Very Low Resource Sentence Alignment: Luhya and Swahili

Everlyn Asiko Chimoto
University of Cape Town, South Africa
African Institute for Mathematical
Sciences
`everlyn@aims.ac.za`

Bruce A. Bassett
University of Cape Town, South Africa
African Institute for Mathematical
Sciences, South Africa
South African Astronomical Observatory
`bruce.a.bassett@gmail.com`

Abstract

Language-agnostic sentence embeddings generated by pre-trained models such as LASER and LaBSE are attractive options for mining large datasets to produce parallel corpora for low-resource machine translation. We test LASER and LaBSE in extracting bitext for two related low-resource African languages: Luhya and Swahili. For this work, we created a new parallel set of nearly 8000 Luhya-English sentences which allows a new zero-shot test of LASER and LaBSE. We find that LaBSE significantly outperforms LASER on both languages. Both LASER and LaBSE however perform poorly at zero-shot alignment on Luhya, achieving just 1.5% and 22.0% successful alignments respectively (P@1 score). We fine-tune the embeddings on a small set of parallel Luhya sentences and show significant gains, improving the LaBSE alignment accuracy to 53.3%. Further, restricting the dataset to sentence embedding pairs with cosine similarity above 0.7 yielded alignments with over 85% accuracy.

1 Introduction

Sentence alignment is the creation of parallel corpora from monolingual data (Gale and Church, 1993; Kay and Roscheisen, 1993). This alignment can be done manually and/or automatically. Manual alignment is laborious and costly hence there has been a lot of work on automatic sentence alignment (Steingrímsson et al., 2021; Schwenk, 2018; Guo et al., 2018). Tasks such as Building Using Comparable Corpora (BUCC) focus on building parallel corpora using neural methods (Zweigenbaum et al., 2017, 2018). Essentially, sentences are aligned to the corresponding translation in another language using language agnostic sentence embeddings with the idea that sentences that are translations of each other will be close in the vector space (Huang et al., 2015; Zhang et al., 2015). These sentence embeddings are generated using pre-trained models such as Language

Agnostic Sentence Representation (LASER) and Language Agnostic BERT Sentence Embeddings (LaBSE) (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019; Feng et al., 2022). LASER and LaBSE have been used to effectively mine bitext from comparable corpora.

As these pre-trained models are effective in mining bitext, we investigate how they would perform on an unseen low-resource language: Luhya, Marama dialect, as well as Swahili. Our main contributions are:

1. We created a Luhya-English parallel corpus of nearly 8000 aligned Luhya[1] and English sentences.

2. An empirical evaluation of LASER and LaBSE on Luhya and Swahili datasets.

3. Fine-tuning Luhya embeddings to improve bitext mining for this unseen language to explore the value of small amounts of parallel sentences for improving zero-shot performance.

2 Multilingual Sentence Embeddings

In this section, we review LASER and LaBSE.

2.1 LASER

Language Agnostic SEntence Representation (LASER) is a framework used to obtain multilingual sentence embeddings (Schwenk and Douze, 2017). It borrows from neural machine translation by utilizing encoders and decoders to generate the sentence embeddings which are of a fixed size in this case 1024 (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014).

The encoder-decoder architecture is shown in Figure 1. The encoder consists of 1-5 stacked BiL-STM layers each of dimension size 512 Artetxe and Schwenk (2019). The output of the encoder is
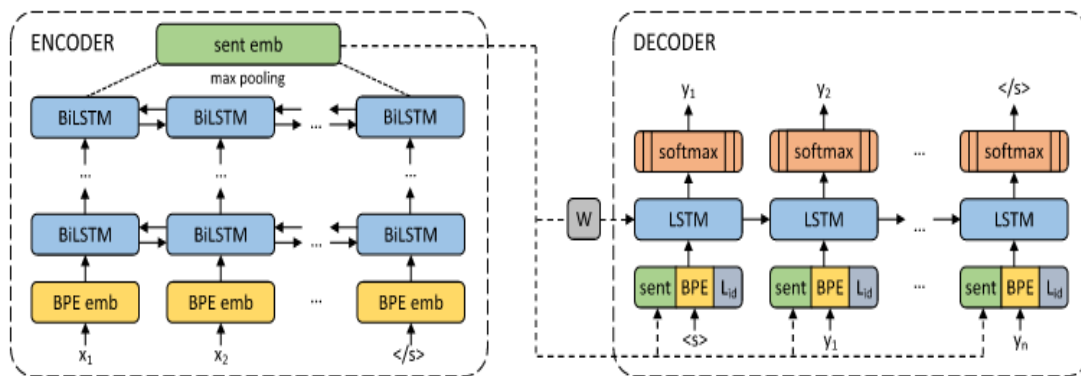
---

