Training Cross-Lingual embeddings for Setswana and Sepedi

Makgatho, Mack
Dept. of Computer Science, University of Pretoria
mack.letladi1@gmail.com

Marivate, Vukosi
Dept. of Computer Science, University of Pretoria
vukosi.marivate@cs.up.ac.za

Sefara, Tshephisho
Council for Scientific and Industrial Research
tsefara@csir.co.za

Wagner, Valencia
Sol Plaatje University
valencia.wagner@spu.ac.za

Abstract

African languages still lag in the advances of Natural Language Processing techniques, one reason being the lack of representative data, having a technique that can transfer information between languages can help mitigate against the lack of data problem. This paper trains Setswana and Sepedi monolingual word vectors and uses VecMap to create cross-lingual embeddings for Setswana-Sepedi in order to do a cross-lingual transfer.

Word embeddings are word vectors that represent words as continuous floating numbers where semantically similar words are mapped to nearby points in n-dimensional space. The idea of word embeddings is based on the distribution hypothesis that states, semantically similar words are distributed in similar contexts (Harris, 1954).

Cross-lingual embeddings leverages monolingual embeddings by learning a shared vector space for two separately trained monolingual vectors such that words with similar meaning are represented by similar vectors. In this paper, we investigate cross-lingual embeddings for Setswana-Sepedi monolingual word vector. We use the unsupervised cross lingual embeddings in VecMap to train the Setswana-Sepedi cross-language word embeddings. We evaluate the quality of the Setswana-Sepedi cross-lingual word representation using a semantic evaluation task. For the semantic similarity task, we translated the WordSim and SimLex tasks into Setswana and Sepedi. We release this dataset as part of this work for other researchers. We evaluate the intrinsic quality of the embeddings to determine if there is improvement in the semantic representation of the word embeddings.

Keywords: cross-lingual embeddings, word embeddings, intrinsic evaluation

1 Introduction

Many African languages have insufficient language resources (data, tools, people) (Abbott & Martinus 2019, Martinus & Abbott 2019, Nekoto et al. 2020, Sefara et al. 2021) and fall into the classification of low resource languages (Ranathunga et al. 2021) in the Natural Language Processing (NLP) field. This lack of resources makes it harder to capitalise on the recent advances in many NLP downstream tasks such as Neural Machine Translation (Cho et al. 2014), Large Language Models (Devlin et al. 2018, Howard & Ruder 2018), Q&A systems (Kwiatkowski et al. 2019), etc. There may be more downstream approaches to deal with some of these challenges such as Transfer Learning (Ruder et al. 2019), Data Augmentation (Marivate & Sefara 2020*a*), Multilingual Models (Hedderich et al. 2020), etc. Additionally, the lack of research attention to existing NLP techniques results in difficulties finding a benchmark (Abbott & Martinus 2019). In this work, we focus on word representations through word embeddings and how we can leverage one language to assist in the representation of another related language. These embeddings can then be used to develop tools for other downstream tasks.

Word Embeddings are a mathematical technique to learn general language vector representations from a large amount of unlabelled text using co-

occurring statistics. In recent years, monolingual word embeddings techniques are increasingly becoming an important resource in NLP. Word embeddings are widely used in NLP problems such as sentiment analysis (Socher et al. 2013), named-entity-recognition (Guo et al. 2014), parts-of-speech tagging, and document retrieval. Word2Vec is a vector training model proposed by Mikolov et al. (2013). Word2Vec produces a low-dimensional real-value vector representing the meaning of a word. The word vector represents grammatical and semantic properties, which results in words with similar semantic relations being close to each other. The word vector representation method incorporates the semantic relationship between words which is not possible through representations such as Bag-Of-Words of TFIDF. Word embeddings are better than both methods because they map all the words in a language into a vector space of a given dimension, the words are converted into vectors and allow multiple linear operations and have the property of preserving analogies (Mikolov et al. 2013, Pennington et al. 2014).

