Unsupervised Cross-lingual Word Embedding Representation for English-isiZulu

Derwin Ngomane[1], Vukosi Marivate[1,2,3], Jade Abbott[2,3], Rooweither Mabuya[3,4]
[1]Department of Computer Science, University of Pretoria;
[2]Lelapa AI; [3]Masakhane NLP;
[4]South African Centre for Digital Language Resources, North-West University
derwin.ngomane@gmail.com, vukosi.marivate@cs.up.ac.za,
jade.abbott@lelapa.ai, roo.mabuya@nwu.ac.za

Abstract

In this study, we investigate the effectiveness of using cross-lingual word embeddings for zero-shot transfer learning between a language with an abundant resource, English, and a language with limited resource, isiZulu. IsiZulu is a part of the South African Nguni language family, which is characterised by complex agglutinating morphology. We use VecMap, an open source tool, to obtain cross-lingual word embeddings. To perform an extrinsic evaluation of the effectiveness of the embeddings, we train a news classifier on labelled English data in order to categorise unlabelled isiZulu data using zero-shot transfer learning. In our study, we found our model to have a weighted average F1-score of 0.34. Our findings demonstrate that VecMap generates modular word embeddings in the cross-lingual space that have an impact on the downstream classifier used for zero-shot transfer learning.

1 Introduction

In the development of egalitarian Natural Language Processing (NLP) systems, cross-lingual word embeddings are gaining prominence. According to distributional theory, words that appear in comparable contexts have semantic commonalities (Villegas et al., 2016). As a result, word embeddings have paved the way for NLP technology advancement.

The use of word embedding representation has provided satisfactory performance for many NLP applications and associated applications (Gutiérrez and Keith, 2018). Word embeddings have been used for feature engineering (Tang et al., 2014) and transfer learning purposes (Bataa and Wu, 2019). Translation models, sentiment analysis tasks and classification tasks have all benefited. These improvements have been primarily in high-resource languages, such as English, due to the abundance of labelled corpora in these languages. The lack of labelled corpora in low-resourced languages has come at a disadvantage in advancing NLP technologies within this space.

Many news publications in South Africa are in English even though South Africa has eleven official languages (Marivate et al., 2020). According to the Statistics South Africa (Stats SA) Census 2011 results, isiZulu was the most spoken home language with 22.7% of the population indicating it as a home language (Statistics South Africa (Stats SA), 2012). IsiZulu forms part of the many indigenous languages in South Africa, and belongs to the Nguni family of languages (Dube and Suleman, 2019). The Nguni language family is characterised by their agglutinative morphology (Keet and Khumalo, 2016).

South Africa is a multilingual nation where the large majority of the population have a secondary language that they use on top of their primary languages. Over the post-apartheid years, the country has seen a population growth of citizens that speak English as a second language. This behaviour has also become prevalent for isiZulu, where citizens have added isiZulu as a second language (Posel and Zeller, 2020). This emphasizes the importance of advancing NLP work for low-resourced languages in South Africa.

In this work we attempt to take advantage of a high-resource corpus and a low-resource corpus in order to learn cross-lingual word embeddings for English and isiZulu. The use of cross-lingual word embeddings would allow us to perform model transfer learning from a high resource language to a low-resource language. The undertaking for isiZulu is especially challenging due to the morphological complexity of the language. Hence, we have to a handle words in a manner that can maximise syntactic and semantic representation of both English and isiZulu in the cross-lingual space.

Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), pages 11–17

In this work, we aim to answer the following research questions:

- Can we use monolingual word embeddings for English and isiZulu to create cross-lingual semantic embedding vectors for both languages?
- Can we use zero-shot transfer learning in the cross-lingual space to use an English news article classifier to classify isiZulu articles?

Additionally, we introduce two new datasets:

- the Umsuka English-isiZulu dataset (Mabuya et al., 2021) which is used in the creation of the cross-lingual vectors.
- An isiZulu South African news classification dataset sourced from the South African Broadcasting Corporation (SABC) data

The remainder of the work is organised as follows: Section 2 will discuss the background and prior work. Section 3 describes the methodology used for the VecMap library and zero-shot learning classifier. Section 4 discusses the experiments and results. Finally, concluding remarks and potential next steps for future research are discussed in Section 5.