[1] Also sometimes written as Luhyia.

Figure 1: The LASER architecture as proposed by Artetxe and Schwenk (2019) consisting of a single encoder and a single decoder. The encoder has 1-5 stacked BiLSTM layers followed by a max pooling. The decoder is an auxiliary component.

max pooled to get the sentence embeddings. On the other hand, the decoder is an auxiliary component that consists of an LSTM layer of dimension 2048. LASER was trained by feeding 93 input languages to the system with a joint Byte Pair Encoding (BPE) with 50k merge operations. While the input to the encoder is just the BPE embedding, the input to the decoder consists of the sentence embedding generated by the encoder, the BPE embedding of the translation as well as the language ID. The encoder does not include the language ID since the goal is to allow the model to learn language-independent representations.

At the time of its release, LASER achieved state-of-the-art results in mining bitext in the BUCC task dataset (Zweigenbaum et al., 2017, 2018) for all language pairs except Chinese-English.

2.2 LaBSE

The Language Agnostic BERT Sentence Embeddings (LaBSE) framework is a cross-lingual approach that utilises a pre-trained BERT model to generate sentence embeddings (Feng et al., 2022; Yang et al., 2019a). The LaBSE model consists of 12-layer transformer dual encoders which share parameters (Guo et al., 2018). These encoders are initialized using pre-trained BERT weights (Devlin et al., 2019). Each encoder is fed source and target text respectively and embeddings are trained by minimizing the translation ranking loss with additive margin softmax (Yang et al., 2019b); see Figure 2 for further details. Each output embedding vector has dimension 768. The LaBSE model was trained on 109 languages and achieved state-of-the-art performance with bitext mining as shown in
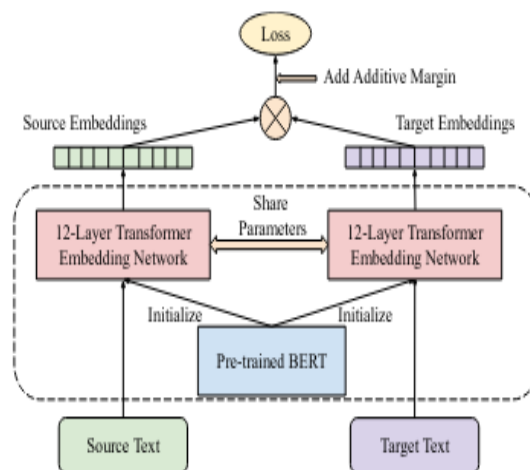
Feng et al. (2022); Heffernan et al. (2022).

Figure 2: The LaBSE architecture as proposed by Feng et al. (2022) which uses a dual encoder, one taking in the source sentence while the other taking in the target sentence. The encoders are initialized using pre-trained BERT weights and the architecture is trained using the translation ranking loss with additive margin.