Cross-lingual word embeddings have been receiving more and more attention from the NLP community, mainly because it has provided a path to effectively align two disjoint monolingual embeddings with no bilingual dictionary for unsupervised techniques or no more than a small bilingual dictionary for supervised techniques (Lample et al. 2018, Artetxe et al. 2020). Cross-lingual techniques also enable knowledge transfer between languages with rich resources and low resources. For languages lacking bilingual parallel corpus with other languages, cross-lingual embeddings can be utilised to train high-quality cross-lingual embeddings (Lample et al. 2018). This can aid in accelerating the progress of applying NLP to low-resourced languages. Artetxe et al. (2018) created the cross-lingual unsupervised or supervised word embedding (VecMap library) approach for training cross-lingual word embedding models. The approaches can be used to construct cross-language word vectors with or without a bilingual dictionary.

The majority of South African languages lag bilingual parallel corpus with other languages. In this work, we aim to investigate how cross-lingual embeddings could be used to improve the state of one or both languages. We used data (corpus) from different domains to train Word2Vec and fastText (Bojanowski et al. 2016) monolingual embeddings. When using VecMap, the two embeddings are aligned. VecMap requires two monolingual word vectors from source and target (Artetxe et al. 2018). To evaluate the effectiveness of the cross-lingual embedding for Setswana and Sepedi, we use intrinsic evaluation (Bakarov 2018) through Setswana and Sepedi versions of WordSim (Finkelstein et al. 2001) and Simlex (Hill et al. 2015). This is following on an approach that has been used for Yoruba and Twi (Alabi et al. 2019). We also release the dataset for this benchmark of human semantic similarity task.

This paper is structured as follows; the next section is a review of related work that is done on cross-lingual word vectors. Followed by data collection in Section 3. Section 4 discusses methodology followed to train cross-lingual word vectors using VecMap. The evaluation of the word vectors is discussed in Section 5. Section 5.1 explains the results while Section 6 discusses the findings and finally, conclusions and future work can be found in Section 7.

2 Background and Related Work

Cross-lingual word embeddings (CLWEs) are becoming popular in NLP for two reasons: Cross-lingual word embeddings can transfer knowledge from rich-resourced languages to low-resourced; The technique can also infer the semantics of words in a multiple language environment. Conneau et al. (2018) show that word embeddings spaces can be aligned without any cross-lingual supervision. The alignment is based on solely unaligned datasets of each language. Using adversarial training, they were able to initialise a linear mapping between a source and a target space, which they use to create

a synthetic parallel dictionary. First, they propose a simple criterion that is used as an unsupervised validation matric. Second, they propose the similarity measure cross-domain similarity local scaling (CSLS), which mitigates the hubness problem and increases the word translation accuracy. The hubness problem is defined by Dinu et al. (2015) as:

> "neighbourhoods of the mapped elements are strongly polluted by hubs, vectors that tend to be near a high proportion of items, pushing their correct labels down the neighbour list."

In the work done by Adams et al. (2017), the research looked at applying CLWEs to Yongning Na, a Sino-Tibetan language. The research focused on determining if the quality of CLWEs depends on having large amounts of data in multiple languages and if initialising the parameters of neural network language models (NMLM) can improve language modelling in a low-resourced context. The research scaled down the available monolingual data of the target language to about 1000 sentences. The quality of intrinsic embedding was assessed by taking into consideration correlation with human judgement on the WordSim353 (Finkelstein et al. 2001) test set. They went further to perform language modelling experiments by initialising the parameters for long short-term memory (LSTM) (Hochreiter & Schmidhuber 1997) by training across different language pairs. The research showed that CLWEs are resilient even when target language training data is scaled-down and that initialisation of NMLM parameters leads to good performance.

