). Zero-shot translation among Indian languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 47–54, Suzhou, China, December. Association for Computational Linguistics.

Huidrom, R., Lepage, Y., and Khomdram, K. (2021). EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67, Online (Virtual Mode), September. INCOMA Ltd.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Koehn, P. (2004). Statistical significance tests for ma-

chine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Lakew, S. M., Cettolo, M., and Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Moseley, C. and Nicolas, A. (2010). *Atlas of the world's languages in danger*. UNESCO, France, 3 edition.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, et al., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Pennington, J., Socher, R., and Manning, C. (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, October. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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, June. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

9. Language Resource References

Rudali Huidrom and Yves Lepage. (2021). *Ema-lon Manipuri Corpus (including word embedding and language model)*. distributed via ELRA: ELRA-Id ELRA-W0316, ISLRN 588-170-827-016-7.