3 Languages

In this section, we provide a brief description of the low-resource languages experimented on: Luhya (Marama dialect) and Swahili. The Marama dialect of the Luhya language, also known as Olumarama, is spoken in Western Kenya in Kakamega and Vihiga region with about 43,000 speakers (Eberhard et al., 2021). The language status of Marama is educational as it is used vigorous both verbally and in broadcast media (Eberhard et al., 2021). On the other hand, Swahili is spoken in East and Central

Africa including countries such as Kenya, Tanzania, DRC, parts of Uganda & Rwanda and has approximately 100 million speakers (Eberhard et al., 2021). Its status is national as several countries use it as their national language (Eberhard et al., 2021). These two languages are both Bantu languages from the Niger-Congo language family (World Atlas of Language Structures, 2021). Being from the same language family, they have the same word order structure, namely sentences follow a Subject-Verb-Object (SVO) ordering. They are also both agglutinative. Since both LASER and LaBSE were trained on Swahili, Luhya makes a very interesting zero-shot example to see how much information is transferred from the raw embeddings.

4 Related work

With the proliferation of neural embedding techniques, there have been various efforts to align sentences in various low-resource languages. Thompson and Koehn (2019) introduce VecAlign which aligns sentences using LASER sentence embeddings (Artetxe and Schwenk, 2019) similarity as well as dynamic programming approximation. They experiment on low-resource language pairs namely: Sinhala-English and Nepali-English. They show that the sentences aligned using this method achieve improvement in machine translation models downstream.

On the other hand, Tien et al. (2021) proposed KC4Align that utilises a multilingual translation system to generate embeddings. The similarity in these embeddings is used to perform paragraph alignment. Sentences are aligned where the sentences appear in the paragraph alignment using similarity scores and sentence length ratio. This method was tested on the Vietnamese-Laos language pair. Focusing on African low-resource language, Schwenk et al. (2021) extracts parallel sentences from Wikipedia using multilingual sentences embeddings for Swahili among other languages.

Regarding Luhya, there has been work on building Luhya datasets. Steimel (2018) work focuses on parts of speech tagging for the Wanga Luhya dialect whereas Wanzare et al. (2022) focuses on producing parallel datasets for several Luhya dialects to English with the help of human translators. This is in contrast to our work which analyses automatic sentence alignment for Luhya.

5 Methodology

5.1 LASER and LaBSE evaluation

LASER and LaBSE were utilised to generate our raw embeddings. Following the embedding generation, we compute the cosine similarity and the Euclidean distance (L2) from each vector in the English embedding set and all the Luhya/Swahili embeddings. Sentences are aligned to the most similar or closest sentence. We took both the Top-1 and Top-3 best alignments to test the performance of the pre-trained models.

We use accuracy as our key metric. An important note is that we evaluate alignment performance by demanding exact matching of sentence indices on both sides. This means that if a sentence appears more than once in one of the languages, and the alignment chooses the "wrong" index, despite the sentence being identical to the "correct" sentence, then this is classified as a fail. This will apply primarily to the Top-1 results. As a consequence, our accuracy estimates should be taken as a lower bound on the true alignment performance.

5.2 Fine-tuning the embeddings

To fine-tune our Luhya embeddings, we added a fully-connected network with a single hidden layer to help learn new weights where the cosine similarity between the new embedding and the English embedding would be greatest. We defined the loss as:

$$\text{Loss}(\mathbf{x}, \mathbf{y}) = 1 - S_C(\tilde{\mathbf{x}}, \mathbf{y})$$

where $\mathbf{x}, \mathbf{y}$ are the raw Luhya and English embeddings, S_C is the cosine similarity and $\tilde{\mathbf{x}}$ are the fine-tuned Luhya embeddings which depend on $\mathbf{w}_{1,2}$, the vectors of new weights introduced by the fine-tuning architecture (Pal and Savvides, 2017). $\mathbf{w}_1$ representing weights to the hidden layer and $\mathbf{w}_2$ representing weights to the output layer. The bottleneck layer size is a hyper-parameter that we vary to explore its impact on performance; see Figure 3.