Artetxe & Schwenk (2019) introduced an architecture that can be used to learn multilingual sentence representations for more than 90 languages. The languages belonged to 30 different families. The research used a single BiLSTM encoder with a shared Byte Pair Encoding (BPE) vocabulary coupled with an auxiliary decoder and trained on parallel corpora. They learn a classifier using English annotated data only and transfer it to any language without modification. The research mainly focused on vector representations of sentences that are general for the input language and the NLP task.

Alabi et al. (2019) worked on massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. Authors compare two types of word embeddings obtained from curated corpora and a language-dependent processing. They move further to collect high quality and noisy data for the two languages. They quantify that improvements that is based on the quality of data and not only on the amount of data. In their experiments, they use different architectures to learn word representations both from characters and surface forms. They evaluate multilingual BERT on a down stream task, specifically named entity recognition and WordSim-353 word pairs dataset.

Feng et al. (2018) investigates a cross-lingual knowledge transfer technique to improve the semantic representation of low-resourced languages and improving low resource named-entity recognition. In their research, neural networks are used to do knowledge transfer from high resource language using bilingual lexicons to improve low resource word representation. They automatically learn semantic projections using a lexicon extension strategy that is designed to address out-of lexicon problem. Finally, they regard word-level entity type distribution features as an external language independent knowledge and incorporate them into their neural architecture. The experiment is done on two low resource languages (Dutch and Spanish) to demonstrate the effectiveness of these additional semantic representations.

Banerjee et al. (2021) show that initialising the embedding layer of Unsupervised Neural Machine Translation (UNMT) models with cross-lingual embeddings shows significant improvements in BLEU score. Authors show that freezing the embedding layer weights lead to better gains compared to updating the embedding layer weights during training. They experimented using Denoising Autoencoder (DAE) and Masked Sequence to Sequence (MASS) for three different unrelated

language pairs (for English-Hindi, English-Bengali, and English-Gujarati). The analysis shows the importance of using cross-lingual embedding as compared to other techniques.

The literature shows that there is a substantial amount of work done on cross-lingual transfer and empirical proof that the method improves the performance of models. The literature does not relay solely on intrinsic evaluation but the solutions are applied to some downstream tasks. In the next section, we detail the data used for conducting experiments.

3 Data collection

Training data is very important for implementing powerful and accurate models, and clean training data can make a difference between a good and great model. The data needs to be very imperative because the quality of the alignment depends on the quality of the monolingual embeddings, i.e. data used to create the initial monolingual embeddings before mapping.

We use data collected from different domains for training word vectors:

- **JW300 bible** (Agić & Vulić 2019): A biblical-domain data set containing parallel corpus for low-resourced languages.
- **Wikipedia**
- **National Centre for Human Language Technology (NCHLT) text corpus** (Eiselen & Puttkammer 2014): The dataset contains clean textual data in Sepedi and Setswana. The data set was constructed by harvesting existing data such as online publications, online news, web crawling and crowd-sourcing.
- **SABC News Data in Setswana and Sepedi** (Marivate et al. 2020, Marivate & Sefara 2020*b*): The data set contains news titles collected from online social media.

National Centre for Human Language Technology (NCHLT) data is used for training monolingual word vectors. For preprocessing, we changed all words to lowercase, removing brackets, digits, punctuations, and white spaces.

Table 1: Corpus size for the Setswana and Sepedi Datasets

	Sepedi	Setswana
Number of tokens:	2133972	3000682
Unique words:	93461	107606

In this section we dealt with how we collected the data used to training our monolingual embeddings for both languages and what approach we took to pre-process the data before training the models. In the next section we discuss the approach taken to train the monolingual embeddings and how VecMap was used to training the cross-lingual embeddings.

4 Training monolingual and cross-lingual embeddings (VecMap)

In this section, we present the methods (frameworks) used to train monolingual and cross-lingual embeddings. We describe the parameters used to train word2Vec and fastText embeddings. We also look into VecMap, the framework that we used to align monolingual embeddings.