2 Background & Related Work

The popular methods that have been used to represent tokens as vectors have been Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) and Pointwise Mutual Information (PMI). However, these techniques create sparse and large matrices and also create vectors that do not capture semantics (Villegas et al., 2016).

Innovative advancements have led to the emergence of new techniques using neural networks. These methods include the popular Word2Vec models (Mikolov et al., 2013) and attention based models (Vaswani et al., 2017). These models have achieved state-of-the-art performance on adaption and transfer learning tasks in NLP (Devlin et al., 2018).

The disadvantage of monolingual word embeddings is that they struggle with transferring between languages that are dissimilar (Ruder et al., 2019b). Multilingual word embeddings trained on multiple languages have also shown to perform poorly for low-resource languages (Wu and Dredze, 2020). Approaches for cross-lingual representation learning have been proposed to address this issue (Ruder et al., 2019a). There are two types of cross-lingual word embedding methods: mapping and post-hoc approaches (Ruder et al., 2019a). Initial mapping approaches required the use of a bilingual mapping dictionary between the languages of study, but recent improvements have introduced adversarial methods. Adversarial approaches presume that the input monolingual spaces are isomorphic and hence reducing the requirement of a bilingual mapping dictionary (Ruder et al., 2019a). However, it has been shown that adverserial methods perform poorly when the monolingual embeddings are not isomorphic (Søgaard et al., 2018).

Downstream NLP tasks have been proven to benefit from cross-lingual word embeddings. For example, cross-lingual word embeddings have increased the performance of machine translation (Conneau and Lample, 2019) and Part-of-Speech (POS) tagging (Kim et al., 2017) tasks as a transfer learning strategy.

VecMap has been used to create a mapping between English and Welsh and the resulting embedding were used to train a zero-shot and few-shot learning Welsh sentiment classifier (Espinosa-Anke et al., 2021). In the South African context, VecMap was used to develop a Sepedi-Setswana cross-lingual word embedding and adopted a semantic assessment method to analyse the similarity between pairs of Setswana and Sepedi terms (Makgatho et al., 2021) and used in a noisy, multilingual question-answering challenge to train an LSTM classification model (Daniel et al., 2019).

Zero-shot transfer learning has not been widely used for South African news classification. Marivate et al. (2020) used an annotated corpus of local news items to create a news classifier for Setswana and Sepedi. Instead of using static word embedding as those used by VecMap, using contextual word embeddings from a fine-tuned BERT model has achieved impressive results for a zero-shot learning Named Entity Recognition (NER) task for isiZulu (Wang et al., 2020). However, this work would require significant computing and has not been applied to news classification. To our knowledge, this is the first work that uses cross-lingual word embeddings generated from VecMap to create a news article classifier from English to isiZulu.

3 Methodology

We discuss the methodology for the development of the cross-lingual word embeddings and the downstream classifier for news classification.

3.1 Data collection & Preprocessing

The data used in this study is from the South African Centre for Digital Language Resources repository. We use this data for the purposes of training our cross-lingual embedding model. The data consists only of text data for both English and isiZulu. According to the technical documentation, the monolingual English corpus [1] contains 35 686 791 tokens while the isiZulu corpus [2] contains 451 154 tokens. We will only consider the source categories that are the same between the languages.

We also use the Umsuka English-isiZulu parallel corpus (Mabuya et al., 2021) which is a open-source, high quality isiZulu parallel corpus from a mixture of domains, taking into account both Southern African and international context. Half the dataset was a random sample of the News Crawl dataset which was then translated into isiZulu. The other half of the dataset was sampled from isiZulu newspapers (Isolezwe[3], Ilanga[4] and Ezasegagasini Metro [5] publications), spanning from 2012 to 2016, as well as novels and short stories, which were then translated into English. Professional translators were used to create the dataset. An initial pilot study of 500 sentences was performed with quality assurance done by an isiZulu linguist to ensure that the quality requirements were understood.