6 Experiments

6.1 Datasets and Alignment

We experiment on both Luhya-English and Swahili-English sentence alignment. The Luhya-English parallel set was created by aligning sentences from the New Testament Bible translations. This dataset consists of 7952 parallel sentences in the Marama dialect of Luhya. This bitext creation was achieved
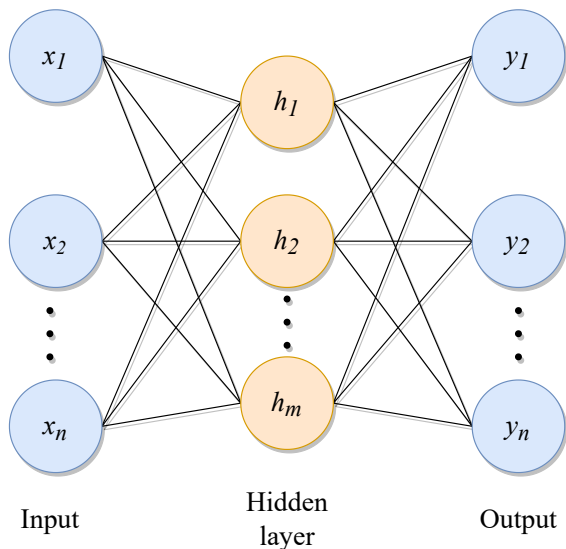
Figure 3: Fine-tuning architecture with one bottleneck layer that can vary in size. Hidden sizes of 32, 64, 96, 128 and 256 were tested. The network takes in the generated embeddings from LaBSE and outputs new vectors that maximize the similarity between the new Luhya embedding and English embedding.

by cleaning, curating and aligning the New Testament of the Bible in Luhya and English. After alignment, the dataset was assessed by three independent Native speakers to assess the quality of alignment. This dataset is the only known digital parallel corpus in Luhya-English for the Marama dialect. See examples of the aligned sentences on Table 1.

The Swahili-English dataset was sourced from the SAWA dataset which contains approximately $89k$ parallel sentences from various domains (De Pauw et al., 2009; Pauw et al., 2011). We sampled 10k parallel sentences from the Bible. This sampling allows comparison of automatic alignment specifically in the religious domain. Data cleaning involved getting rid of characters from different text encodings, removing both extra white spaces and verse numbers from all the datasets.

6.2 Results and Discussion

We utilize LASER and LaBSE to test out zero-shot bitext mining on Luhya-English dataset. Luhya is not included in the initial training of these models. We take both cosine similarity and Euclidian distance of the English embedding to the Luhya embedding. The results can be seen on table 2 where we can see we see LaBSE outperforming LASER on Luhya-English alignment by matching up to 22% of the sentences correctly whereas LASER

only matched 0.02%. We also note that increasing the number of sentences that are match from 1 to 3 does not increase the performance of LASER in bitext mining while LaBSE performance increases by 10%. The top-3 result means we consider accurate alignment if the correct alignment was among the top 3 matched sentences.

The performance of LaBSE shows that there are great gains achieved by utilising a pre-trained model in the sentence embedding model. LaBSE model utilises BERT in its training offering cross-lingual benefit that results in up to 22% accurate alignment on a language it has not seen before. LASER on the other hand was trained from scratch and does not provide great results in aligning Luhya, an unseen language.

Considering the performance on Swahili, we see that prior knowledge of a language greatly helps in performance. Swahili performs better than Luhya in the alignment with LaBSE embeddings resulting in near perfect alignment (See table 2). The performance of LASER with Swahili does not correspond with the results by Artetxe and Schwenk (2019) where F1 scores of above 90% were recorded. In our case, the F1 score is equivalent to the accuracy as the number of extracted parallel sentences is equivalent to the number of gold standard alignments. Contrary to the LASER results, the results of LaBSE outperforming LASER corresponds with the results from Feng et al. (2022); Heffernan et al. (2022). We also observed that cosine similarity performs marginally better than Euclidean distance on average and hence is used for our fine-tuning experiments.