CLWEs have proved to perform very well for low-resourced languages. The main idea is to do a cross-lingual transfer from the source language to the target, such that we have a single representation for a pair of languages where semantically similar words are closer to one another. In order to use VecMap two monolingual embeddings are required, we train fastText and word2Vec vectors. We use the following parameters for fastText and word2Vec in Table 2. The definition of the parameters are as follows: skipGram - training method, dim - size of word vectors, minCount - minimal number of word occurrences, ws - size of the context window, and epoch - number of epochs or iterations.

4.1 Word2Vec

The word2Vec (Mikolov et al. 2013) algorithm is a two-layer neural network that vectorises words to

Table 2: Parameters for FastText and Word2Vec

Parameter	Value
skip-gram	true
dim	300
minCount	1
ws	4
epoch	100

processes text. The algorithm takes as input a text corpus and returns feature vectors that represent words in that corpus as a set of vectors. Word2Vec trains words against neighbouring words based on a window size context. It trains the words using two methods: skip-gram or continuous bag of words (CBOW), skip-gram uses a word to predict a target context and CBOW uses context to predict a target word. The experiment uses skip-gram to train monolingual embeddings. We use word vectors that were trained using Word2Vec. These correspond to monolingual embeddings of dimension 300 trained on Sepedi and Setswana corpora.

4.2 FastText

FastText (Bojanowski et al. 2016) is a supervised prediction-based technique based on the word2Vec family of algorithms (Mikolov et al. 2013). It predicts tags through context and represents each word as an *n*-gram of characters, instead of learning vectors for words directly. The fastText model has three layers: input layer, hidden layer, and output layer. Input is a number of words and their *n*-gram features, these features are used to represent a single document. The hidden layer is the superimposed average of multiple feature vectors. The hidden layer solves the maximum likelihood function, then constructs a Huffman tree according to the weights and model parameters of each category, and uses the Huffman tree as the output.

4.3 VecMap

VecMap (Artetxe et al. 2020) is an open-source framework to learn CLWEs written in Python. There are two techniques to do cross-lingual embeddings with VecMap, supervised (recommended if you have a large training dictionary) and unsupervised (recommended if you have no seed dictionary and do not want to rely on identical words). In this work, we align word embedding using VecMap[1]. The approach is fully unsupervised. The steps we followed to build our cross-lingual word embeddings model are motivated by the authors of VecMap Artetxe et al. (2020). The assumption is that we have a monolingual corpus for source and target languages. The word representations is learned independently for each language (monolingual embeddings for each language), and then mapped to a common vector space.

In this section, we presented word2Vec, fastText and VecMap. We also described the parameters used to train word2Vec and fastText embeddings. In the next section, we present experimental results and perform some analyses.

5 Evaluation

We evaluate the quality of Setswana and Sepedi word vector representations on two different benchmarks Simlex and WordSim. The datasets (Simlex and WordSim) contain pairs of Setswana and Sepedi words that have been assigned similarity ratings by humans. They give a similarity score between a pair of words corresponding to their relatedness. Cosine similarity is used to collect a score from the model in order to check how close the score is to the human score, we use Spearman to measure correlation. Spearman index measure the dependence of two variables, the correlation of two statistical variables is evaluated using monotonic equation. We manually translate the WordSim and Simlex word pairs dataset from English into Setswana and Sepedi. We are releasing a dataset of Setswana and Sepedi translated WordSim and Simlex as part of this project at `https://github.com/dsfsi/embedding-eval-data` and archived on Zenodo at `https://zenodo.org/record/5673974`.

Table 3: FastText Monolingual Results

Monolingual fastText	Coverage	Spearman
Sepedi(Simlex)	94.58	40.39
Sepedi(WordSim)	81.29	46.15
Setswana(Simlex)	95.22	33.23
Setswana(WordSim)	95.38	44.80

Table 4: Word2Vec Monolingual Results

Monolingual word2Vec	Coverage	Spearman
Sepedi(Simlex)	79.49	25.96
Sepedi(WordSim)	84.49	23.57
Setswana(Simlex)	95.32	31.52
Setswana(WordSim)	95.38	35.11

5.1 Results

This section presents the results of the experiments conducted to show the efficiency of the proposed technique with a couple of experiments. We first present the monolingual evaluation task for word2Vec and fastText and then present the cross-lingual evaluation task for Setswana and Sepedi. The evaluations of cross-lingual evaluation task is based on two embedding methods fastText and word2Vec.