Table 1 presents the sources used and the number of tokens. In total, we have 1 320 393 and 2 121 127 tokens for isiZulu and English respectively.

Source	isiZulu	English
Hansard	100 392	1 758 616
Hotel Websites	156 143	197 670
Information Guides	7 658	12 564
Internet	21 001	32 857
Other	82 488	5 415
Umsuka	952 711	114 005

Table 1: Tokens by Source and Language

We eliminate stopwords and remove punctuation in the original corpora [6]. Additionally, we perform lemmatization on the English corpus using the WordNet lemmatizer (Miller, 1995). We trained the news classifier on British Broadcasting Corporation (BBC) news data sourced from Kaggle [7]. There are 1 490 English articles for training and 736 English articles evaluation purposes. We use SABC news articles for the isiZulu evaluation. Two isiZulu speakers annotated 219 SABC news articles, which were reviewed by an isiZulu linguist [8]. Table 2 presents the distribution of the datasets.

Category	BBC English	SABC isiZulu
Sport	346	47
Business	336	25
Politics	274	111
Entertainment	273	36
Tech	261	0

Table 2: Frequency of categories

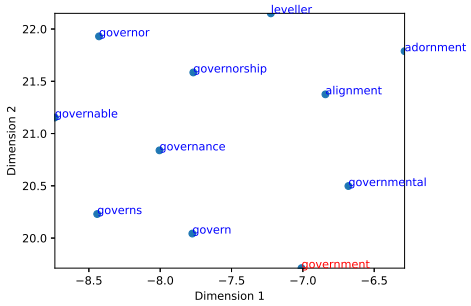
3.2 Monolingual Word Embeddings

The processed corpora that we have described in Section 3.1, are used to develop monolingual embeddings using the Python Gensim module [9]. We make use of the FastText architecture to handle the agglutinative morphology of isiZulu (Bojanowski et al., 2017).

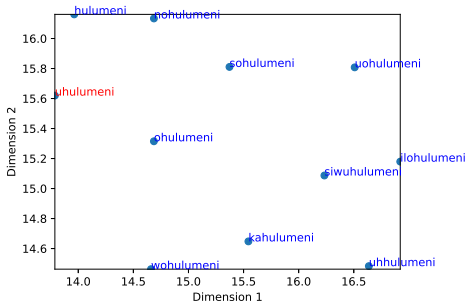
We generate vectors of 64 dimensions for each token in the corpus with a context window size of 3. These hyperparameters were chosen because of the limited vocabulary size of our corpora, and it has been shown that a shorter context window captures the syntactic representation of the word and a larger context window captures more topical representation (Levy and Goldberg, 2014). To demonstrate the learned representation of the corpus we use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to visualise the 10 closest words from both the monolingual corpora. The chosen tokens are "government" for English and the isiZulu translation "uhulumeni". In Figure 1, we observe that the closest tokens to "uhulumeni" are synonyms of "uhulumeni". Similarly,

[1] https://repo.sadilar.org/handle/20.500.12185/466
[2] https://repo.sadilar.org/handle/20.500.12185/338
[3] https://www.isolezwe.co.za
[4] https://ilanganews.co.za/
[5] https://www.durban.gov.za/pages/government/documents
[6] List of isiZulu stopwords: https://github.com/stopwords-iso/stopwords-zu
[7] https://www.kaggle.com/c/learn-ai-bbc
[8] Data available at https://github.com/dsfsi/izindaba-zesizulu and https://zenodo.org/record/5674236
[9] https://radimrehurek.com/gensim/index.html

we also observe the same for “government”.



(a) Representation of “government”



(b) Representation of “uhulumeni”

Figure 1: UMAP representation of the monolingual embeddings

3.3 News Article Classifier

The generated cross-lingual word embeddings are used to train a classifier on the English corpus. We compare four algorithms and assess their performance using cross-validation. The models we train are Naive Bayes, Support Vector Machines, Logistic Regression, and Gradient Boosted Machines. The best model is then used to classify unseen isiZulu news articles in the cross-lingual embedding space.