6.3 Fine-tuning LaBSE Luhya embeddings

Owing to the good performance of 22% on zero-shot alignment, we fine-tune the LaBSE Luhya embedding to evaluate the extent one needs to go to see improvements. Initially, we added one additional layer without the bottleneck and trained this network with 50% of the Luhya dataset while testing on the other 50%. This achieved an accuracy of 40.22%. However, we did not pursue this network further as the number of trained parameters was too large to offer value for a small data as the Luhya dataset. We added a bottleneck layer whose input was the 768-sized embedding and the output layer was of size 768. Our experimental setup aimed to investigate what amount of correctly aligned sentences are needed to fine-tune the embeddings and

4

Luhya	English
Nebutswa omukholi , womumukunda oyo na-mukalusia ari , ' Omwami , lekha , kubekhwoho omuyika kuno khandi , nasi ndalakwachila , nekurakhwo imbolela.	But he answered and said to him , ' Sir , let it alone this year also , until I dig around it and fertilize it.
Ne , nali emakombe nanyasibungwa muno , ya-henga ikulu ne , nalola Aburahamu nende Lazaro nibali halala ehale.	And being in torments in Hades , he lifted up his eyes and saw Abraham afar off , and Lazarus in his bosom.
Saulo namenya ninabo nayaala muliira lia Yesu , mu Yerusalemu obularia likhuwa liosi liosi tawe.	So he was with them at Jerusalem , coming in and going out.

Table 1: Sample aligned sentences from our bible dataset.

		LASER		LaBSE	
		Top-1	*Top-3*	*Top-1*	*Top-3*
Luhya-English	Cos. Sim.	0.02	0.02	0.22	0.32
	L2	0.00	0.01	0.22	0.32
Swahili-English	Cos. Sim.	0.50	0.55	0.97	1.00
	L2	0.45	0.55	0.97	1.00

Table 2: Alignment accuracy for Luhya and Swahili using the raw LASER and LaBSE embeddings (no fine-tuning). The Top-1 (Top-3) columns represent the accuracy of correctly aligned sentences based on the sentences with the top 1 (3) most similar embeddings based either on cosine similarity ("Cos. Sim") or Euclidean distance ("L2"). LaBSE performs better than LASER on both languages, correctly aligning 22.0% of Luhya-English sentences and 97.1% of Swahili-English sentences. LASER performs poorly when aligning Luhya-English with only 1.5% being aligned correctly.

see improvement as well as what is the optimal hidden size to achieve improvement in alignment. We split the Luhya-English dataset into 5-folds. At each iteration of training, one fold of 1591 parallel sentences was used for testing while the other folds were used for training. To investigate how many sentences are required to see improvements, we used 10% of the training set to fine-tune the embeddings and evaluated the model with the test set and continuously added 10% until the whole training set was used up. The training set consisted of 6361 parallel sentences.

Figure 5 shows the results in which we see that regardless on the hidden size only 20% of the training data set, about 1272 sentence, is sufficient to result in the improvement of the alignment. Also, looking at the different hidden sizes 128 and 96 hidden sizes were the best with no distinct difference between the two. As much as the hidden size of 128 is comparable with 96, hidden size of 96 works with fewer parameters thus more efficient. Increasing the hidden size beyond 128 degraded the performance as evidence by the performance of hidden size 256. Also, smaller hidden sizes did not offer much gain.

Training Size	Threshold	Precision	Recall	F1 Score
At 10%	-0.2	0.22	0.22	0.22
	0.54	0.29	0.18	0.22
	0.68	0.54	0.03	0.06
At 20%	-0.2	0.34	0.34	0.34
	0.5	0.38	0.31	0.34
	0.69	0.72	0.05	0.09
At 40%	-0.2	0.44	0.44	0.44
	0.52	0.49	0.41	0.45
	0.72	0.83	0.06	0.11
At 60%	-0.2	0.48	0.48	0.48
	0.53	0.54	0.43	0.48
	0.73	0.88	0.06	0.11
At 80%	-0.2	0.51	0.51	0.51
	0.5	0.55	0.49	0.52
	0.75	0.9	0.05	0.09
At 100%	-0.2	0.53	0.53	0.53
	0.5	0.56	0.51	0.53
	0.75	0.92	0.05	0.09