In Table 3 and Table 4, we show the Spearman's correlation for word vectors trained on fastText and word2vec. The correlation scores calculate the similarity between word vectors. Table 5 and Table 6 scores are obtained from using Setswana and Sepedi monolingual vectors and using VecMap to align the two vectors to the same vector space.

The results at Table 3 and Table 4 show the coverage and Spearman results. Coverage refers to the total number of in vocabulary words (words that are found both in the model and evaluation dataset). We can see that the coverage is lower for word2Vec but a little higher for fastText (we expected coverage for fastText to be 100 percent). The Simlex and WordSim similarity score for monolingual fastText embeddings in Table 3 is higher, this is expected due to the coverage percentage also being very high as compared to the coverage value in Table 4.

Table 5: Word2Vec Crosslingual Results

Monolingual word2Vec	Coverage	Spearman
Setswana-Sepedi(Simlex)	90.76	31.14
Setswana-Sepedi(WordSim)	68.56	40.87

Table 6: FastText Crosslingual Results

Crosslingual fastText	Coverage	Spearman
Setswana-Sepedi(Simlex)	91.19	30.44
Setswana-Sepedi(WordSim)	68.84	36.33

6 Discussion

The main purpose of this research is to show that it is possible to do cross-lingual transfer from the source language to the target. In essence we wanted to check if cross-lingual alignment can improve the word representation for the target language. The results on Table 4 shows that the Spearman's correlation value for the target language when using word2Vec is low, this is also due to coverage percentage, but fastText based-embeddings perform better on Table 3 and has a higher coverage percentage, as stated upove we expected 100 percent coverage. Table 5 shows that we improved the representation of words after cross-lingual alignment for word2Vec based-embeddings. The Spearman's value has increased for both Simlex and Wordsim. We expected to improve the results for fastText embeddings but in this case word2Vec actually yielded better results.

7 Conclusion

In this paper, VecMap was used to align Setswana-Sepedi to the same vector space. Through this work, we wanted to use cross-lingual (VecMap) technique to enable knowledge transfer between languages with rich resources and low resources. The results show that it is possible to align two monolingual embeddings to get cross-lingual embeddings. We mapped Setswana to Sepedi and used Spearman's to check correlation. Interestingly we get different results on fastText and word2Vec-based embeddings though we used the same data to train the embeddings.

In future work, it would be interesting to use the cross-lingual embedding on a downstream task like translation or sentiment analysis specifically for low-resourced languages.

8 Acknowledgements

We would like to acknowledge ABSA for sponsoring the industry chair and it's related activities to the project.

References

Abbott, J. & Martinus, L. (2019), Benchmarking neural machine translation for southern african languages, *in* 'Proceedings of the 2019 Workshop on Widening NLP', pp. 98–101.

Adams, O., Makarucha, A., Neubig, G., Bird, S. & Cohn, T. (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, Long Papers', pp. 937–947.

Agić, Ž. & Vulić, I. (2019), JW300: A wide-coverage parallel corpus for low-resource languages, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3204–3210.
URL: *https://aclanthology.org/P19-1310*

Alabi, J. O., Amponsah-Kaakyire, K., Adelani, D. I. & España-Bonet, C. (2019), 'Massive vs. curated word embeddings for low-resourced languages. the case of yor\ub\'a and twi', *arXiv preprint arXiv:1912.02481* .

Artetxe, M., Labaka, G. & Agirre, E. (2018), A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 789–798.

Artetxe, M., Ruder, S. & Yogatama, D. (2020), On the cross-lingual transferability of monolingual representations, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 4623–4637.