4 Experiments & Results

In this section of the paper we describe the experiments and outcomes conducted in an effort to answer the research questions. We describe the created cross-lingual word embeddings and the experiments undertaken to develop the news article classifier.

4.1 Cross-lingual Word Embeddings

We use the VecMap library [10] created by Artetxe et al. (2018) to generate the cross-lingual word embeddings. In VecMap, we employ the unsupervised cross-lingual learning method. The algorithm-generated UMAP representation of the 10 closest words to “government” and “uhulumeni” are depicted in Figure 2. Some of the words from the monolingual representations provided in Figure 1 were retained by VecMap. However, the algorithm developed a language-based clustering with the exception of identifying “powerfully” as being closer to isiZulu words.

Modularity is the phenomenon in which language clustering occurs in the cross-lingual space (Fujinuma et al., 2019), as depicted in Figure 2. Fujinuma et al. (2019) claim that cross-lingual embeddings that reflect modularity have a negative effect on downstream tasks.

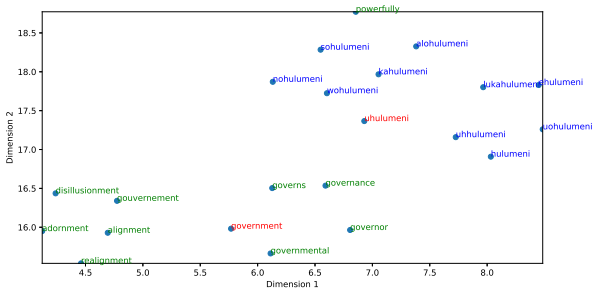


Figure 2: UMAP representation of the VecMap cross-lingual embeddings

4.2 Zero-shot transfer learning

This section presents the zero-shot transfer learning task of identifying isiZulu news headlines from an English news classifier. We initially train our model on labelled English news articles using the pre-generated cross-lingual word embeddings from the previous section.

4.2.1 Experiment Setup

The input data needs to be converted into its vector representation as per the word embeddings from VecMap. Since each token has 64 dimensions, we represent a sentence as the average of all the token vector representations. We use a 64 zero dimensional vector to represent tokens that are out of vocabulary.

We use 75% of the data for training and 25% for evaluation purposes in the English BBC data. In order to select the best model, we run a repeated 5-fold stratified cross-evaluation on the training data. Based on the cross-validation procedure, the

[10] https://github.com/artetxem/vecmap

LightGBM model outperforms all the other models.

4.2.2 Results

The LightGBM model achieves an accuracy of 76.4% on the evaluation set. The classifier achieved an above 70% accuracy for most of the classes, except for entertainment that obtained an accuracy of 68%.

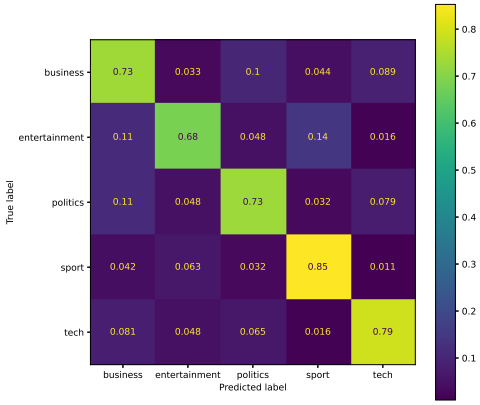


Figure 3: Confusion matrix of the evaluation set

We use the LightGBM model to classify the isiZulu news articles. The resulting performance of the model is presented in Table 3.

Category	Precision	Recall	F1-score
Business	0.07	0.04	0.05
Entertainment	0.15	0.08	0.11
Politics	0.49	0.52	0.51
Sport	0.25	0.32	0.28
Tech	0.0	0.0	0.0
weighted avg	**0.33**	**0.35**	**0.34**

Table 3: Model performance on SABC isiZulu news

The model performed unsatisfactory across all the titles. However, as depicted in Table 3 the model seems to perform better for politics related titles and getting an F1-score of 0.51. Politics contains a larger sample size in our isiZulu data. The other titles had a lower performance, which may be explained by modularity, whereby cross-lingual embeddings that cluster on language tend to perform poorly for downstream tasks (Fujinuma et al., 2019).