Table 3: Precision, Recall and F1 score as the cosine similarity score threshold is increased with different training sizes. -0.2 threshold represents not setting a threshold at all, this results in all sentences being aligned however the precision is very low as some sentences are wrongly aligned. As the threshold increases fewer sentences are considered aligned, however there are more accurate alignment. Hence, the precision increases while the recall and F1 score decrease.
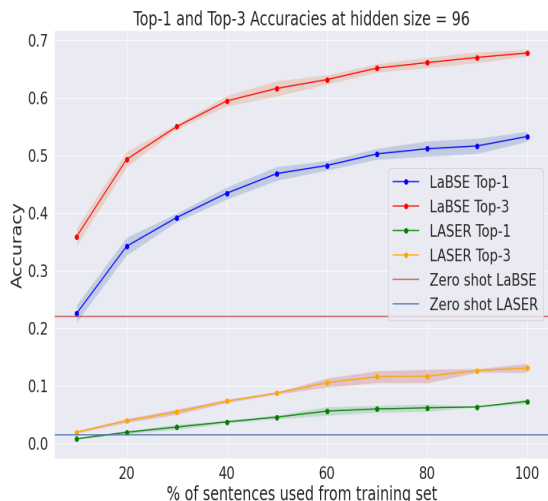
Figure 4: LaBSE outperforms LASER in both the Top-1 and Top-3 results from fine-tuning the respective embeddings (with a hidden layer dimension of 96) for Luhya. Results are shown as a function of the percentage of the full training set of 6361 sentences used for training. For LaBSE the Top-1 (Top-3) accuracy reaches 53% (68%). Error bars are standard deviations estimated from 5-fold cross-validation.
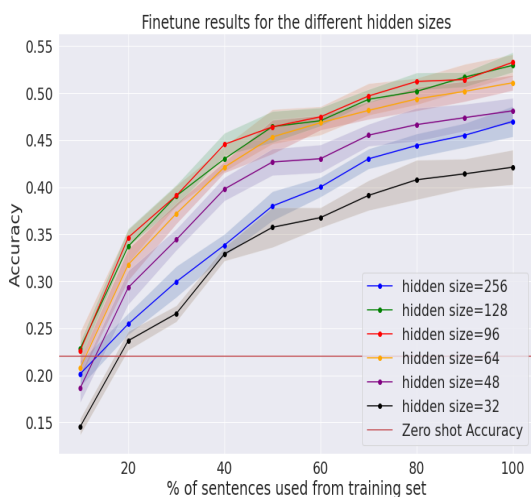

Figure 5: Top-1 results from fine-tuning the Luhya embeddings from LaBSE using different hidden layer sizes, as a function of the percentage of the full training set of 6361 sentences, together with the zero-shot result of 22.0%. Results are the mean of 5-fold cross validation with error bands given by $\pm 1\sigma$, where σ is the standard deviation of the 5 folds. The best hidden sizes are for hidden layers of dimensionality 96 and 128.

As we note good results with a hidden size of 96, we perform both the top-1 and the top-3 evaluation for both LASER and LaBSE Luhya embedding fine-tuning. Figure 4 shows that fine-tuning with hidden size 96 results in an accuracy of up to 68%. These results show that with little effort, LaBSE

embeddings can be used to effectively mine bitext of languages not seen during training. This is practical for various very low-resource languages.

To assess the accuracy of aligned sentences after fine-tuning, we analyse the accuracy against different similarity score thresholds along with different training set size. Figure 6 shows that fine-tuning with 40% of the training dataset gets an accuracy of above 80% meaning with 2500 sentences one can get >80% accuracy in alignment past 0.65 threshold. Table 3 shows that setting a high similarity threshold results in higher precision but we see a significant drop in recall. Using the full training set, we see that without setting a threshold, the precision and recall are 53%. However, setting the threshold to 0.75 improves the precision to 92% while the recall drops to just 5%. Setting a higher threshold results in more accurate alignments but selects fewer sentence pairs. Choosing the optimal threshold to balance precision and recall will depend on the task being considered.
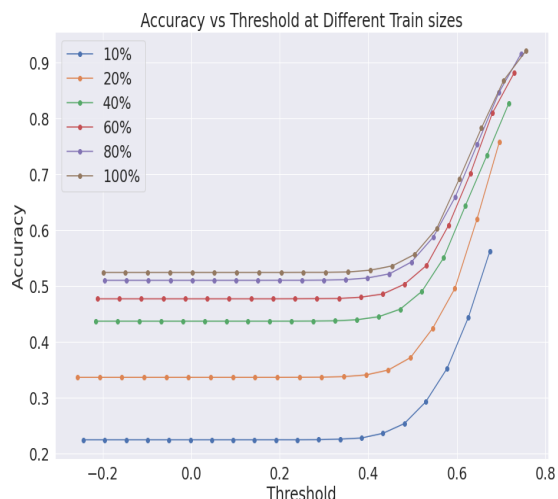