Artetxe, M. & Schwenk, H. (2019), 'Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond', *Transactions of the Association for Computational Linguistics* **7**, 597–610.

Bakarov, A. (2018), 'A survey of word embeddings evaluation methods', *arXiv preprint arXiv:1801.09536* .

Banerjee, T., au2, R. M. V. & Bhattacharyya, P. (2021), 'Crosslingual embeddings are essential in unmt for distant languages: An english to indoaryan case study'.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016), 'Enriching word vectors with subword information', *arXiv preprint arXiv:1607.04606* .

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014), 'Learning phrase representations using rnn encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078* .

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. (2018), 'Word translation without parallel data'.

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

Dinu, G., Lazaridou, A. & Baroni, M. (2015), 'Improving zero-shot learning by mitigating the hubness problem'.

Eiselen, R. & Puttkammer, M. J. (2014), Developing text resources for ten south african languages., *in* 'LREC', pp. 3698–3703.

Feng, X., Feng, X., Qin, B., Feng, Z. & Liu, T. (2018), Improving low resource named entity recognition using cross-lingual knowledge transfer, *in* 'Proceedings of the Twenty-Seventh In-

ternational Joint Conference on Artificial Intelligence, IJCAI-18', International Joint Conferences on Artificial Intelligence Organization, pp. 4071–4077.
URL: *https://doi.org/10.24963/ijcai.2018/566*

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2001), Placing search in context: The concept revisited, *in* 'Proceedings of the 10th international conference on World Wide Web', pp. 406–414.

Guo, J., Che, W., Wang, H. & Liu, T. (2014), Revisiting embedding features for simple semi-supervised learning, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 110–120.

Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U. & Klakow, D. (2020), Transfer learning and distant supervision for multilingual transformer models: A study on african languages, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 2580–2591.

Hill, F., Reichart, R. & Korhonen, A. (2015), 'Simlex-999: Evaluating semantic models with (genuine) similarity estimation', *Computational Linguistics* **41**(4), 665–695.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**, 1735–1780.

Howard, J. & Ruder, S. (2018), Universal language model fine-tuning for text classification, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 328–339.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K. et al. (2019), 'Natural questions: a benchmark for question answering research', *Transactions of the Association for Computational Linguistics* **7**, 453–466.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L. & Jégou, H. (2018), Word translation without parallel data, *in* 'International Conference on Learning Representations'.

Marivate, V. & Sefara, T. (2020*a*), Improving short text classification through global augmentation methods, *in* 'International Cross-Domain Conference for Machine Learning and Knowledge Extraction', Springer, pp. 385–399.

Marivate, V. & Sefara, T. (2020*b*), 'South african news data'.
URL: *https://doi.org/10.5281/zenodo.3668495*

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. (2020), Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi, *in* 'Proceedings of the first workshop on Resources for African Indigenous Languages', pp. 15–20.

Martinus, L. & Abbott, J. Z. (2019), 'A focus on neural machine translation for african languages', *arXiv preprint arXiv:1906.05685* .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F. et al. (2020), Participatory research for low-resourced machine translation: A case study in african languages, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings', pp. 2144–2160.

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, Vol. 14, pp. 1532–1543.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M. & Kaur, R. (2021), 'Neural machine translation for low-resource lan-

guages: A survey', *arXiv preprint arXiv:2106.15115* .

Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. (2019), Transfer learning in natural language processing, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials', pp. 15–18.

Sefara, T. J., Zwane, S. G., Gama, N., Sibisi, H., Senoamadi, P. N. & Marivate, V. (2021), Transformer-based machine translation for low-resourced languages embedded with language identification, *in* '2021 Conference on Information Communications Technology and Society (ICTAS)', IEEE, pp. 127–132.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y. & Potts, C. (2013), Recursive deep models for semantic compositionality over a sentiment treebank, *in* 'Proceedings of the 2013 conference on empirical methods in natural language processing', pp. 1631–1642.