5 Conclusion

In this work we use VecMap to create cross-lingual word embeddings between English and isiZulu. We have shown that VecMap generates modular word embeddings in the cross-lingual space due to the monolingual word embeddings generated by FastText. Hence, we generate cross-lingual embeddings that are used to train a classifier that performs good on English news. However, we were unable to transfer the learning on unseen isiZulu news articles.

In future work, we would like to examine the performance of VecMap using a larger corpus for isiZulu. Additionally, it would also be advantageous to apply the modularity metric to optimize the hyperparameters of FastText in order to generate appropriate monolingual embeddings for the task.

Limitations

In this section of the paper we describe the limitations of the paper. We made design choices based on the corpus and the resources available for making the research possible.

The work presents a corpus that is of limited size for isiZulu. This is due to the lack of resources for the language. The other work that has attempted to build monolingual word embeddings for isiZulu is by Dlamini et al. (2021). However, the results were not published in a publicly accessible resource that would allow us to compare the embeddings generated by our work.

VecMap does not scale well without the use of a GPU, and hence hyperparameter searching was not done for this work. However, using vectors with 128 dimensions and a larger window size of 10 as suggested by Ri and Tsuruoka (2020) resulted in a performance decrease, even for the English news articles.

We also note that downstream model was trained using European news article titles and that can have an impact on the performance of identifying events that are uniquely for South African news.

Author Contributions

Derwin Ngomane: Data curation; formal analysis; investigation; methodology; writing – original draft; writing – review and editing. **Vukosi Marivate**: Conceptualization; data curation; resources; supervision; validation; writing – review and editing. **Jade Abbott**: Supervision; data curation; validation; writing – review and editing. **Rooweither Mabuya**: Supervision; data curation; writing – review and editing.

Acknowledgements

We would like to acknowledge funding from Facebook (Machine Translation Gift), the ABSA Chair of Data Science and the Google Research scholar program.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Enkhbold Bataa and Joshua Wu. 2019. An investigation of transfer learning-based sentiment analysis in japanese. *arXiv preprint arXiv:1905.09642*.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Jeanne E Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 948–953.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

Sibonelo Dlamini, Edgar Jembere, Anban Pillay, and Brett van Niekerk. 2021. isizulu word embeddings. In *2021 Conference on Information Communications Technology and Society (ICTAS)*, pages 121–126. IEEE.

Meluleki Dube and Hussein Suleman. 2019. Language identification for south african bantu languages using rank order statistics. In *International Conference on Asian Digital Libraries*, pages 283–289. Springer.

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14):6541.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. *arXiv*.

Luis Gutiérrez and Brian Keith. 2018. A systematic literature review on word embeddings. In *International Conference on Software Process Improvement*, pages 132–141. Springer.

C. Maria Keet and Langa Khumalo. 2016. Grammar rules for the isiZulu complex verb. *arXiv*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka english - isizulu parallel corpus. Https://doi.org/10.5281/zenodo.5035171.

Mack Makgatho, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner. 2021. Training cross-lingual embeddings for setswana and sepedi. *arXiv preprint arXiv:2111.06230*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho B. Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Dorrit Posel and Jochen Zeller. 2020. *Language use and language shift in post-apartheid South Africa*, pages 288–309. Cambridge University Press Cambridge.

Ryokan Ri and Yoshimasa Tsuruoka. 2020. Revisiting the context window for cross-lingual word embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 995–1005, Online. Association for Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019a. Unsupervised cross-lingual representation

learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019b. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Statistics South Africa (Stats SA). 2012. The South Africa I know, The Home I understand. https://www.statssa.gov.za/census/census_2011/census_products/NW_Municipal_Report.pdf.pdf.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

María Paula Villegas, María José Garciarena Ucelay, Juan Pablo Fernández, Miguel A Álvarez Carmona, Marcelo Luis Errecalde, and Leticia Cagnina. 2016. Vector-based word representations for sentiment analysis: a comparative study. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? *arXiv preprint arXiv:2005.09093*.