Figure 6: Top-1 accuracies as a function of cosine similarity threshold and training dataset percentage. We see that accuracy correlates with similarity, allowing the curation of high-accuracy datasets suitable for use in machine translation.

Conclusion

Low-resource languages lack digitized parallel data needed for developing machine translation models. Sentence embedding models offer a potential way to create parallel data cheaply for these low and very low-resource languages but themselves need parallel data for their training.

In this work we present a new dataset consisting of 7952 sentences translating the Luhya (Marama)

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Top-1	0.22	0.34	0.39	0.43	0.47	0.48	0.50	0.51	0.52	**0.53**
Top-3	0.36	0.49	0.55	0.59	0.62	0.63	0.65	0.66	0.67	**0.68**

Table 4: Fine-tuning results with a hidden layer size of 96 for the LaBSE embeddings. With 100% of the training data, (6362 sentences), the Top-1 accuracy is 53.2%, while the Top-3 accuracy is 67.8%.

dialect into English. We use this data, together with a similar Swahili dataset, to explore transfer learning and fine-tuning based on raw embeddings from the LASER and LaBSE algorithms for alignment on these languages.

We show that LaBSE significantly outperforms LASER on both Swahili and Luhya but that both struggle with zero-shot learning on Luhya, achieving alignment accuracies of 22.0% and 1.5% respectively.

We also show that fine-tuning with as little as 1200 correctly aligned Luhya sentences can result in models with significantly improved sentence alignments. In addition, setting a minimum similarity score threshold results in datasets with much more accurate alignments, useful for curating high-quality parallel corpora for machine translations. However, this comes at the cost of significantly reducing the number of aligned sentences. We leave it to future work to investigate active learning for the choice of sentences for fine-tuning that will result in the greatest gains in performance.

Acknowledgements

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The SAWA corpus: A parallel corpus English - Swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16, Athens, Greece. Association for Computational Linguistics.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig(eds.). 2021. Ethnologue: Languages of the world. twenty-fourth edition.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *ArXiv*, abs/2205.12654.

Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha P. Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1084–1088, Lisbon, Portugal. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Dipan K. Pal and Marios Savvides. 2017. Copernican loss : Learning a discriminative cosine embedding.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2011. Exploring the sawa corpus: collection and deployment of a parallel corpus english—swahili. *Language Resources and Evaluation*, 45:331–344.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Kenneth Steimel. 2018. Part of speech tagging in Luyia: A Bantu macrolanguage. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 46–54, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective bitext extraction from comparable corpora using a combination of three different approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode). INCOMA Ltd.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Ha Nguyen Tien, Dat Nguyen Huu, Huong Le Thanh, Vinh Nguyen Van, and Minh Nguyen Quang. 2021. KC4Align: Improving sentence alignment method for low-resource language pairs. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 354–363, Shanghai, China. Association for Computational Lingustics.

Lilian D.A Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, Barack Wanjawa, and Lawrence Muchemi. 2022. KenTrans: A Parallel Corpora for Swahili and local Kenyan Languages.

World Atlas of Language Structures, 2021. Accessed July 2021. Languages. Wals, `https://wals.info/languoid`.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *CoRR*, abs/1902.08564.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Matthew Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019b. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *ArXiv*, abs/1902.08564.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2015. Towards machine translation in semantic vector space. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 14(2).

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan.