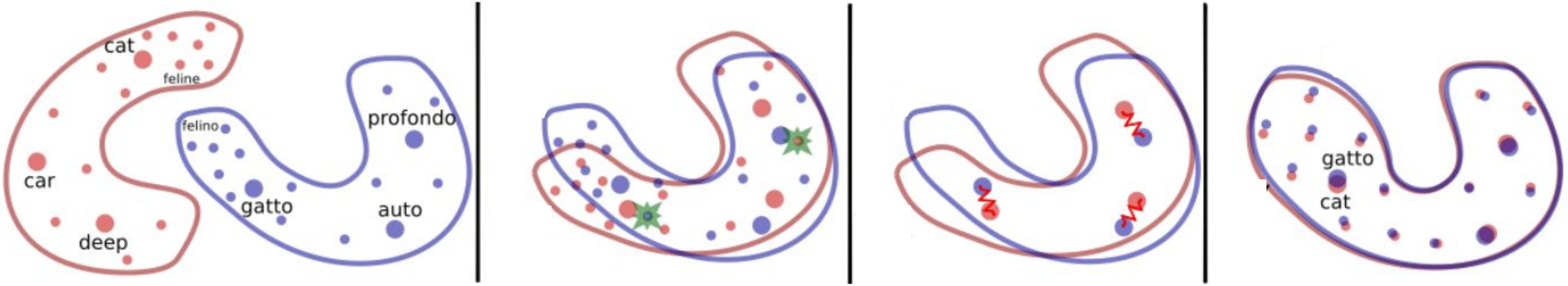
Multilingual Word Embedding Alignment for Sinhala

229405N-M.K.I.Wickramasinghe
(Supervisor: Dr. Nisansa de Silva)



Content

- Introduction
- Research Problem
- Literature Survey
- Research Phases
- Summary

Introduction

Introduction

- Embeddings are the basic ingredient in many kinds of natural language processing tasks.
- In multilingual tasks unaligned embedding spaces are a huge burden. [1]
- The alignment is required for two kinds of embedding models.
  - Embedding models separately trained on monolingual data
  - Multilingual models trained on parallel multilingual data
- Multilingual model training process implicitly encourages for the alignment [2, 3, 4]
- For monolingual models, the alignment has to be done as a separate task [5, 6, 7]
- Monolingual embedding alignment is still vital since,
  - Monolingual models are lightweight
  - Can be run using simpler libraries and frameworks
  - Using multilingual models may be redundant due to supporting many languages [2, 3, 4]
  - Accuracy can be compromised due to the support of many languages in multilingual models [2]
  - The accuracy for low-resource languages can be less compared to high-resource languages due to training data imbalance in multilingual models [2]
  - Pretraining or fine-tuning a multilingual model is time and resource consuming [2,3,4]

[1] A. Kalinowski and Y. An, 'A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs', arXiv preprint arXiv:2010. 13688, 2020.
[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," ACL,2022.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," NIPS, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," ACL, 2020.
[5] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," arXiv preprint arXiv:1309.4168, 2013.
[6] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized word embedding and orthogonal transform for bilingual word translation," NAACL, 2015.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," EMNLP, 2018.

Research Problem

Research Problem

- Monolingual word embedding models have been there for decades. [8, 9]
- Aligned word embedding models are not available for many languages[1]. [7]

The main focus of the research is to find the best aligned word embedding model (find the transformation matrix) between Sinhala and English languages.

  - To facilitate the above, as an intermediate goal, we shall build a Sinhala-English parallel word dataset/ dictionary
  - This will serve as the anchor dataset for Sinhala-English supervised word embedding alignment

- We study the classical monolingual embedding alignment techniques, novel multilingual embeddings and hybrid embedding alignment techniques
- Further we study measurement of the degree of alignment

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," EMNLP, 2014.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," EMNLP, 2018.
1 https://fasttext.cc/docs/en/aligned-vectors.html

Literature Survey

Word Embedding Techniques

- Different vector representations for words have been there from early days and they were statistical and human crafted representations.
  - One-hot-encoding
  - Count vectorizing
  - TF-IDF [10]

- The idea of generating word embeddings without direct human interaction (complex embedding representations) was introduced in 2013 by Mikolov et al. [8] by introducing Word2Vec.

- After that two similar models were introduced,
  - GloVe [9]
  - FastText [11]

- The beauty of these new word embeddings is that the embeddings:
  - Gives a global representation of words (gives a fixed embedding for a given word) [8, 9, 11]
  - Perform word analogy arithmetic (***Paris - France + Rome = Italy*** [8])

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” EMNLP, 2014.
[10] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.
[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.

Contextual Embeddings

- The meaning of a word changes according to its context (where that word occurs in a sentence and what the other words in the sentence). This is the classical *sense disambiguation problem* [12].
- Therefore, having a global vector representation for a word is not a good approach in cases where a context related representations are needed
- Word2Vec [8], Glove [9] and FastText [11] embeddings are global embedding representations where earlier TF-IDF [10], one-hot etc. do have some context representations but not powerful enough
- ELMo[13] which is a deep bi-LSTM based word embedding generator is the first context based embedding model that became popular
- After the introduction of Transformers [14], a revolutionary improvement happened in the contextual embedding representations, where researches could achieve state of the art accuracies and efficiencies in embedding generation.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” EMNLP, 2014.
[10] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.
[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.
[12] H. T. Ng and H. B. Lee, ‘Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach’, arXiv preprint cmp-lg/9606032, 1996.
[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” 2018. [Online].Available: https://arxiv.org/abs/1802.05365
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” NIPS, 2017.

Transformer-based Embeddings

- BERT[15] is the first transformer-based embedding generator which showed state of the art results at the first place.
- Then so many BERT variations were released after that; such as RoBERTa [16], ALBERT [17], ELECTRA [18] etc. where each of them showed improved results in accuracy or efficiency.
- Other than word embedding models, sentence-embedding models were also introduced as BERT extensions of which Sentence-BERT (S-BERT) [19] was the pioneer.
- After S-BERT many S-BERT variations were released and by now there are multitudes of[1] word and sentence embedding models out there that achieve better results than the initial BERT and S-BERT. [20, 21, 22]

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.
[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” arXiv preprint arXiv:1909.11942, 2019.
[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pretraining text encoders as discriminators rather than generators,” arXiv preprint arXiv:2003.10555, 2020.
[19] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” EMNLP, 2019.
[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’, arXiv preprint arXiv:1910. 01108, 2019.
[21] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, ‘MPNet: Masked and Permuted Pre-training for Language Understanding’, NIPS, 2020.
[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, ‘Xlnet: Generalized autoregressive pretraining for language understanding’, NIPS, 2019.

[1] https://www.sbert.net/docs/pretrained_models.html

Multilingual Embeddings

- The next advancement in word and sentence embeddings is having a single model for multiple languages.
- For word embeddings there are mBERT [23], XLM [3], XLM-R [4] etc. and for sentence embeddings there are LASER [24, 25], LaBSE [2] etc.
- The beauty of multilingual models is that they have a single embedding space for all the languages it supports. Thus, we can perform mathematical operations on the embeddings across languages, providing much reprieve for multilingual tasks.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," ACL, 2022.
[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," NIPS, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," ACL, 2020.
[23] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation,", EMNLP,2020
[24] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,", ACL,2019
[25] . Heffernan, O. Çelebi, and H. Schwenk, "Bitext mining using distilled sentence representations for low-resource languages," EMNLP,2022

Embedding Alignment (For Monolingual Embeddings)

- Aligned Embeddings are vital for multilingual tasks where embeddings of multiple languages share a single embedding space so that multilingual tasks can be performed irrespective of the language.
- Mikolov et al. [5] aligned two Word2Vec word embedding spaces assuming a simple linear mapping between the two embedding spaces
- Xing et al. [6], showed that better alignment results can be achieved by assuming an orthogonal mapping between two embedding spaces.
- VecMap [28] proposes a series of linear transformations to align two embedding spaces
- RCSLS by Joulin et al. [7] have addressed the so called *hubness issue* where some words appear too frequently in the neighborhoods of other words, by introducing an improved loss function for alignment called Cross-domain similarity local scaling (CSLS).
- Li et al. [29] have used the InfoNCEloss [30] (a contrastive loss) to iteratively improve weakly aligned word embeddings.
- All the above techniques are supervised alignment techniques which need to have a parallel word dictionary to decide the alignment matrix.
- Unsupervised techniques, while not as prevalent, do exist. Some are based on traditional statistical methods [23] while others are based on adversarial approaches [24].

[5] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” arXiv preprint arXiv:1309.4168, 2013.
[6] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” NAACL, 2015.
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” EMNLP, 2018.
[26] E. Grave, A. Joulin, and Q. Berthet, “Unsupervised alignment of embeddings with wasserstein procrustes,” in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 1880–1890.
[27] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” arXiv preprint arXiv:1710.04087, 2017.
[28] M. Artetxe, G. Labaka, and E. Agirre, “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018
[29] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vulí c, “Improving word translation via two-stage contrastive learning,” ACL, 2022
[30] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

Word Embedding Alignment for Sinhala

- Smith et al. [40] have published EN-Si alignment matrix along with 77 other languages using procrustes alignment technique in Si→En direction
- Liyanage et al. [46] have done a study on English and Sinhala embedding alignment using VecMap for bilingual lexicon induction task and affecting factors for the alignment

[40] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in International Conference on Learning Representations, 2016
[46]. Liyanage, S. Ranathunga, and S. Jayasena, “Bilingual lexical induction for sinhala-english using cross lingual embedding spaces,” in 2021 Moratuwa Engineering Research Conference (MERCon), 2021, pp. 579–584.

Alignment Datasets for Sinhala Language

- For supervised embedding alignment we need an alignment dataset which help to identify corresponding points in the two embedding spaces. These datasets are parallel datasets [7].
- For supervised word embedding alignment what we need is a parallel word dataset or a dictionary dataset [7].
- Sinhala, being a low-resource language does not have much such resources available at the moment [31].
- The dictionary Subasa Ingiya [31] is one of them which is a small dictionary that contains about 36000 pairs and contains not only word pairs but also phrases.
- We came across several multilingual parallel corpora that contain Sinhala as a language, such as the works by Guzmán et al. [32, 33], Hameed et al. [34], Bañón et al. [35] and Vasantharajan and Thayasivam [36] that are comprised of sentence and paragraph level parallel entries.
- They are well suited for higher-level multilingual tasks such as Machine Translation (MT) but, not for lower-level tasks such as word embedding alignment. [32, 34].

[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," EMNLP, 2018.
[31] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research," arXiv preprint arXiv:1906.02358, 2019.
[32] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," EMNLP, 2019.
[33] M. R. Costa-jussà et al., 'No language left behind: Scaling human-centered machine translation', arXiv preprint arXiv:2207. 04672, 2022.
[34] R. A. Hameed, N. Pathirennehelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, "Automatic creation of a sentence aligned sinhala-tamil parallel corpus," in Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), 2016, pp. 124–132.
[35] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn et al., "Paracrawl: Web-scale acquisition of parallel corpora," ACL, 2020.
[36] C. Vasantharajan and U. Thayasivam, "Tamizhi-net ocr: Creating a quality large scale tamil-sinhala-english parallel corpus using deep learning based printed character recognition (pcr)," arXiv preprint arXiv:2109.05952, 2021.
[37] A. Wasala and R. Weerasinghe, "Ensitip: a tool to unlock the english web," in 11th international conference on humans and computers, Nagaoka University of Technology, Japan, 2008

Aligned Embeddings in Multilingual Models

- Unlike in monolingual models, in multilingual models the embeddings get aligned in the training process itself.
- This is achieved by using alignment supportive training objective in the training process.
- We find two main methods of training multilingual models.
- One type of models have used multiple monolingual models to extend knowledge for building a multilingual model through knowledge distillation (mBERT [23], LASER [24, 25]) while,
- Another type of models have used large corpora of monolingual and multilingual parallel datasets to pre-train large multilingual models using training objectives such as Multilingual Masked Language Modelling (mMLM) and Translation Language Modeling (TLM) (LaBSE [2], XLM [3], XLM-R [4])

[23] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,”, EMNLP,2020
[24] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,”, ACL,2019
[25] . Heffernan, O. Çelebi, and H. Schwenk, “Bitext mining using distilled sentence representations for low-resource languages,” EMNLP,2022
[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic bert sentence embedding,” ACL,2022.
[3] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” NIPS, 2019.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” ACL, 2020.

Cross-Lingual Embedding Evaluation Techniques

- There are many tasks used by the research community to evaluate the quality of cross-lingual embeddings.
- One of the most popular methods is bilingual lexicon induction (BLI) or word translation task [38].
- There are other tasks as well such as cross-lingual natural language inference (XNLI) [39], cross-lingual semantic word similarity [27], sentence translation retrieval [27] and cross-lingual question-answering [4]

[27] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," arXiv preprint arXiv:1710.04087, 2017.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," ACL, 2020.
[38] A. Irvine and C. Callison-Burch, "A comprehensive analysis of bilingual lexicon Induction," ACL, 2017
[39] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," EMNLP, 2018

BLI Benchmarking Datasets

- BLI checks from a parallel data set, how many target translations of source words can be found using the aligned embedding spaces [27, 40]
- MUSE [27] is one the largest collection of bilingual dictionary collections with 110 language pairs.
- One issue with MUSE dataset is that 90 language pairs consist of English as one language and only 20 non-English language pairs are there.
- XLing [41] is another BLI dataset consisting of 8 languages and 56 BLI directions in total.
- PanLex-BLI [42] is another large-scale BLI dataset that consists of 210 BLI directions of 15 low-resource languages.
- None of these datasets contain Sinhala as a language

[27] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” arXiv preprint arXiv:1710.04087, 2017.
[40] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in International Conference on Learning Representations, 2016
[41] G. Glavaš, R. Litschko, S. Ruder, and I. Vulí c, “How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions,” ACL, 2019
[42] I. Vulí c, G. Glavaš, R. Reichart, and A. Korhonen, “Do we really need fully unsupervised cross-lingual embeddings?” EMNLP, 2019

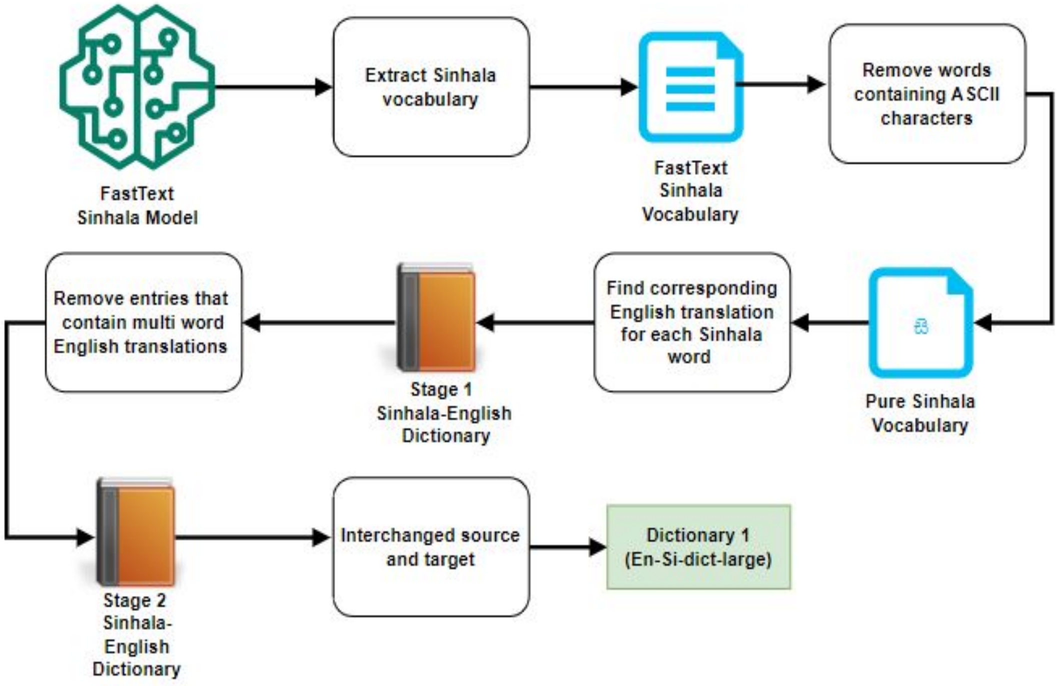
Research Phases

Phase 1:
Dataset Creation

Phase 1

Sinhala-English Parallel Dictionary Dataset

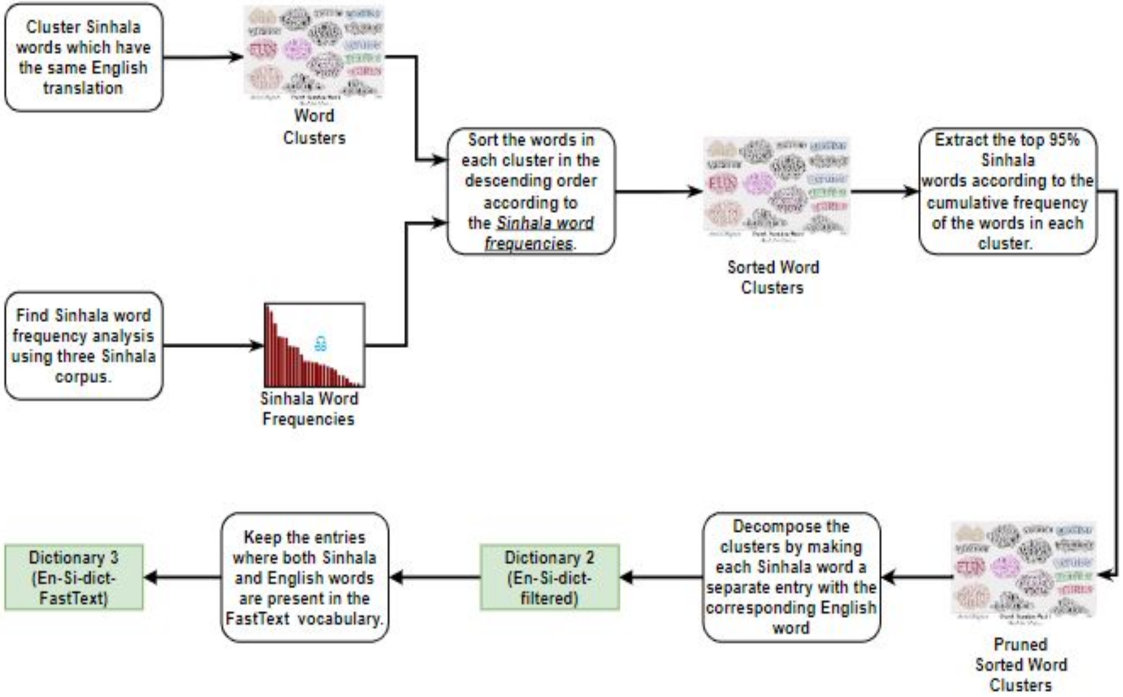
- We created 3 large-scale Sinhala-English parallel word datasets that facilitate word-level NLP tasks such as lexicon induction and supervised word embedding alignment
- We have created a data generation pipeline for the dataset creation process
- We used these datasets to create the alignment dataset for Sinhala-English word embedding alignment tasks we carried out
- This is the first pipeline of building our dictionary


FastText Sinhala Model
Extract Sinhala vocabulary
FastText Sinhala Vocabulary
Remove words containing ASCII characters
Pure Sinhala Vocabulary
Find corresponding English translation for each Sinhala word
Stage 1 Sinhala-English Dictionary
Remove entries that contain multi word English translations
Stage 2 Sinhala-English Dictionary
Interchanged source and target
Dictionary 1 (En-Si-dict-large)

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Sinhala-English Parallel Dictionary Dataset

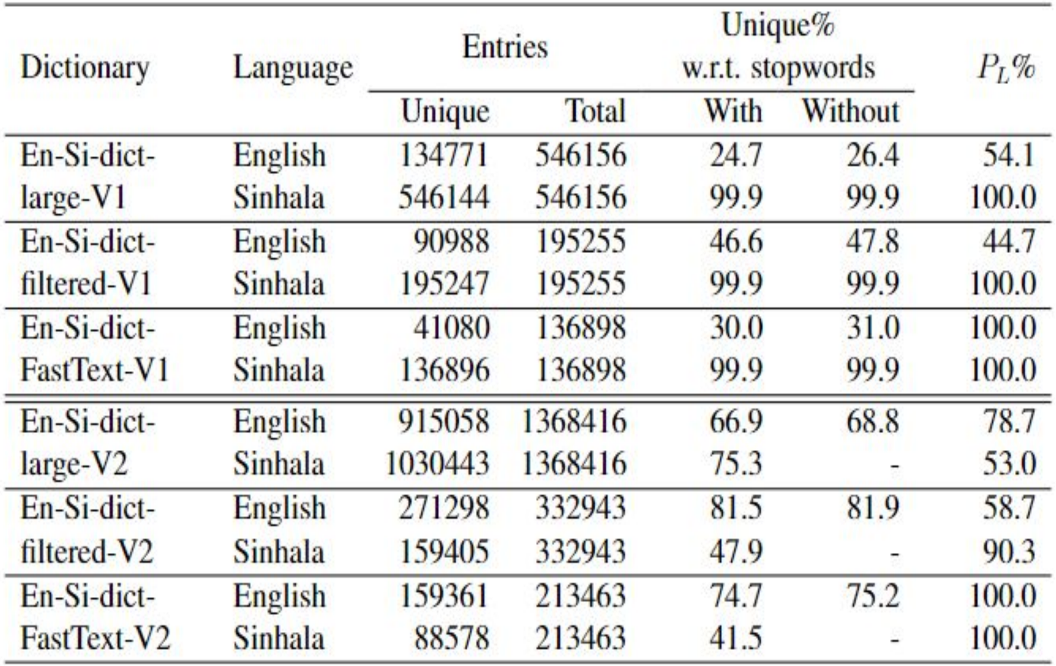
- We created 3 large-scale Sinhala-English parallel word datasets that facilitate word-level NLP tasks such as lexicon induction and supervised word embedding alignment
- We have created a data generation pipeline for the dataset creation process
- We used these datasets to create the alignment dataset for Sinhala-English word embedding alignment tasks we carried out
- This is the second pipeline of further filtering the originally built dataset into the next two versions


Cluster Sinhala words which have the same English translation
Word Clusters
Sort the words in each cluster in the descending order according to the Sinhala word frequencies.
Sorted Word Clusters
Extract the top 95% Sinhala words according to the cumulative frequency of the words in each cluster.
Find Sinhala word frequency analysis using three Sinhala corpus.
Sinhala Word Frequencies
Pruned Sorted Word Clusters
Decompose the clusters by making each Sinhala word a separate entry with the corresponding English word
Dictionary 2 (En-Si-dict-filtered)
Keep the entries where both Sinhala and English words are present in the FastText vocabulary.
Dictionary 3 (En-Si-dict-FastText)

[43] K. Wickramasinghe and N. de Silva, "Sinhala-English parallel word dictionary dataset," in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Sinhala-English Parallel Dictionary Dataset - Statistics

- Following table shows the statistics of our datasets
- Our largest dataset consists of ~1.3M translation pairs



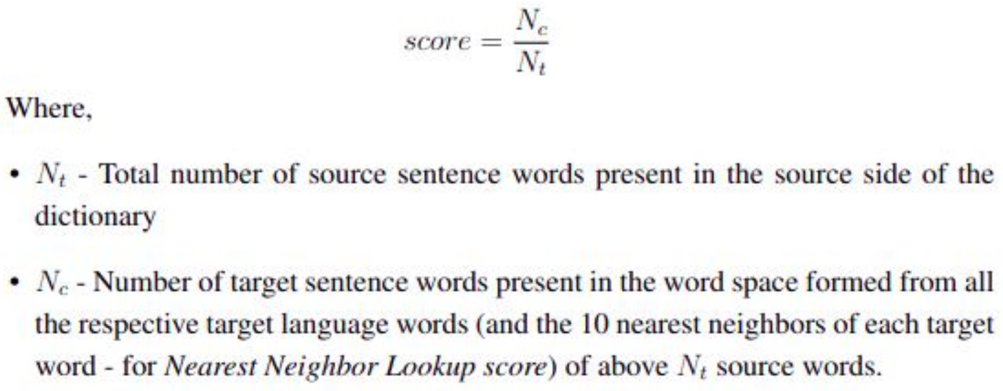
Dictionary	Language	Entries		Unique% w.r.t. stopwords		$P_L\%$
		Unique	Total	With	Without	
En-Si-dict-large-V1	English	134771	546156	24.7	26.4	54.1
	Sinhala	546144	546156	99.9	99.9	100.0
En-Si-dict-filtered-V1	English	90988	195255	46.6	47.8	44.7
	Sinhala	195247	195255	99.9	99.9	100.0
En-Si-dict-FastText-V1	English	41080	136898	30.0	31.0	100.0
	Sinhala	136896	136898	99.9	99.9	100.0
En-Si-dict-large-V2	English	915058	1368416	66.9	68.8	78.7
	Sinhala	1030443	1368416	75.3	-	53.0
En-Si-dict-filtered-V2	English	271298	332943	81.5	81.9	58.7
	Sinhala	159405	332943	47.9	-	90.3
En-Si-dict-FastText-V2	English	159361	213463	74.7	75.2	100.0
	Sinhala	88578	213463	41.5	-	100.0

- V1 - datasets created by translating the Sinhala FastText vocabulary to English
- V2 - datasets created by using both Sinhala and English FastText vocabularies

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Dataset Validation

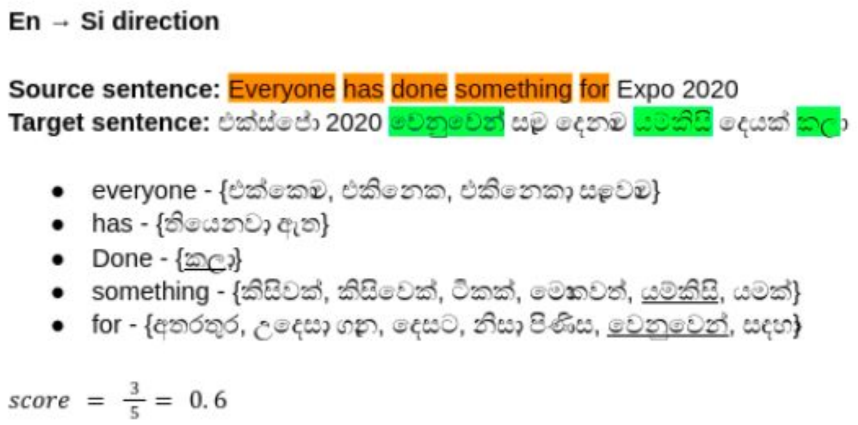
- To validate our datasets, we have defined two scoring criteria inspired by the widely used ROUGE-1 metric [82].
- The per-sentence score is calculated as follows.



$$score = \frac{N_c}{N_t}$$

Where,

- N_t - Total number of source sentence words present in the source side of the dictionary
- N_c - Number of target sentence words present in the word space formed from all the respective target language words (and the 10 nearest neighbors of each target word - for *Nearest Neighbor Lookup score*) of above N_t source words.



En → Si direction

Source sentence: Everyone has done something for Expo 2020
Target sentence: එක්ස්පෝ 2020 වෙනුවෙන් සෑම දෙනාම යමකිසි දෙයක් කළා

- everyone - {එක්කෙම, එකිනෙක, එකිනෙකා සළවම}
- has - {තියෙනවා ඇත}
- Done - {කළා}
- something - {කිසිවක්, කිසිවෙක්, ටිකක්, මොනවත්, යමකිසි, යමක්}
- for - {අතරතුර, උදෙසා ගැන, දෙසට, නිසා පිණිස, වෙනුවෙන්, සඳහා}

$$score = \frac{3}{5} = 0.6$$

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Phase 1

Dataset Validation

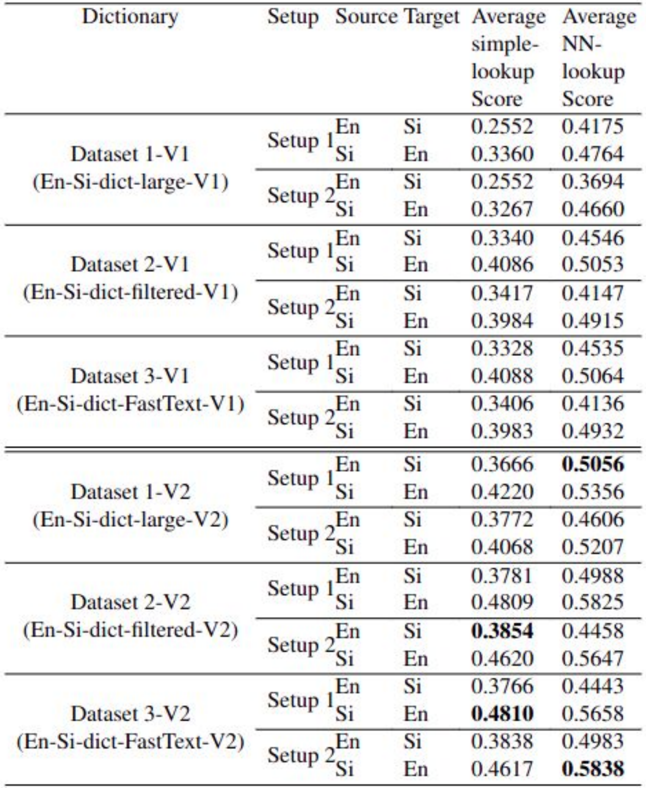
- We have validated our created datasets using the score we have defined
- We used two parallel and aligned datasets for that
  - English-Sinhala WikiMedia Dataset
  - English-Sinhala TED-2020 Dataset
- We validated these datasets in two setups
  - Setup 1 - All sentences as they are (Including stop-words)
  - Setup 2 - Excluding stop-words (used Lakmal et al. [45] for Si and Spacy for En)
- The scoring scheme we have defined is a ROUGE-1 like scheme
- Rule of thumb ROUGE-1
  - > 0.5 - Good
  - 0.4-0.5 - Moderate

[45] D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, "Word embedding evaluation for sinhala," in LREC, 2020, pp. 1874–1881.

Phase 1

Dataset Validation - WikiMedia Dataset

- Here we present the validation results using the English-Sinhala WikiMedia dataset.
- The dataset contains ~7.9k parallel sentences from Wikipedia translations
- In ROUGE-1 perspective our results are in the good and moderate regions in most of the cases

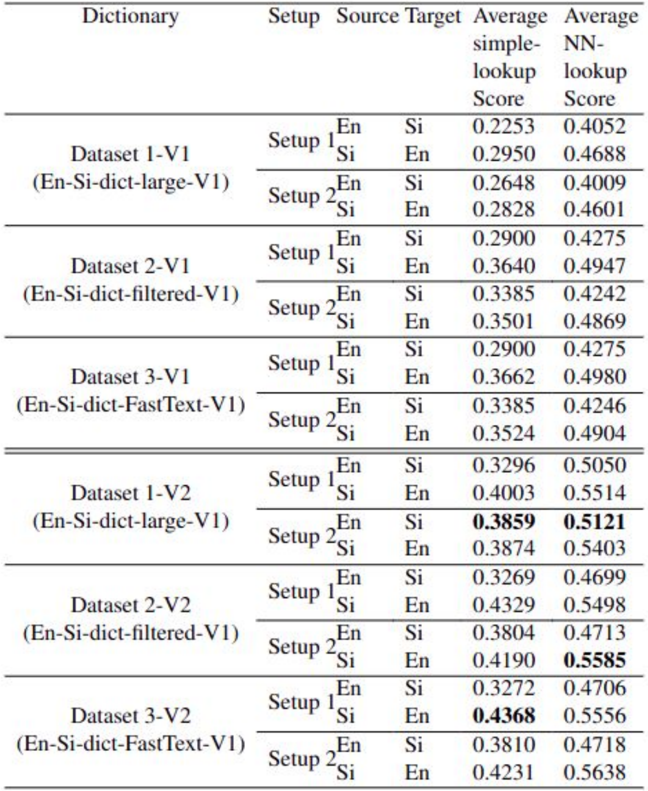


Dictionary	Setup	Source	Target	Average simple-lookup Score	Average NN-lookup Score
Dataset 1-V1 (En-Si-dict-large-V1)	Setup 1	En	Si	0.2552	0.4175
		Si	En	0.3360	0.4764
	Setup 2	En	Si	0.2552	0.3694
		Si	En	0.3267	0.4660
Dataset 2-V1 (En-Si-dict-filtered-V1)	Setup 1	En	Si	0.3340	0.4546
		Si	En	0.4086	0.5053
	Setup 2	En	Si	0.3417	0.4147
		Si	En	0.3984	0.4915
Dataset 3-V1 (En-Si-dict-FastText-V1)	Setup 1	En	Si	0.3328	0.4535
		Si	En	0.4088	0.5064
	Setup 2	En	Si	0.3406	0.4136
		Si	En	0.3983	0.4932
Dataset 1-V2 (En-Si-dict-large-V2)	Setup 1	En	Si	0.3666	**0.5056**
		Si	En	0.4220	0.5356
	Setup 2	En	Si	0.3772	0.4606
		Si	En	0.4068	0.5207
Dataset 2-V2 (En-Si-dict-filtered-V2)	Setup 1	En	Si	0.3781	0.4988
		Si	En	0.4809	0.5825
	Setup 2	En	Si	**0.3854**	0.4458
		Si	En	0.4620	0.5647
Dataset 3-V2 (En-Si-dict-FastText-V2)	Setup 1	En	Si	0.3766	0.4443
		Si	En	**0.4810**	0.5658
	Setup 2	En	Si	0.3838	0.4983
		Si	En	0.4617	**0.5838**

- Setup 1: Stopwords included for lookup
- Setup 2: Stopwords excluded from lookup

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Dataset Validation - TED-2020 Dataset



Dictionary	Setup	Source	Target	Average simple-lookup Score	Average NN-lookup Score
Dataset 1-V1 (En-Si-dict-large-V1)	Setup 1	En	Si	0.2253	0.4052
		Si	En	0.2950	0.4688
	Setup 2	En	Si	0.2648	0.4009
		Si	En	0.2828	0.4601
Dataset 2-V1 (En-Si-dict-filtered-V1)	Setup 1	En	Si	0.2900	0.4275
		Si	En	0.3640	0.4947
	Setup 2	En	Si	0.3385	0.4242
		Si	En	0.3501	0.4869
Dataset 3-V1 (En-Si-dict-FastText-V1)	Setup 1	En	Si	0.2900	0.4275
		Si	En	0.3662	0.4980
	Setup 2	En	Si	0.3385	0.4246
		Si	En	0.3524	0.4904
Dataset 1-V2 (En-Si-dict-large-V1)	Setup 1	En	Si	0.3296	0.5050
		Si	En	0.4003	0.5514
	Setup 2	En	Si	**0.3859**	**0.5121**
		Si	En	0.3874	0.5403
Dataset 2-V2 (En-Si-dict-filtered-V1)	Setup 1	En	Si	0.3269	0.4699
		Si	En	0.4329	0.5498
	Setup 2	En	Si	0.3804	0.4713
		Si	En	0.4190	**0.5585**
Dataset 3-V2 (En-Si-dict-FastText-V1)	Setup 1	En	Si	0.3272	0.4706
		Si	En	**0.4368**	0.5556
	Setup 2	En	Si	0.3810	0.4718
		Si	En	0.4231	0.5638

- Here we present the validation results using the English-Sinhala TED-2020 dataset
- This dataset consists of ~1k parallel sentences from TED and TED-X transcripts
- Here also, in ROUGE-1 perspective our results are in the good and moderate regions in most of the cases

- Setup 1: Stopwords included for lookup
- Setup 2: Stopwords excluded from lookup

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Publication

- Sinhala-English Parallel Word Dictionary Dataset
    - This paper was published in the 2023 IEEE 17th International Conference on Industrial and Information Systems (**ICIIS-2023**)
    - Introduces our large-scale Sinhala-English parallel dataset along with the dataset generation pipeline [43].

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Phase 2:
Traditional Embedding Alignment Techniques

Embedding Alignment Techniques - Traditional

- Orthogonal Procrustes Analysis [40]
  - The orthogonal transformation matrix is approximated using SVD components of the product of the two unaligned embedding spaces
- RCSLS [7]
  - Minimize the Cross-domain Similarity Local Scaling (CSLS) loss as the optimization criterion
  - This method tries to do a symmetric alignment addressing the so-called *Hubness Issue*
- VecMap [28]
  - Align two embedding spaces using a series of linear transformations
  - Whitening → Orthogonal Mapping → Re-weighting → De-whitening → Dimensionality Reduction
- Contrastive Alignment (C1) [29]
  - Align two embedding spaces by optimizes a special contrastive loss called InfoNCEloss [30]
  - In general, contrastive learning tries to bring the similar points closer while pushing away the the negative points
  - This is the stage 1 of a two-stage contrastive alignment technique

[40] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in International Conference on Learning Representations, 2016
[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” EMNLP, 2018.
[28] M. Artetxe, G. Labaka, and E. Agirre, “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018
[29] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vulí c, “Improving word translation via two-stage contrastive learning,” ACL, 2022
[30] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

Phase 2

Novel Alignment Dataset Generation Method

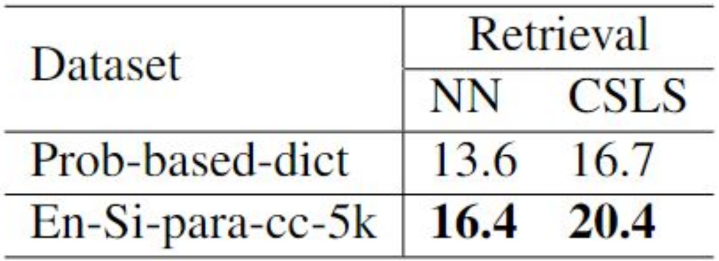
- We proposed a novel alignment dataset creation method where we do not find alignment datasets for supervised word embedding alignment
- Here we make use of available large scale corpora and find the source and target language word pairs that maximizes the coexisting probability
- The optimization criteria is as follows

$$\max_{src,tgt}\left[P\left(src|tgt\right)P\left(tgt|src\right)\right]\implies\max_{src,tgt}\left[\frac{P(src,tgt)^2}{P(source)P(target)}\right]\implies\max_{src,tgt}\left[\frac{count(src,tgt)^2}{count(src).count(tgt)}\right]$$

[44] K. Wickramasinghe and N. de Silva, “Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language,” PACLIC37, 2023

Novel Alignment Dataset Creation Method

- We have generated an alignment dataset using the method we proposed
- There we find the word pairs from large corpora that maximizes the coexisting probability and select them as translation pairs (**Prob-based-dict**)
- We built another alignment dataset using our large dictionary dataset (**En-Si-para-cc-5k**) and compared them using an embedding alignment task
- We aligned English and Sinhala FastText embeddings using the Procrustes alignment technique
- We observed that the new dataset has performed fairly well compared to the other dataset which assures the validity of our method



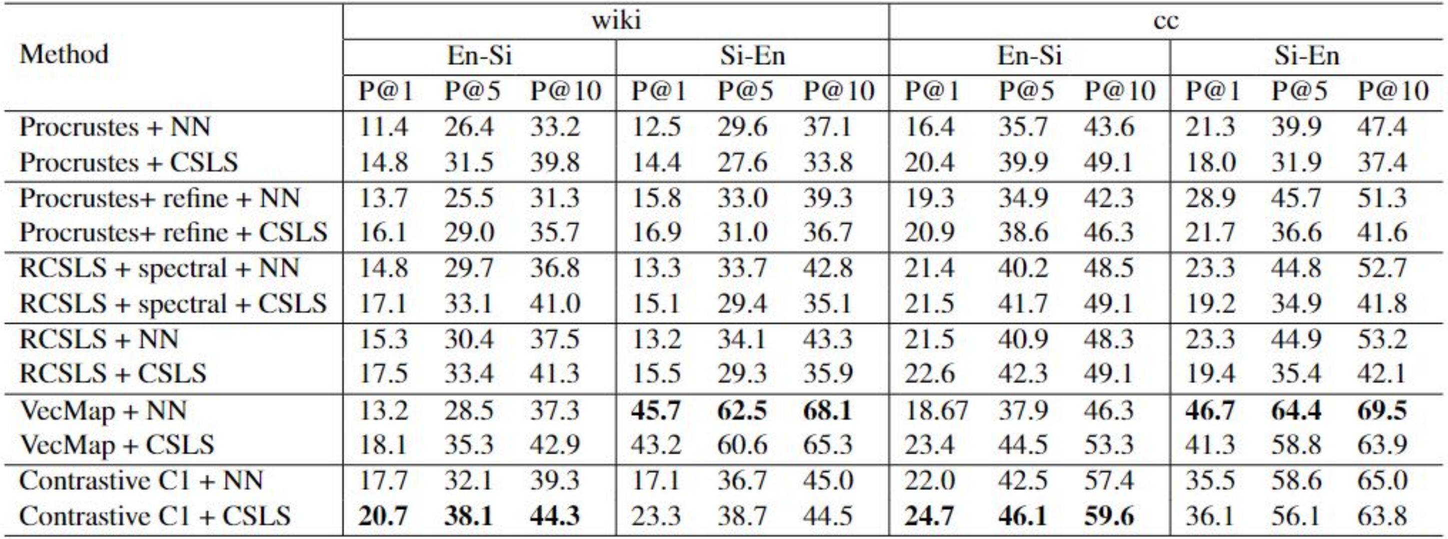
Dataset	Retrieval	
	NN	CSLS
Prob-based-dict	13.6	16.7
En-Si-para-cc-5k	**16.4**	**20.4**

BLI Alignment results of EN→Si direction with cc-Fasttext embeddings

Phase 2

English-Sinhala Word Embedding Alignment - Traditional Methods

- We have experimented with the available traditional embedding alignment techniques to align the English and Sinhala FastText word embedding spaces
- Following Table shows the @1, @5 and @10 BLI retrieval score for different traditional word embedding alignment techniques



Method	wiki						cc					
	En-Si			Si-En			En-Si			Si-En		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Procrustes + NN	11.4	26.4	33.2	12.5	29.6	37.1	16.4	35.7	43.6	21.3	39.9	47.4
Procrustes + CSLS	14.8	31.5	39.8	14.4	27.6	33.8	20.4	39.9	49.1	18.0	31.9	37.4
Procrustes+ refine + NN	13.7	25.5	31.3	15.8	33.0	39.3	19.3	34.9	42.3	28.9	45.7	51.3
Procrustes+ refine + CSLS	16.1	29.0	35.7	16.9	31.0	36.7	20.9	38.6	46.3	21.7	36.6	41.6
RCSLS + spectral + NN	14.8	29.7	36.8	13.3	33.7	42.8	21.4	40.2	48.5	23.3	44.8	52.7
RCSLS + spectral + CSLS	17.1	33.1	41.0	15.1	29.4	35.1	21.5	41.7	49.1	19.2	34.9	41.8
RCSLS + NN	15.3	30.4	37.5	13.2	34.1	43.3	21.5	40.9	48.3	23.3	44.9	53.2
RCSLS + CSLS	17.5	33.4	41.3	15.5	29.3	35.9	22.6	42.3	49.1	19.4	35.4	42.1
VecMap + NN	13.2	28.5	37.3	**45.7**	**62.5**	**68.1**	18.67	37.9	46.3	**46.7**	**64.4**	**69.5**
VecMap + CSLS	18.1	35.3	42.9	43.2	60.6	65.3	23.4	44.5	53.3	41.3	58.8	63.9
Contrastive C1 + NN	17.7	32.1	39.3	17.1	36.7	45.0	22.0	42.5	57.4	35.5	58.6	65.0
Contrastive C1 + CSLS	**20.7**	**38.1**	**44.3**	23.3	38.7	44.5	**24.7**	**46.1**	**59.6**	36.1	56.1	63.8

Publication

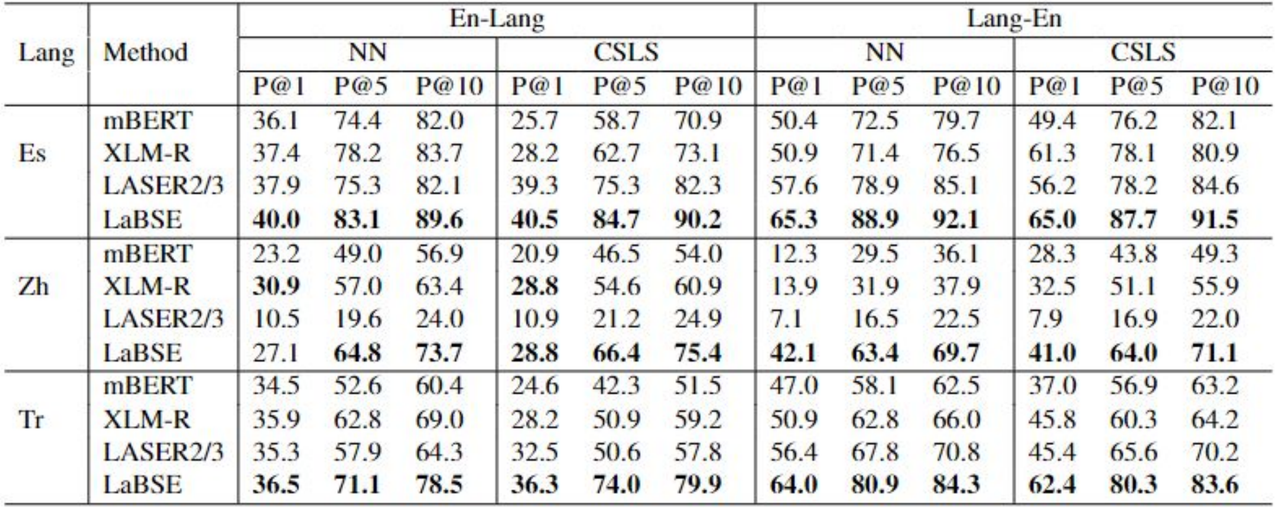
- Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language
    - This was published in the Proceedings of the 37th Pacific Asia Conference on Language, Information, and Computation (**PACLIC37-2023**) in 2023
    - Present the traditional word embedding alignment results obtained for Sinhala-English pair
    - The novel alignment dataset generation technique using large-scale parallel corpora [44].

[44] K. Wickramasinghe and N. de Silva, “Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language,” PACLIC37, 2023

Phase 3:
Multilingual Embeddings and Combined Alignment Techniques

Multilingual Embedding Alignment Evaluation

- We carried out the BLI evaluation for the currently best performing multilingual embedding models
- We selected the best model out of them for the next experiments
- From the results we selected **LaBSE** as the candidate multilingual model for rest of the experiments



Lang	Method	En-Lang						Lang-En					
		NN			CSLS			NN			CSLS		
		P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Es	mBERT	36.1	74.4	82.0	25.7	58.7	70.9	50.4	72.5	79.7	49.4	76.2	82.1
	XLM-R	37.4	78.2	83.7	28.2	62.7	73.1	50.9	71.4	76.5	61.3	78.1	80.9
	LASER2/3	37.9	75.3	82.1	39.3	75.3	82.3	57.6	78.9	85.1	56.2	78.2	84.6
	LaBSE	**40.0**	**83.1**	**89.6**	**40.5**	**84.7**	**90.2**	**65.3**	**88.9**	**92.1**	**65.0**	**87.7**	**91.5**
Zh	mBERT	23.2	49.0	56.9	20.9	46.5	54.0	12.3	29.5	36.1	28.3	43.8	49.3
	XLM-R	**30.9**	57.0	63.4	**28.8**	54.6	60.9	13.9	31.9	37.9	32.5	51.1	55.9
	LASER2/3	10.5	19.6	24.0	10.9	21.2	24.9	7.1	16.5	22.5	7.9	16.9	22.0
	LaBSE	27.1	**64.8**	**73.7**	**28.8**	**66.4**	**75.4**	**42.1**	**63.4**	**69.7**	**41.0**	**64.0**	**71.1**
Tr	mBERT	34.5	52.6	60.4	24.6	42.3	51.5	47.0	58.1	62.5	37.0	56.9	63.2
	XLM-R	35.9	62.8	69.0	28.2	50.9	59.2	50.9	62.8	66.0	45.8	60.3	64.2
	LASER2/3	35.3	57.9	64.3	32.5	50.6	57.8	56.4	67.8	70.8	45.4	65.6	70.2
	LaBSE	**36.5**	**71.1**	**78.5**	**36.3**	**74.0**	**79.9**	**64.0**	**80.9**	**84.3**	**62.4**	**80.3**	**83.6**

Embedding Alignment Techniques - Multilingual and Hybrid

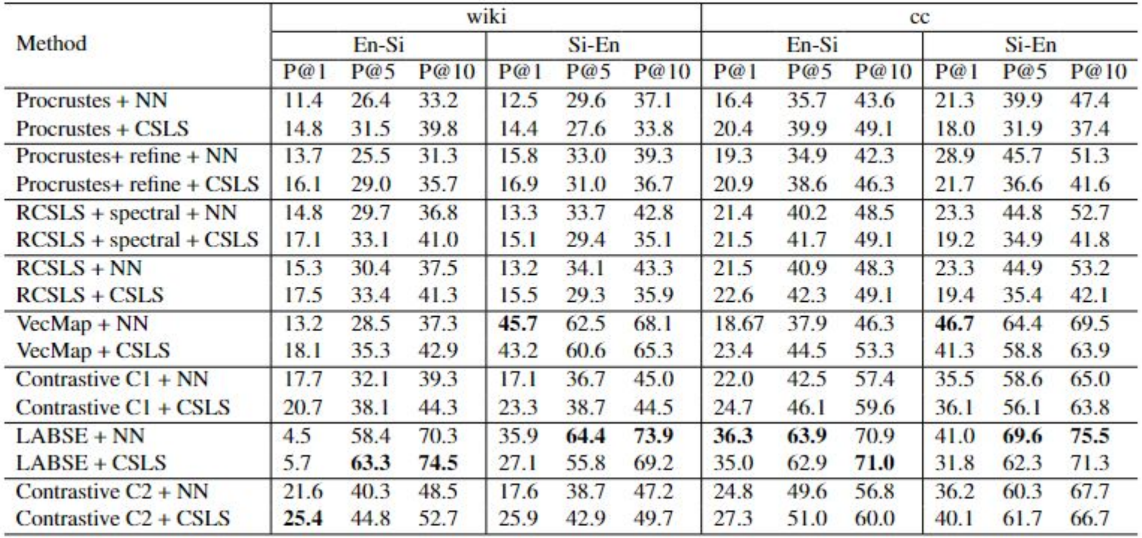
- LaBSE
  - LaBSE is a transformer-based multilingual sentence embedding model
  - 112 different languages share a common embedding space in LaBSE
  - Therefore LaBSE embeddings are already in the aligned state
- Two-stage Contrastive Alignment (C2) [29]
  - This is the second stage of C1 contanstive alignment we talked in the previous slide
  - Here, using the alignment dataset and the stage 1 aligned embeddings, positive and negative pairs are selected for contrastive fine tuning of a multilingual embedding model
  - The C1 embedding space is then re-mapped onto the fine tuned C2 space
  - The final aligned embedding representation of an input word is calculated as a linear combination of its C1-based vector mapped to a C2, and its C2-based vector

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic bert sentence embedding,” ACL, 2022.
[29] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vulí c, “Improving word translation via two-stage contrastive learning,” ACL, 2022

Phase 3

English-Sinhala Word Embedding Alignment - All Methods

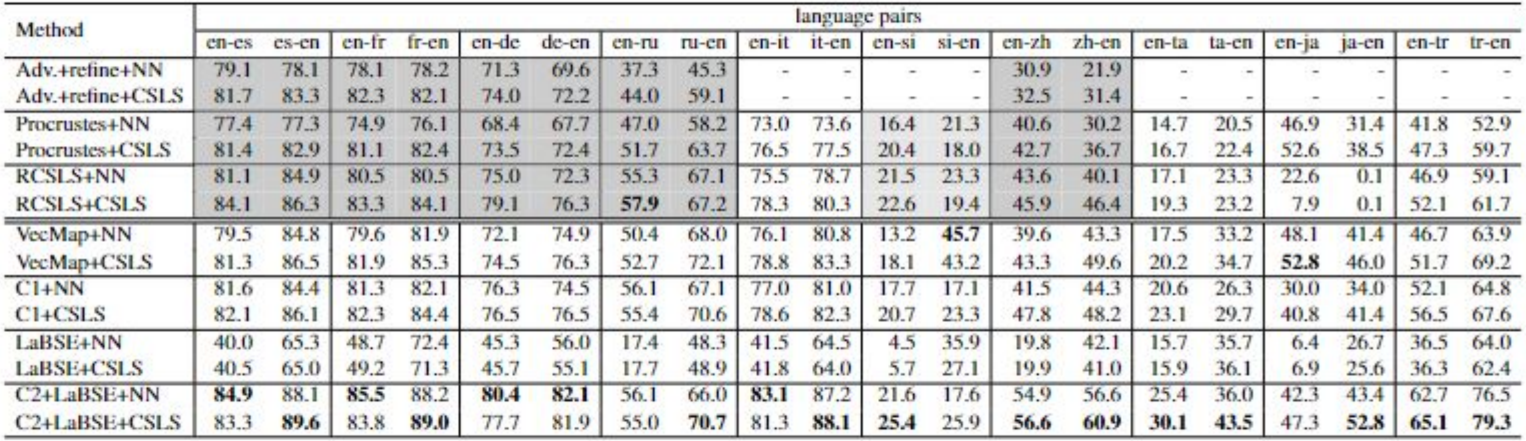
- Here we present all the techniques we have tried to align the English and Sinhala FastText word embedding spaces
- Following Table shows the @1, @5 and @10 BLI retrieval score for different traditional word embedding alignment techniques



Method	wiki						cc					
	En-Si			Si-En			En-Si			Si-En		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Procrustes + NN	11.4	26.4	33.2	12.5	29.6	37.1	16.4	35.7	43.6	21.3	39.9	47.4
Procrustes + CSLS	14.8	31.5	39.8	14.4	27.6	33.8	20.4	39.9	49.1	18.0	31.9	37.4
Procrustes+ refine + NN	13.7	25.5	31.3	15.8	33.0	39.3	19.3	34.9	42.3	28.9	45.7	51.3
Procrustes+ refine + CSLS	16.1	29.0	35.7	16.9	31.0	36.7	20.9	38.6	46.3	21.7	36.6	41.6
RCSLS + spectral + NN	14.8	29.7	36.8	13.3	33.7	42.8	21.4	40.2	48.5	23.3	44.8	52.7
RCSLS + spectral + CSLS	17.1	33.1	41.0	15.1	29.4	35.1	21.5	41.7	49.1	19.2	34.9	41.8
RCSLS + NN	15.3	30.4	37.5	13.2	34.1	43.3	21.5	40.9	48.3	23.3	44.9	53.2
RCSLS + CSLS	17.5	33.4	41.3	15.5	29.3	35.9	22.6	42.3	49.1	19.4	35.4	42.1
VecMap + NN	13.2	28.5	37.3	**45.7**	62.5	68.1	18.67	37.9	46.3	**46.7**	64.4	69.5
VecMap + CSLS	18.1	35.3	42.9	43.2	60.6	65.3	23.4	44.5	53.3	41.3	58.8	63.9
Contrastive C1 + NN	17.7	32.1	39.3	17.1	36.7	45.0	22.0	42.5	57.4	35.5	58.6	65.0
Contrastive C1 + CSLS	20.7	38.1	44.3	23.3	38.7	44.5	24.7	46.1	59.6	36.1	56.1	63.8
LABSE + NN	4.5	58.4	70.3	35.9	**64.4**	**73.9**	**36.3**	**63.9**	70.9	41.0	**69.6**	**75.5**
LABSE + CSLS	5.7	**63.3**	**74.5**	27.1	55.8	69.2	35.0	62.9	**71.0**	31.8	62.3	71.3
Contrastive C2 + NN	21.6	40.3	48.5	17.6	38.7	47.2	24.8	49.6	56.8	36.2	60.3	67.7
Contrastive C2 + CSLS	**25.4**	44.8	52.7	25.9	42.9	49.7	27.3	51.0	60.0	40.1	61.7	66.7

Study on Other Languages

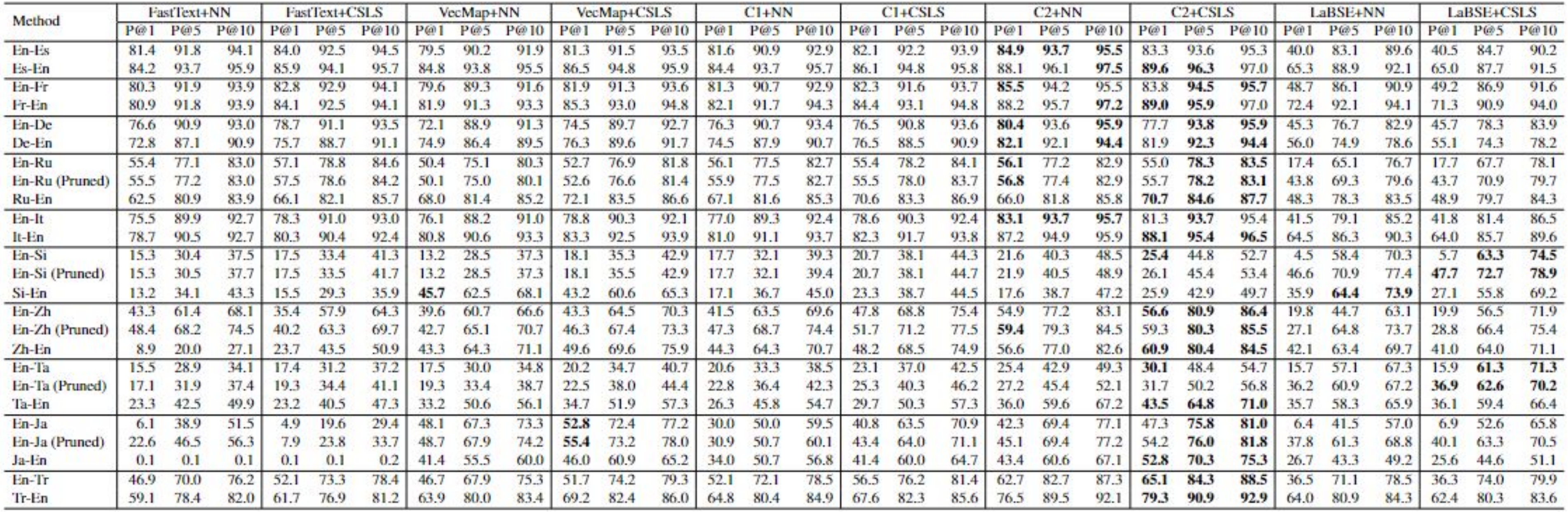
- We extended our experiments to 8 other languages to ensure the consistency of the results
- Following table shows the @1 BLI scores for 10 language pairs



Method	language pairs																			
	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-it	it-en	en-si	si-en	en-zh	zh-en	en-ta	ta-en	en-ja	ja-en	en-tr	tr-en
Adv.+refine+NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	45.3	-	-	-	-	30.9	21.9	-	-	-	-	-	-
Adv.+refine+CSLS	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	-	-	-	-	32.5	31.4	-	-	-	-	-	-
Procrustes+NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	73.0	73.6	16.4	21.3	40.6	30.2	14.7	20.5	46.9	31.4	41.8	52.9
Procrustes+CSLS	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	76.5	77.5	20.4	18.0	42.7	36.7	16.7	22.4	52.6	38.5	47.3	59.7
RCSLS+NN	81.1	84.9	80.5	80.5	75.0	72.3	55.3	67.1	75.5	78.7	21.5	23.3	43.6	40.1	17.1	23.3	22.6	0.1	46.9	59.1
RCSLS+CSLS	84.1	86.3	83.3	84.1	79.1	76.3	**57.9**	67.2	78.3	80.3	22.6	19.4	45.9	46.4	19.3	23.2	7.9	0.1	52.1	61.7
VecMap+NN	79.5	84.8	79.6	81.9	72.1	74.9	50.4	68.0	76.1	80.8	13.2	**45.7**	39.6	43.3	17.5	33.2	48.1	41.4	46.7	63.9
VecMap+CSLS	81.3	86.5	81.9	85.3	74.5	76.3	52.7	72.1	78.8	83.3	18.1	43.2	43.3	49.6	20.2	34.7	**52.8**	46.0	51.7	69.2
C1+NN	81.6	84.4	81.3	82.1	76.3	74.5	56.1	67.1	77.0	81.0	17.7	17.1	41.5	44.3	20.6	26.3	30.0	34.0	52.1	64.8
C1+CSLS	82.1	86.1	82.3	84.4	76.5	76.5	55.4	70.6	78.6	82.3	20.7	23.3	47.8	48.2	23.1	29.7	40.8	41.4	56.5	67.6
LaBSE+NN	40.0	65.3	48.7	72.4	45.3	56.0	17.4	48.3	41.5	64.5	4.5	35.9	19.8	42.1	15.7	35.7	6.4	26.7	36.5	64.0
LaBSE+CSLS	40.5	65.0	49.2	71.3	45.7	55.1	17.7	48.9	41.8	64.0	5.7	27.1	19.9	41.0	15.9	36.1	6.9	25.6	36.3	62.4
C2+LaBSE+NN	**84.9**	88.1	**85.5**	88.2	**80.4**	**82.1**	56.1	66.0	**83.1**	87.2	21.6	17.6	54.9	56.6	25.4	36.0	42.3	43.4	62.7	76.5
C2+LaBSE+CSLS	83.3	**89.6**	83.8	**89.0**	77.7	81.9	55.0	**70.7**	81.3	**88.1**	**25.4**	25.9	**56.6**	**60.9**	**30.1**	**43.5**	47.3	**52.8**	**65.1**	**79.3**

Study on Other Languages (cont.)

- This table has the extended results for all language pairs and all alignment techniques



Method	FastText+NN			FastText+CSLS			VecMap+NN			VecMap+CSLS			C1+NN			C1+CSLS			C2+NN			C2+CSLS			LaBSE+NN			LaBSE+CSLS		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
En-Es	81.4	91.8	94.1	84.0	92.5	94.5	79.5	90.2	91.9	81.3	91.5	93.5	81.6	90.9	92.9	82.1	92.2	93.9	**84.9**	**93.7**	**95.5**	83.3	93.6	95.3	40.0	83.1	89.6	40.5	84.7	90.2
Es-En	84.2	93.7	95.9	85.9	94.1	95.7	84.8	93.8	95.5	86.5	94.8	95.9	84.4	93.7	95.7	86.1	94.8	95.8	88.1	96.1	**97.5**	**89.6**	**96.3**	97.0	65.3	88.9	92.1	65.0	87.7	91.5
En-Fr	80.3	91.9	93.9	82.8	92.9	94.1	79.6	89.3	91.6	81.9	91.3	93.6	81.3	90.7	92.9	82.3	91.6	93.7	**85.5**	94.2	95.5	83.8	**94.5**	**95.7**	48.7	86.1	90.9	49.2	86.9	91.6
Fr-En	80.9	91.8	93.9	84.1	92.5	94.1	81.9	91.3	93.3	85.3	93.0	94.8	82.1	91.7	94.3	84.4	93.1	94.8	88.2	95.7	**97.2**	**89.0**	**95.9**	97.0	72.4	92.1	94.1	71.3	90.9	94.0
En-De	76.6	90.9	93.0	78.7	91.1	93.5	72.1	88.9	91.3	74.5	89.7	92.7	76.3	90.7	93.4	76.5	90.8	93.6	**80.4**	93.6	**95.9**	77.7	**93.8**	**95.9**	45.3	76.7	82.9	45.7	78.3	83.9
De-En	72.8	87.1	90.9	75.7	88.7	91.1	74.9	86.4	89.5	76.3	89.6	91.7	74.5	87.9	90.7	76.5	88.5	90.9	**82.1**	92.1	**94.4**	81.9	**92.3**	**94.4**	56.0	74.9	78.6	55.1	74.3	78.2
En-Ru	55.4	77.1	83.0	57.1	78.8	84.6	50.4	75.1	80.3	52.7	76.9	81.8	56.1	77.5	82.7	55.4	78.2	84.1	**56.1**	77.2	82.9	55.0	**78.3**	**83.5**	17.4	65.1	76.7	17.7	67.7	78.1
En-Ru (Pruned)	55.5	77.2	83.0	57.5	78.6	84.2	50.1	75.0	80.1	52.6	76.6	81.4	55.9	77.5	82.7	55.5	78.0	83.7	**56.8**	77.4	82.9	55.7	**78.2**	**83.1**	43.8	69.3	79.6	43.7	70.9	79.7
Ru-En	62.5	80.9	83.9	66.1	82.1	85.7	68.0	81.4	85.2	72.1	83.5	86.6	67.1	81.6	85.3	70.6	83.3	86.9	66.0	81.8	85.8	**70.7**	**84.6**	**87.7**	48.3	78.3	83.5	48.9	79.7	84.3
En-It	75.5	89.9	92.7	78.3	91.0	93.0	76.1	88.2	91.0	78.8	90.3	92.1	77.0	89.3	92.4	78.6	90.3	92.4	**83.1**	**93.7**	**95.7**	81.3	**93.7**	95.4	41.5	79.1	85.2	41.8	81.4	86.5
It-En	78.7	90.5	92.7	80.3	90.4	92.4	80.8	90.6	93.3	83.3	92.5	93.9	81.0	91.1	93.7	82.3	91.7	93.8	87.2	94.9	95.9	**88.1**	**95.4**	**96.5**	64.5	86.3	90.3	64.0	85.7	89.6
En-Si	15.3	30.4	37.5	17.5	33.4	41.3	13.2	28.5	37.3	18.1	35.3	42.9	17.7	32.1	39.3	20.7	38.1	44.3	21.6	40.3	48.5	**25.4**	44.8	52.7	4.5	58.4	70.3	5.7	**63.3**	**74.5**
En-Si (Pruned)	15.3	30.5	37.7	17.5	33.5	41.7	13.2	28.5	37.3	18.1	35.5	42.9	17.7	32.1	39.4	20.7	38.1	44.7	21.9	40.5	48.9	26.1	45.4	53.4	46.6	70.9	77.4	**47.7**	**72.7**	**78.9**
Si-En	13.2	34.1	43.3	15.5	29.3	35.9	**45.7**	62.5	68.1	43.2	60.6	65.3	17.1	36.7	45.0	23.3	38.7	44.5	17.6	38.7	47.2	25.9	42.9	49.7	35.9	**64.4**	**73.9**	27.1	55.8	69.2
En-Zh	43.3	61.4	68.1	35.4	57.9	64.3	39.6	60.7	66.6	43.3	64.5	70.3	41.5	63.5	69.6	47.8	68.8	75.4	54.9	77.2	83.1	**56.6**	**80.9**	**86.4**	19.8	44.7	63.1	19.9	56.5	71.9
En-Zh (Pruned)	48.4	68.2	74.5	40.2	63.3	69.7	42.7	65.1	70.7	46.3	67.4	73.3	47.3	68.7	74.4	51.7	71.2	77.5	**59.4**	79.3	84.5	59.3	**80.3**	**85.5**	27.1	64.8	73.7	28.8	66.4	75.4
Zh-En	8.9	20.0	27.1	23.7	43.5	50.9	43.3	64.3	71.1	49.6	69.6	75.9	44.3	64.3	70.7	48.2	68.5	74.9	56.6	77.0	82.6	**60.9**	**80.4**	**84.5**	42.1	63.4	69.7	41.0	64.0	71.1
En-Ta	15.5	28.9	34.1	17.4	31.2	37.2	17.5	30.0	34.8	20.2	34.7	40.7	20.6	33.3	38.5	23.1	37.0	42.5	25.4	42.9	49.3	**30.1**	48.4	54.7	15.7	57.1	67.3	15.9	**61.3**	**71.3**
En-Ta (Pruned)	17.1	31.9	37.4	19.3	34.4	41.1	19.3	33.4	38.7	22.5	38.0	44.4	22.8	36.4	42.3	25.3	40.3	46.2	27.2	45.4	52.1	31.7	50.2	56.8	36.2	60.9	67.2	**36.9**	**62.6**	**70.2**
Ta-En	23.3	42.5	49.9	23.2	40.5	47.3	33.2	50.6	56.1	34.7	51.9	57.3	26.3	45.8	54.7	29.7	50.3	57.3	36.0	59.6	67.2	**43.5**	**64.8**	**71.0**	35.7	58.3	65.9	36.1	59.4	66.4
En-Ja	6.1	38.9	51.5	4.9	19.6	29.4	48.1	67.3	73.3	**52.8**	72.4	77.2	30.0	50.0	59.5	40.8	63.5	70.9	42.3	69.4	77.1	47.3	**75.8**	**81.0**	6.4	41.5	57.0	6.9	52.6	65.8
En-Ja (Pruned)	22.6	46.5	56.3	7.9	23.8	33.7	48.7	67.9	74.2	**55.4**	73.2	78.0	30.9	50.7	60.1	43.4	64.0	71.1	45.1	69.4	77.2	54.2	**76.0**	**81.8**	37.8	61.3	68.8	40.1	63.3	70.5
Ja-En	0.1	0.1	0.1	0.1	0.1	0.2	41.4	55.5	60.0	46.0	60.9	65.2	34.0	50.7	56.8	41.4	60.0	64.7	43.4	60.6	67.1	**52.8**	**70.3**	**75.3**	26.7	43.3	49.2	25.6	44.6	51.1
En-Tr	46.9	70.0	76.2	52.1	73.3	78.4	46.7	67.9	75.3	51.7	74.2	79.3	52.1	72.1	78.5	56.5	76.2	81.4	62.7	82.7	87.3	**65.1**	**84.3**	**88.5**	36.5	71.1	78.5	36.3	74.0	79.9
Tr-En	59.1	78.4	82.0	61.7	76.9	81.2	63.9	80.0	83.4	69.2	82.4	86.0	64.8	80.4	84.9	67.6	82.3	85.6	76.5	89.5	92.1	**79.3**	**90.9**	**92.9**	64.0	80.9	84.3	62.4	80.3	83.6

Improvements to BLI

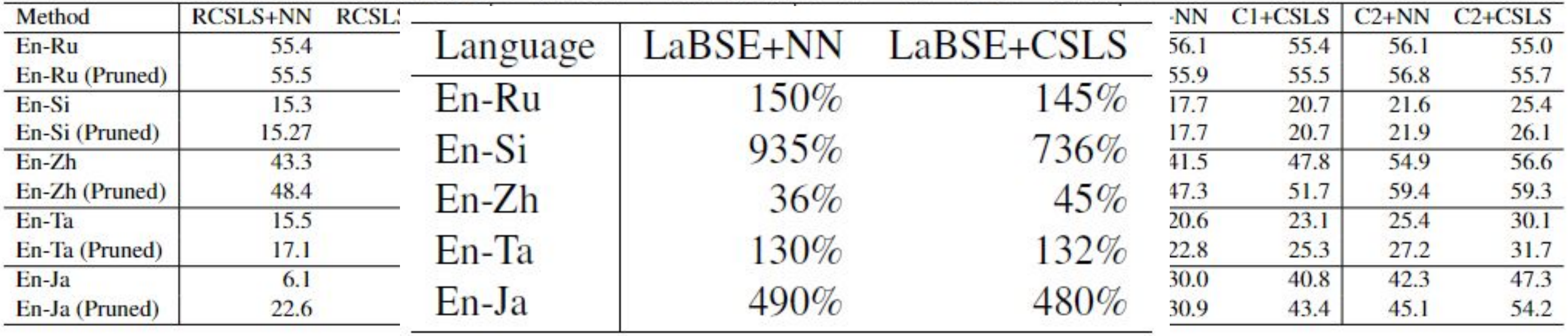
- We measure the degree of alignment between two embedding spaces using Bilingual Lexicon Induction (BLI)
- BLI try to retrieve target translations of given source words from two aligned embedding spaces
- We observed the standard BLI task lacks giving a good measure of the degree of the alignment specially for inflected language case and multilingual model embedding case
- We propose two amendments for BLI in above two cases

Improvements to BLI - Vocabulary Pruning

- The vocabulary of any embedding model is not pure.
- It contains words/tokens of some other languages as well due to the mixed usage of different languages together (code-mixed).
- We wanted to evaluate how good the alignment is considering only that exact language.
- When it comes to BLI, code-mixed usage can negatively affect the results if we consider only the top-1 match. This becomes even worse in multilingual model cases since multilingual models could give the top-1 match from any language.
- The simple solution we propose for this issue is pruning the vocabulary by removing tokens that do not belong to the language of interest before BLI evaluation. Therefore vocabulary pruning idea removes the unnecessary burden added by code-mixed vocabularies.
- This can be simply done for languages that have different character sets than English alphabetical characters.
- For that, we simply removed all the ASCII characters from the FastText vocabularies of Russian (Ru), Sinhala (Si), Chinese (Zh), Tamil (Ta), and Japanese (Ja) and conducted the evaluation.

Improvements to BLI - Vocabulary Pruning

- Here we present how our proposed vocabulary pruning for BLI is effective especially when it comes to multilingual embedding models
- We observe a massive improvement in LaBSE embeddings when vocabulary pruning is used
- The minimum @1 improvements for LaBSE embeddings are **Ru-145%, Si-736%, Zh-36%, Ta-130%, Ja-480%**

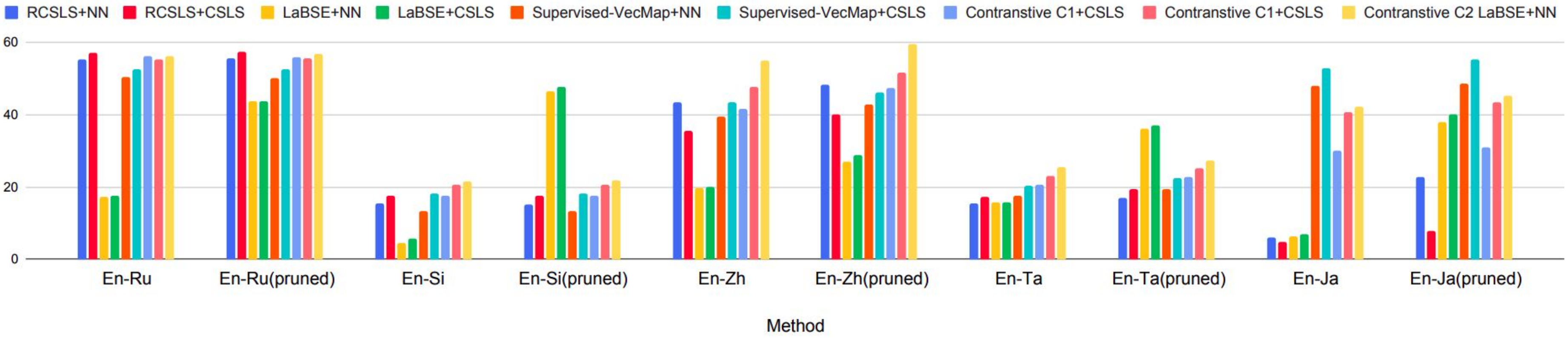


Method	RCSLS+NN	RCSL	NN	C1+CSLS	C2+NN	C2+CSLS
En-Ru	55.4		56.1	55.4	56.1	55.0
En-Ru (Pruned)	55.5		55.9	55.5	56.8	55.7
En-Si	15.3		17.7	20.7	21.6	25.4
En-Si (Pruned)	15.27		17.7	20.7	21.9	26.1
En-Zh	43.3		41.5	47.8	54.9	56.6
En-Zh (Pruned)	48.4		47.3	51.7	59.4	59.3
En-Ta	15.5		20.6	23.1	25.4	30.1
En-Ta (Pruned)	17.1		22.8	25.3	27.2	31.7
En-Ja	6.1		30.0	40.8	42.3	47.3
En-Ja (Pruned)	22.6		30.9	43.4	45.1	54.2

Language	LaBSE+NN	LaBSE+CSLS
En-Ru	150%	145%
En-Si	935%	736%
En-Zh	36%	45%
En-Ta	130%	132%
En-Ja	490%	480%

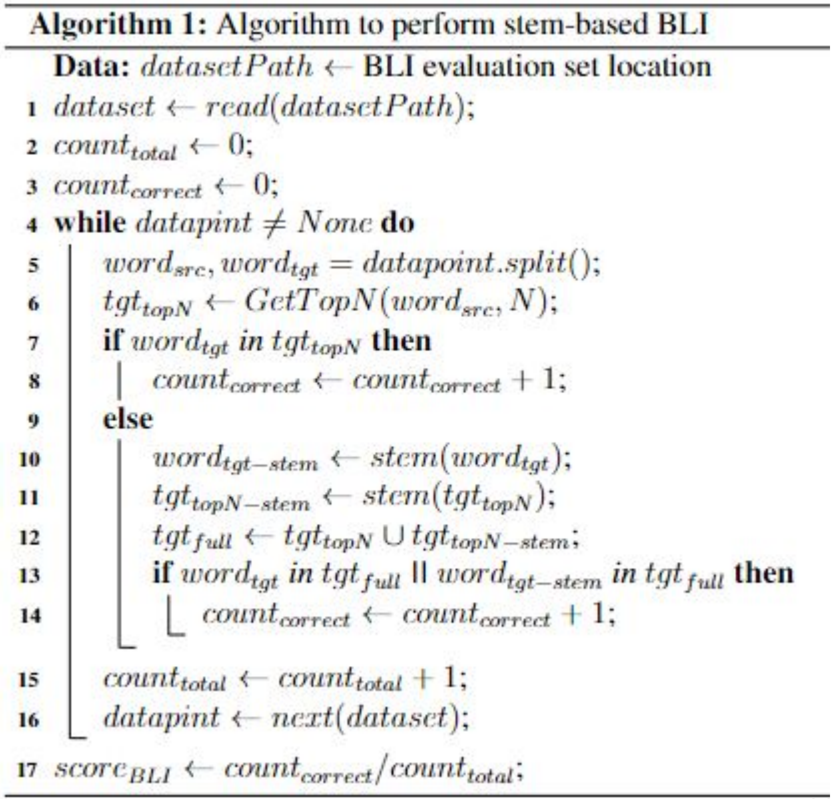
Improvements to BLI - Vocabulary Pruning (cont.)

- Following chart shows the same results graphically
- We can see that LaBSE embeddings (Yellow and Green) have gained a huge improvement compared to other techniques after the pruning



Improvements to BLI - Stem-based BLI

- Languages such as Sinhala (Si) are highly inflected [8] and therefore contain many variations for a given single word.
- Even though the model alignment happens properly it may not be properly reflected through the BLI task due to not having exact matches as expected in the test sets.
- Therefore we experimented with a soft matching rather than performing an exact match.
- Here instead of finding the exact Sinhala translation for a given English word from the test set, we query for the stem of the expected Sinhala translation word
- We referred to this as **Stem-based BLI**

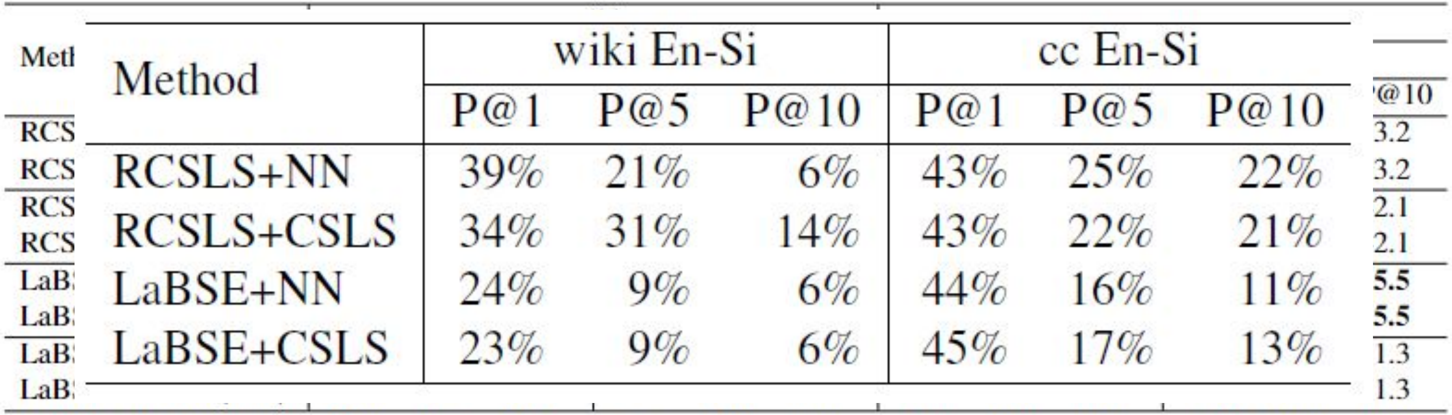


Algorithm 1: Algorithm to perform stem-based BLI

```
Data: $datasetPath \leftarrow$ BLI evaluation set location
$dataset \leftarrow read(datasetPath)$;
$count_{total} \leftarrow 0$;
$count_{correct} \leftarrow 0$;
while $datapint \neq None$ do
    $word_{src}, word_{tgt} = datapoint.split()$;
    $tgt_{topN} \leftarrow GetTopN(word_{src}, N)$;
    if $word_{tgt}$ in $tgt_{topN}$ then
        $count_{correct} \leftarrow count_{correct} + 1$;
    else
        $word_{tgt-stem} \leftarrow stem(word_{tgt})$;
        $tgt_{topN-stem} \leftarrow stem(tgt_{topN})$;
        $tgt_{full} \leftarrow tgt_{topN} \cup tgt_{topN-stem}$;
        if $word_{tgt}$ in $tgt_{full}$ || $word_{tgt-stem}$ in $tgt_{full}$ then
            $count_{correct} \leftarrow count_{correct} + 1$;
    $count_{total} \leftarrow count_{total} + 1$;
    $datapint \leftarrow next(dataset)$;
$score_{BLI} \leftarrow count_{correct}/count_{total}$;
```

Improvements to BLI - Stem-based BLI

- Here the impact of our proposed **Stem-based BLI** (given as ***soft*** in the table) has been presented
- The table shows how stem-based BLI is appropriate for evaluating inflected languages like Sinhala
- We observe **23%, 9%, 6%** minimum improvements for @1, @5 and @10 BLI scores respectively
- The stem-based matching is not needed for Si-En direction since En is not a highly inflected language [45]

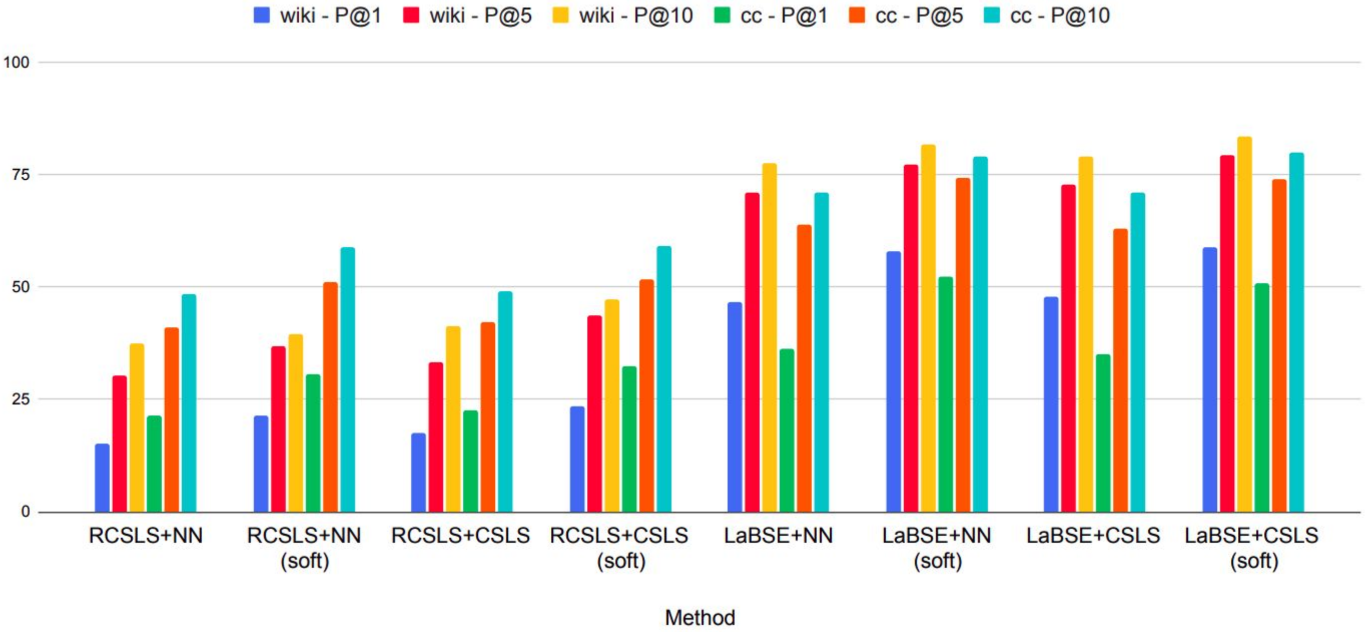


Method	wiki En-Si			cc En-Si		
	P@1	P@5	P@10	P@1	P@5	P@10
RCSLS+NN	39%	21%	6%	43%	25%	22%
RCSLS+CSLS	34%	31%	14%	43%	22%	21%
LaBSE+NN	24%	9%	6%	44%	16%	11%
LaBSE+CSLS	23%	9%	6%	45%	17%	13%

[45] M. S. Mauˇcec, Z. Kaˇciˇc, and B. Horvat, “Modelling highly inflected languages,” Information Sciences, 2004

Improvements to BLI - Stem-based BLI (cont.)

- Following chart shows the graphical representation of the results.



[45] M. S. Mauˇcec, Z. Kaˇciˇc, and B. Horvat, “Modelling highly inflected languages,” Information Sciences, 2004

Publication

- Aligned or Multilingual: A Comparative Study in Word Embedding Paradigms (under review)
  - This is our next paper which we are willing to propose in The 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP-2024**).
  - This paper discusses how good traditional monolingual alignment techniques, multilingual embeddings, and hybrid alignment techniques in terms of the different criteria.
  - Also, we introduce two novel modifications for the standard Bilingual Lexicon Induction to measure the degree of alignment more realistically in certain scenarios.

[44] K. Wickramasinghe and N. de Silva, "Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language," PACLIC37, 2023

Summary

Summary

- Built large-scale Sinhala-English dictionary dataset with ~1.3M translation pairs
- Achieved the best available Sinhala-English FastText word embedding alignment results so far
- Proposed a novel conditional probability statistics based alignment dictionary building technique using large parallel corpora
- Experimented embedding alignment with 10 language-pairs
- Evaluated the traditional word embedding alignment techniques, multilingual embeddings and combined alignment techniques

Summary (cont.)

- Study the effectiveness of using BLI for measuring the degree of the alignment of embedding spaces
  - Found that standard BLI does not measure the True Degree of Alignment in certain cases
  - Proposed a vocabulary pruning for BLI which gives better insight when evaluating multilingual word embeddings
  - Proposed a stem-based BLI technique for evaluating the alignment of inflected languages

Publications

Phase 1

Sinhala-English Parallel Word Dictionary Dataset

- This paper was published in the 2023 IEEE 17th International Conference on Industrial and Information Systems (**ICIIS-2023**) which introduces our large-scale Sinhala-English parallel dataset along with the dataset generation pipeline [43].

Phase 2

Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language

- This was published in the Proceedings of the 37th Pacific Asia Conference on Language, Information, and Computation (**PACLIC37-2023**) in 2023 which introduces a part of our Sinhala-English embedding alignment results and a novel alignment dataset generation technique using large-scale parallel corpora [44].

Phase 3

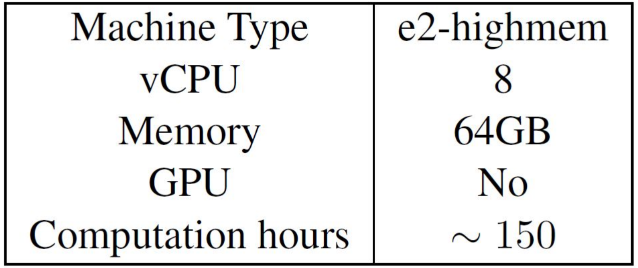
Aligned or Multilingual: A Comparative Study in Word Embedding Paradigms (under review)

- This is our next paper which we are willing to propose in The 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP-2024**). This paper discusses how good traditional monolingual alignment techniques, multilingual embeddings, and hybrid alignment techniques in terms of the different criteria.
- Also, we introduce two novel modifications for the standard Bilingual Lexicon Induction to measure the degree of alignment more realistically in certain scenarios.

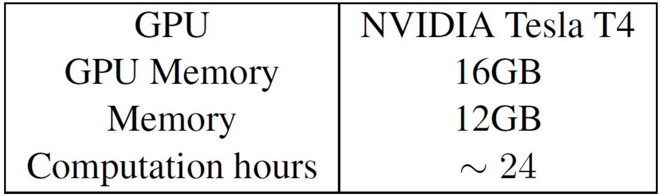
[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023
[44] K. Wickramasinghe and N. de Silva, “Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language,” PACLIC37, 2023

Computation Cost

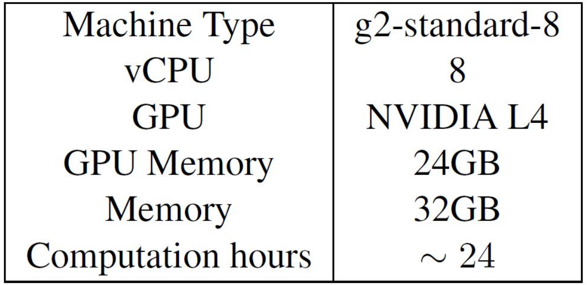
- For monolingual embedding alignment tasks, we used Google Cloud Platform (GCP) Virtual Machine (VM). 1st table shows the specifications and the computation cost of the environment.
- For multilingual model experiments, we used Google Colab Free Tier GPU environment. The 2nd table shows the specifications and the computation cost of the environment.
- For C2 combined alignment, we used a GCP VM equipped with an Nvidia GPU. The 3rd table has the relevant information



Machine Type	e2-highmem
vCPU	8
Memory	64GB
GPU	No
Computation hours	~ 150



GPU	NVIDIA Tesla T4
GPU Memory	16GB
Memory	12GB
Computation hours	~ 24



Machine Type	g2-standard-8
vCPU	8
GPU	NVIDIA L4
GPU Memory	24GB
Memory	32GB
Computation hours	~ 24

Acknowledgement

- My supervisor, Dr. Nisansa de Silva for guiding me throughout the research
- Staff members of the Department of Computer Science and Engineering for conducting lectures and exams and for all the major and minor support given to make the MSc program a success
- Emojot (Pvt) Ltd for providing the GCP GPU access for a part of the research experiments
- All the staff of the University of Moratuwa who support making this MSc program a success
- My colleague MSc students for all the help and support
- Panel of Examiners for evaluating our effort
- Finally, I acknowledge the research community and the open-source community worldwide for helping to take forward scientific research

References

References

[1] A. Kalinowski and Y. An, ‘A Survey of Embedding Space Alignment Methods for Language and Knowledge Graphs’, arXiv preprint arXiv:2010. 13688, 2020.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic bert sentence embedding,” ACL,2022.

[3] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” NIPS, 2019.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” ACL, 2020.

[5] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” arXiv preprint arXiv:1309.4168, 2013.

[6] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” NAACL, 2015.

[7] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” EMNLP, 2018.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” EMNLP, 2014.

[10] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the association for computational linguistics, vol. 5, pp. 135–146, 2017.

[12] H. T. Ng and H. B. Lee, ‘Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach’, arXiv preprint cmp-lg/9606032, 1996.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” 2018. [Online].Available: https://arxiv.org/abs/1802.05365

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” NIPS, 2017.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” arXiv preprint arXiv:1909.11942, 2019.

[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pretraining text encoders as discriminators rather than generators,” arXiv preprint arXiv:2003.10555, 2020.

[19] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” EMNLP, 2019.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’, arXiv preprint arXiv:1910. 01108, 2019.

[21] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, ‘MPNet: Masked and Permuted Pre-training for Language Understanding’, NIPS, 2020.

[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, ‘Xlnet: Generalized autoregressive pretraining for language understanding’, NIPS, 2019.

1 https://www.sbert.net/docs/pretrained_models.html

References

[23] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,”, EMNLP,2020

[24] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,”, ACL,2019

[25] . Heffernan, O. Çelebi, and H. Schwenk, “Bitext mining using distilled sentence representations for low-resource languages,” EMNLP,2022

[26] E. Grave, A. Joulin, and Q. Berthet, “Unsupervised alignment of embeddings with wasserstein procrustes,” in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 1880–1890.

[27] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” arXiv preprint arXiv:1710.04087, 2017.

[28] M. Artetxe, G. Labaka, and E. Agirre, “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018

[29] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vuli´c, “Improving word translation via two-stage contrastive learning,” ACL, 2022

[30] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

[31] N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,” arXiv preprint arXiv:1906.02358, 2019.

[32] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, “The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english,” EMNLP, 2019.

[33] M. R. Costa-jussà et al., ‘No language left behind: Scaling human-centered machine translation’, arXiv preprint arXiv:2207. 04672, 2022.

[34] R. A. Hameed, N. Pathirennehelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, “Automatic creation of a sentence aligned sinhala-tamil parallel corpus,” in Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), 2016, pp. 124–132.

[35] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn et al., “Paracrawl: Web-scale acquisition of parallel corpora,” ACL, 2020.

[36] C. Vasantharajan and U. Thayasivam, “Tamizhi-net ocr: Creating a quality large scale tamil-sinhala-english parallel corpus using deep learning based printed character recognition (pcr),” arXiv preprint arXiv:2109.05952, 2021.

[37] A. Wasala and R. Weerasinghe, “Ensitip: a tool to unlock the english web,” in 11th international conference on humans and computers, Nagaoka University of Technology, Japan, 2008

[38] A. Irvine and C. Callison-Burch, “A comprehensive analysis of bilingual lexicon Induction,” ACL, 2017

[39] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, “XNLI: Evaluating cross-lingual sentence representations,” EMNLP, 2018

[40] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in International Conference on Learning Representations, 2016

[41] G. Glavaš, R. Litschko, S. Ruder, and I. Vuli´c, “How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions,” ACL, 2019

[42] I. Vuli´c, G. Glavaš, R. Reichart, and A. Korhonen, “Do we really need fully unsupervised cross-lingual embeddings?” EMNLP, 2019

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

[44] K. Wickramasinghe and N. de Silva, “Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language,” PACLIC37, 2023

[45] M. S. Mauˇcec, Z. Kaˇciˇc, and B. Horvat, “Modelling highly inflected languages,” Information Sciences, 2004

[46]. Liyanage, S. Ranathunga, and S. Jayasena, “Bilingual lexical induction for sinhala-english using cross lingual embedding spaces,” in 2021 Moratuwa Engineering Research Conference (MERCon), 2021, pp. 579–584.

Thank You

Questions

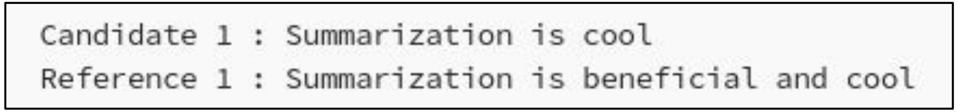
?

ROUGE-N

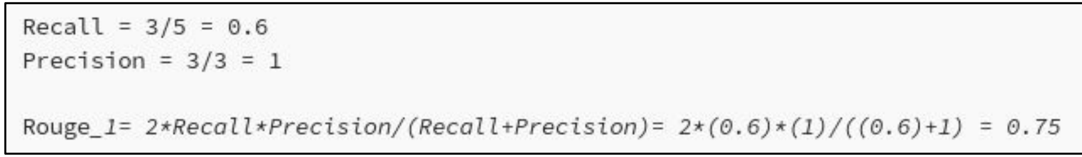
$$RECALL = \frac{Overlapping\, number\, of\, n{-}grams}{Number\, of\, n{-}grams\, in\, the\, reference}$$

$$PRECISION = \frac{Overlapping\, number\, of\, n{-}grams}{Number\, of\, n{-}grams\, in\, the\, candidate}$$

Recall and Precision Equations



```
Candidate 1 : Summarization is cool
Reference 1 : Summarization is beneficial and cool
```



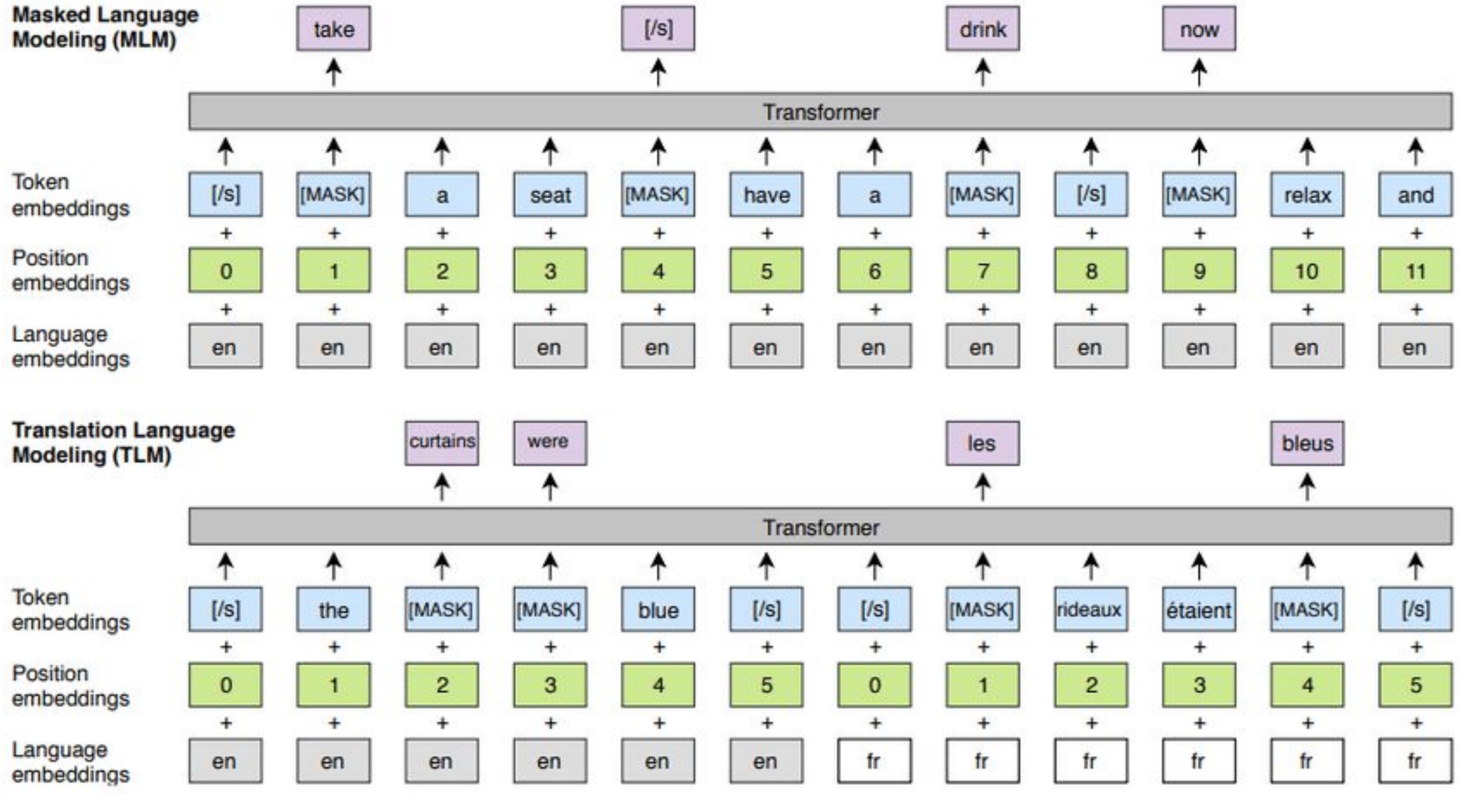
```
Recall = 3/5 = 0.6
Precision = 3/3 = 1

Rouge_1= 2*Recall*Precision/(Recall+Precision)= 2*(0.6)*(1)/((0.6)+1) = 0.75
```

To finalize calculation we also need to calculate F1 scores (Harmonic mean) :

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

Masked Language Modelling



Masked Language Modeling (MLM)

		take			[/s]			drink		now		
	Transformer											
Token embeddings	[/s]	[MASK]	a	seat	[MASK]	have	a	[MASK]	[/s]	[MASK]	relax	and
Position embeddings	0	1	2	3	4	5	6	7	8	9	10	11
Language embeddings	en	en	en	en	en	en	en	en	en	en	en	en

Translation Language Modeling (TLM)

			curtains	were				les			bleus	
	Transformer											
Token embeddings	[/s]	the	[MASK]	[MASK]	blue	[/s]	[/s]	[MASK]	rideaux	étaient	[MASK]	[/s]
Position embeddings	0	1	2	3	4	5	0	1	2	3	4	5
Language embeddings	en	en	en	en	en	en	fr	fr	fr	fr	fr	fr

Alignment Optimization Criteria

Linear Criterion

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - y_i\|^2$$

Orthogonal Criterion

$$\max_{W} \sum_{i} (Wx_i)^T y_i$$

Orthogonal Procrustes Criterion

$$\min_{W \in O_d} \|WX - Y\|^2 = UV^T$$

$$\text{Where } U\Sigma V^T = SVD(YX^T)$$

Alignment Optimization Criteria (cont.)

RCSLS Criterion

$$\underset{W\in O_d}{min}\frac{1}{n}\sum_{i=1}^{n}-2x_i^TW^Ty_i+\frac{1}{k}\sum_{y_j\in N_Y(Wx_i)}x_i^TW^Ty_j+\frac{1}{k}\sum_{Wx_j\in N_X(y_i)}x_j^TW^Ty_i$$

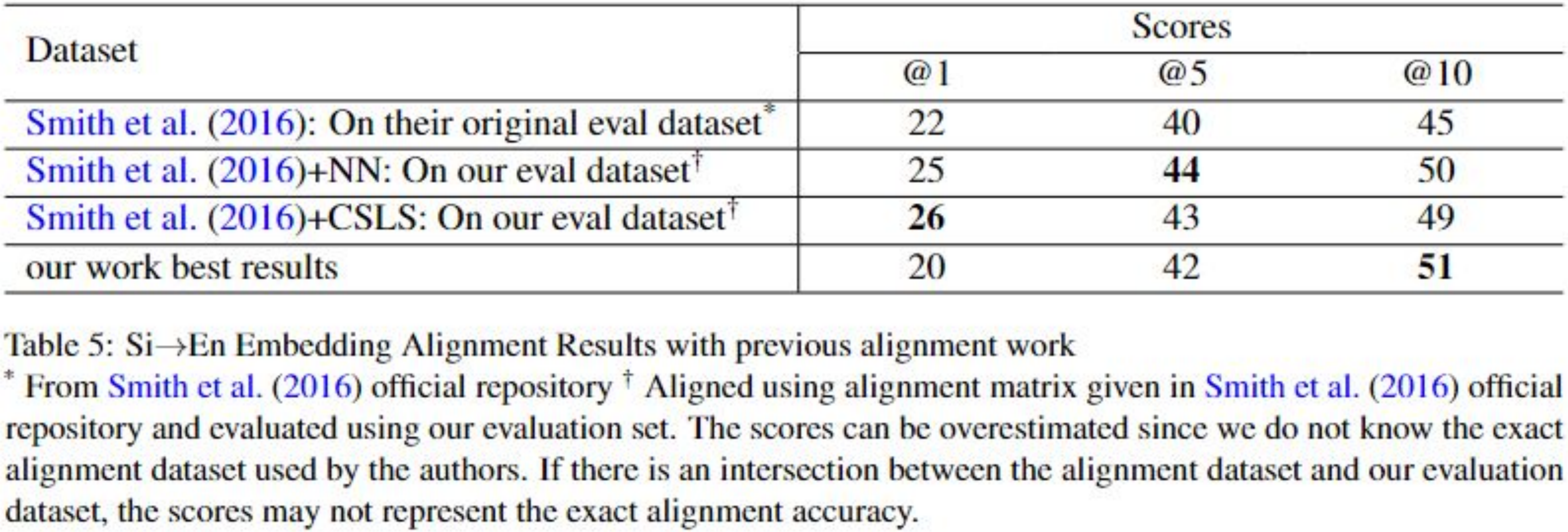
Contrastive Alignment (C1) - InfoNCE Loss

$$\mathcal{L}_{\mathrm{N}}=-\mathop{\mathbb{E}}_{X}\left[\log\frac{f_k(x_{t+k},c_t)}{\sum_{x_j\in X}f_k(x_j,c_t)}\right]$$

Contrastive Alignment (C2) - paper

$$(1-\lambda)\frac{\mathbf{v}_w\boldsymbol{W}}{\left\|\mathbf{v}_w\boldsymbol{W}\right\|_2}+\lambda\frac{f_{\theta'}(w)}{\left\|f_{\theta'}(w)\right\|_2},$$

Previous Work for Sinhala



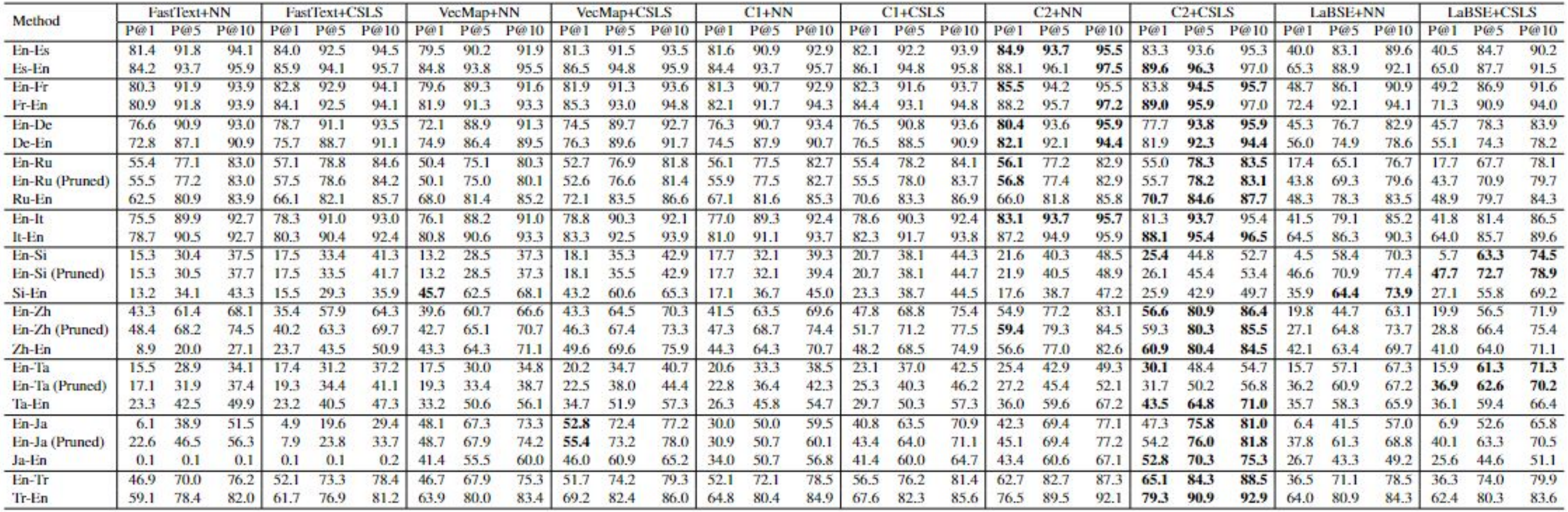
Dataset	Scores		
	@1	@5	@10
Smith et al. (2016): On their original eval dataset*	22	40	45
Smith et al. (2016)+NN: On our eval dataset†	25	**44**	50
Smith et al. (2016)+CSLS: On our eval dataset†	**26**	43	49
our work best results	20	42	**51**

Table 5: Si→En Embedding Alignment Results with previous alignment work

* From Smith et al. (2016) official repository † Aligned using alignment matrix given in Smith et al. (2016) official repository and evaluated using our evaluation set. The scores can be overestimated since we do not know the exact alignment dataset used by the authors. If there is an intersection between the alignment dataset and our evaluation dataset, the scores may not represent the exact alignment accuracy.

All Language Study

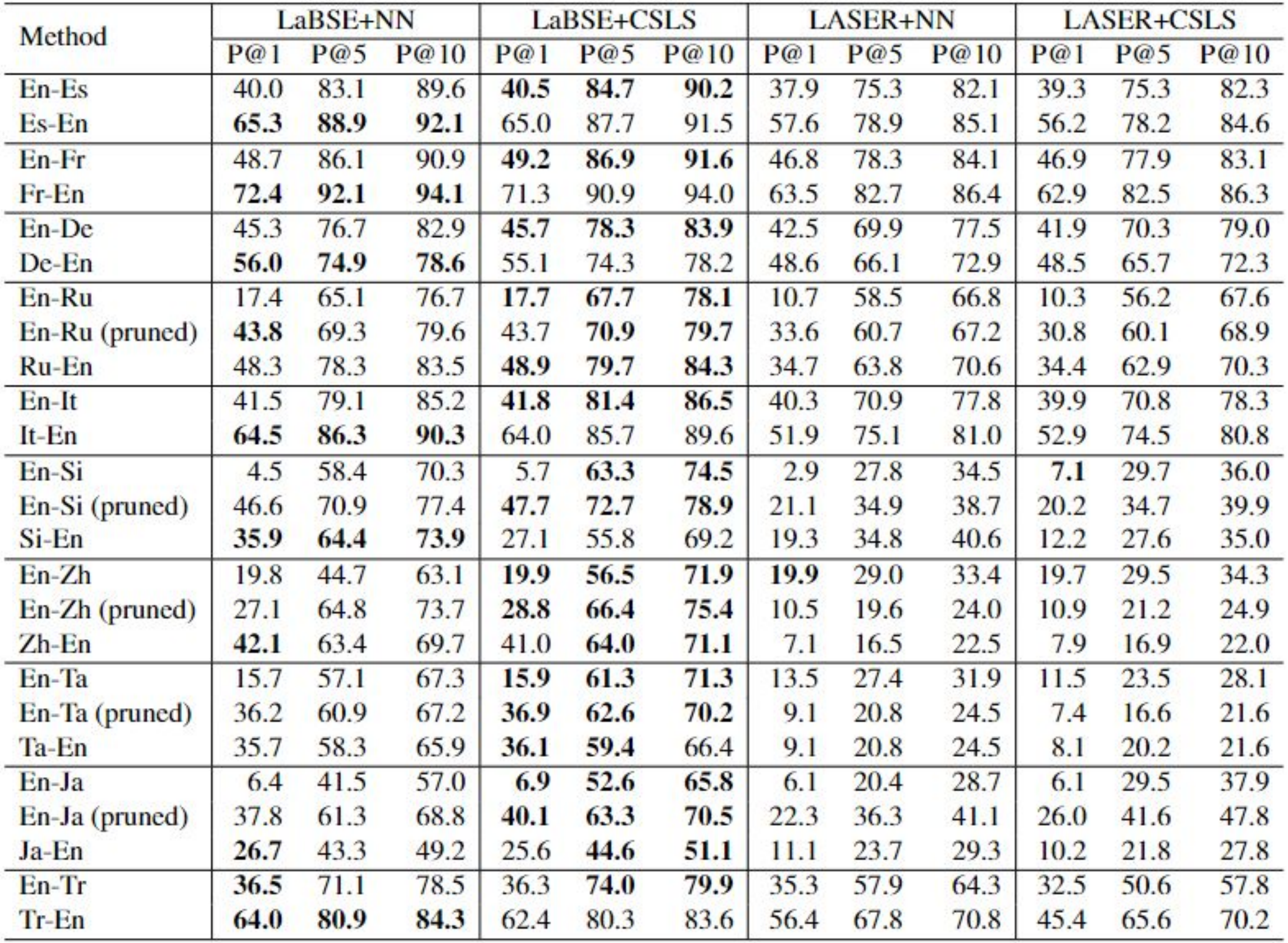
- This table has the extended results for all language pairs and all alignment techniques



Method	FastText+NN			FastText+CSLS			VecMap+NN			VecMap+CSLS			C1+NN			C1+CSLS			C2+NN			C2+CSLS			LaBSE+NN			LaBSE+CSLS		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
En-Es	81.4	91.8	94.1	84.0	92.5	94.5	79.5	90.2	91.9	81.3	91.5	93.5	81.6	90.9	92.9	82.1	92.2	93.9	**84.9**	**93.7**	**95.5**	83.3	93.6	95.3	40.0	83.1	89.6	40.5	84.7	90.2
Es-En	84.2	93.7	95.9	85.9	94.1	95.7	84.8	93.8	95.5	86.5	94.8	95.9	84.4	93.7	95.7	86.1	94.8	95.8	88.1	96.1	**97.5**	**89.6**	**96.3**	97.0	65.3	88.9	92.1	65.0	87.7	91.5
En-Fr	80.3	91.9	93.9	82.8	92.9	94.1	79.6	89.3	91.6	81.9	91.3	93.6	81.3	90.7	92.9	82.3	91.6	93.7	**85.5**	94.2	95.5	83.8	**94.5**	**95.7**	48.7	86.1	90.9	49.2	86.9	91.6
Fr-En	80.9	91.8	93.9	84.1	92.5	94.1	81.9	91.3	93.3	85.3	93.0	94.8	82.1	91.7	94.3	84.4	93.1	94.8	88.2	95.7	**97.2**	**89.0**	**95.9**	97.0	72.4	92.1	94.1	71.3	90.9	94.0
En-De	76.6	90.9	93.0	78.7	91.1	93.5	72.1	88.9	91.3	74.5	89.7	92.7	76.3	90.7	93.4	76.5	90.8	93.6	**80.4**	93.6	**95.9**	77.7	**93.8**	**95.9**	45.3	76.7	82.9	45.7	78.3	83.9
De-En	72.8	87.1	90.9	75.7	88.7	91.1	74.9	86.4	89.5	76.3	89.6	91.7	74.5	87.9	90.7	76.5	88.5	90.9	**82.1**	92.1	**94.4**	81.9	**92.3**	**94.4**	56.0	74.9	78.6	55.1	74.3	78.2
En-Ru	55.4	77.1	83.0	57.1	78.8	84.6	50.4	75.1	80.3	52.7	76.9	81.8	56.1	77.5	82.7	55.4	78.2	84.1	**56.1**	77.2	82.9	55.0	**78.3**	**83.5**	17.4	65.1	76.7	17.7	67.7	78.1
En-Ru (Pruned)	55.5	77.2	83.0	57.5	78.6	84.2	50.1	75.0	80.1	52.6	76.6	81.4	55.9	77.5	82.7	55.5	78.0	83.7	**56.8**	77.4	82.9	55.7	**78.2**	**83.1**	43.8	69.3	79.6	43.7	70.9	79.7
Ru-En	62.5	80.9	83.9	66.1	82.1	85.7	68.0	81.4	85.2	72.1	83.5	86.6	67.1	81.6	85.3	70.6	83.3	86.9	66.0	81.8	85.8	**70.7**	**84.6**	**87.7**	48.3	78.3	83.5	48.9	79.7	84.3
En-It	75.5	89.9	92.7	78.3	91.0	93.0	76.1	88.2	91.0	78.8	90.3	92.1	77.0	89.3	92.4	78.6	90.3	92.4	**83.1**	**93.7**	**95.7**	81.3	**93.7**	95.4	41.5	79.1	85.2	41.8	81.4	86.5
It-En	78.7	90.5	92.7	80.3	90.4	92.4	80.8	90.6	93.3	83.3	92.5	93.9	81.0	91.1	93.7	82.3	91.7	93.8	87.2	94.9	95.9	**88.1**	**95.4**	**96.5**	64.5	86.3	90.3	64.0	85.7	89.6
En-Si	15.3	30.4	37.5	17.5	33.4	41.3	13.2	28.5	37.3	18.1	35.3	42.9	17.7	32.1	39.3	20.7	38.1	44.3	21.6	40.3	48.5	**25.4**	44.8	52.7	4.5	58.4	70.3	5.7	**63.3**	**74.5**
En-Si (Pruned)	15.3	30.5	37.7	17.5	33.5	41.7	13.2	28.5	37.3	18.1	35.5	42.9	17.7	32.1	39.4	20.7	38.1	44.7	21.9	40.5	48.9	26.1	45.4	53.4	46.6	70.9	77.4	**47.7**	**72.7**	**78.9**
Si-En	13.2	34.1	43.3	15.5	29.3	35.9	**45.7**	62.5	68.1	43.2	60.6	65.3	17.1	36.7	45.0	23.3	38.7	44.5	17.6	38.7	47.2	25.9	42.9	49.7	35.9	**64.4**	**73.9**	27.1	55.8	69.2
En-Zh	43.3	61.4	68.1	35.4	57.9	64.3	39.6	60.7	66.6	43.3	64.5	70.3	41.5	63.5	69.6	47.8	68.8	75.4	54.9	77.2	83.1	**56.6**	**80.9**	**86.4**	19.8	44.7	63.1	19.9	56.5	71.9
En-Zh (Pruned)	48.4	68.2	74.5	40.2	63.3	69.7	42.7	65.1	70.7	46.3	67.4	73.3	47.3	68.7	74.4	51.7	71.2	77.5	**59.4**	79.3	84.5	59.3	**80.3**	**85.5**	27.1	64.8	73.7	28.8	66.4	75.4
Zh-En	8.9	20.0	27.1	23.7	43.5	50.9	43.3	64.3	71.1	49.6	69.6	75.9	44.3	64.3	70.7	48.2	68.5	74.9	56.6	77.0	82.6	**60.9**	**80.4**	**84.5**	42.1	63.4	69.7	41.0	64.0	71.1
En-Ta	15.5	28.9	34.1	17.4	31.2	37.2	17.5	30.0	34.8	20.2	34.7	40.7	20.6	33.3	38.5	23.1	37.0	42.5	25.4	42.9	49.3	**30.1**	48.4	54.7	15.7	57.1	67.3	15.9	**61.3**	**71.3**
En-Ta (Pruned)	17.1	31.9	37.4	19.3	34.4	41.1	19.3	33.4	38.7	22.5	38.0	44.4	22.8	36.4	42.3	25.3	40.3	46.2	27.2	45.4	52.1	31.7	50.2	56.8	36.2	60.9	67.2	**36.9**	**62.6**	**70.2**
Ta-En	23.3	42.5	49.9	23.2	40.5	47.3	33.2	50.6	56.1	34.7	51.9	57.3	26.3	45.8	54.7	29.7	50.3	57.3	36.0	59.6	67.2	**43.5**	**64.8**	**71.0**	35.7	58.3	65.9	36.1	59.4	66.4
En-Ja	6.1	38.9	51.5	4.9	19.6	29.4	48.1	67.3	73.3	**52.8**	72.4	77.2	30.0	50.0	59.5	40.8	63.5	70.9	42.3	69.4	77.1	47.3	**75.8**	**81.0**	6.4	41.5	57.0	6.9	52.6	65.8
En-Ja (Pruned)	22.6	46.5	56.3	7.9	23.8	33.7	48.7	67.9	74.2	**55.4**	73.2	78.0	30.9	50.7	60.1	43.4	64.0	71.1	45.1	69.4	77.2	54.2	**76.0**	**81.8**	37.8	61.3	68.8	40.1	63.3	70.5
Ja-En	0.1	0.1	0.1	0.1	0.1	0.2	41.4	55.5	60.0	46.0	60.9	65.2	34.0	50.7	56.8	41.4	60.0	64.7	43.4	60.6	67.1	**52.8**	**70.3**	**75.3**	26.7	43.3	49.2	25.6	44.6	51.1
En-Tr	46.9	70.0	76.2	52.1	73.3	78.4	46.7	67.9	75.3	51.7	74.2	79.3	52.1	72.1	78.5	56.5	76.2	81.4	62.7	82.7	87.3	**65.1**	**84.3**	**88.5**	36.5	71.1	78.5	36.3	74.0	79.9
Tr-En	59.1	78.4	82.0	61.7	76.9	81.2	63.9	80.0	83.4	69.2	82.4	86.0	64.8	80.4	84.9	67.6	82.3	85.6	76.5	89.5	92.1	**79.3**	**90.9**	**92.9**	64.0	80.9	84.3	62.4	80.3	83.6

LaBSE vs LASER

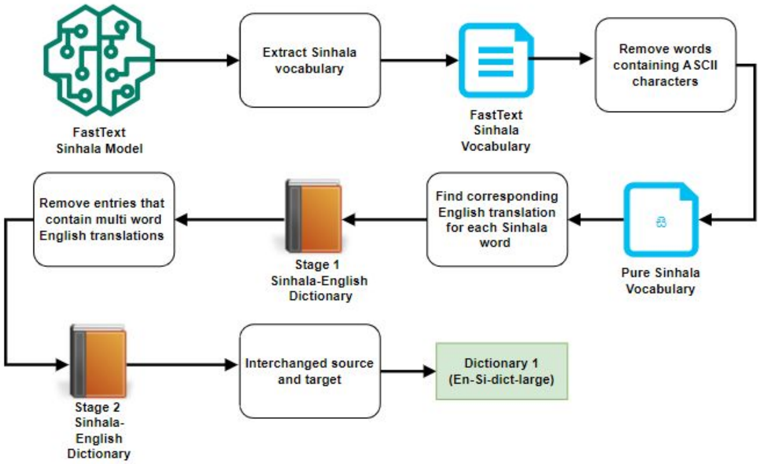
- Extended results comparison between LaBSE and LASER

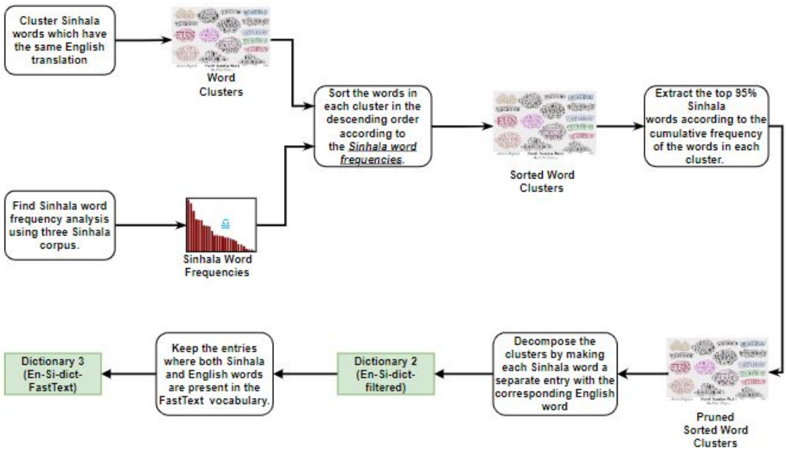


Method	LaBSE+NN			LaBSE+CSLS			LASER+NN			LASER+CSLS		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
En-Es	40.0	83.1	89.6	**40.5**	**84.7**	**90.2**	37.9	75.3	82.1	39.3	75.3	82.3
Es-En	**65.3**	**88.9**	**92.1**	65.0	87.7	91.5	57.6	78.9	85.1	56.2	78.2	84.6
En-Fr	48.7	86.1	90.9	**49.2**	**86.9**	**91.6**	46.8	78.3	84.1	46.9	77.9	83.1
Fr-En	**72.4**	**92.1**	**94.1**	71.3	90.9	94.0	63.5	82.7	86.4	62.9	82.5	86.3
En-De	45.3	76.7	82.9	**45.7**	**78.3**	**83.9**	42.5	69.9	77.5	41.9	70.3	79.0
De-En	**56.0**	**74.9**	**78.6**	55.1	74.3	78.2	48.6	66.1	72.9	48.5	65.7	72.3
En-Ru	17.4	65.1	76.7	**17.7**	**67.7**	**78.1**	10.7	58.5	66.8	10.3	56.2	67.6
En-Ru (pruned)	**43.8**	69.3	79.6	43.7	**70.9**	**79.7**	33.6	60.7	67.2	30.8	60.1	68.9
Ru-En	48.3	78.3	83.5	**48.9**	**79.7**	**84.3**	34.7	63.8	70.6	34.4	62.9	70.3
En-It	41.5	79.1	85.2	**41.8**	**81.4**	**86.5**	40.3	70.9	77.8	39.9	70.8	78.3
It-En	**64.5**	**86.3**	**90.3**	64.0	85.7	89.6	51.9	75.1	81.0	52.9	74.5	80.8
En-Si	4.5	58.4	70.3	5.7	**63.3**	**74.5**	2.9	27.8	34.5	**7.1**	29.7	36.0
En-Si (pruned)	46.6	70.9	77.4	**47.7**	**72.7**	**78.9**	21.1	34.9	38.7	20.2	34.7	39.9
Si-En	**35.9**	**64.4**	**73.9**	27.1	55.8	69.2	19.3	34.8	40.6	12.2	27.6	35.0
En-Zh	19.8	44.7	63.1	**19.9**	**56.5**	**71.9**	**19.9**	29.0	33.4	19.7	29.5	34.3
En-Zh (pruned)	27.1	64.8	73.7	**28.8**	**66.4**	**75.4**	10.5	19.6	24.0	10.9	21.2	24.9
Zh-En	**42.1**	63.4	69.7	41.0	**64.0**	**71.1**	7.1	16.5	22.5	7.9	16.9	22.0
En-Ta	15.7	57.1	67.3	**15.9**	**61.3**	**71.3**	13.5	27.4	31.9	11.5	23.5	28.1
En-Ta (pruned)	36.2	60.9	67.2	**36.9**	**62.6**	**70.2**	9.1	20.8	24.5	7.4	16.6	21.6
Ta-En	35.7	58.3	65.9	**36.1**	**59.4**	66.4	9.1	20.8	24.5	8.1	20.2	21.6
En-Ja	6.4	41.5	57.0	**6.9**	**52.6**	**65.8**	6.1	20.4	28.7	6.1	29.5	37.9
En-Ja (pruned)	37.8	61.3	68.8	**40.1**	**63.3**	**70.5**	22.3	36.3	41.1	26.0	41.6	47.8
Ja-En	**26.7**	43.3	49.2	25.6	**44.6**	**51.1**	11.1	23.7	29.3	10.2	21.8	27.8
En-Tr	**36.5**	71.1	78.5	36.3	**74.0**	**79.9**	35.3	57.9	64.3	32.5	50.6	57.8
Tr-En	**64.0**	**80.9**	**84.3**	62.4	80.3	83.6	56.4	67.8	70.8	45.4	65.6	70.2

Sinhala-English Parallel Dictionary Dataset

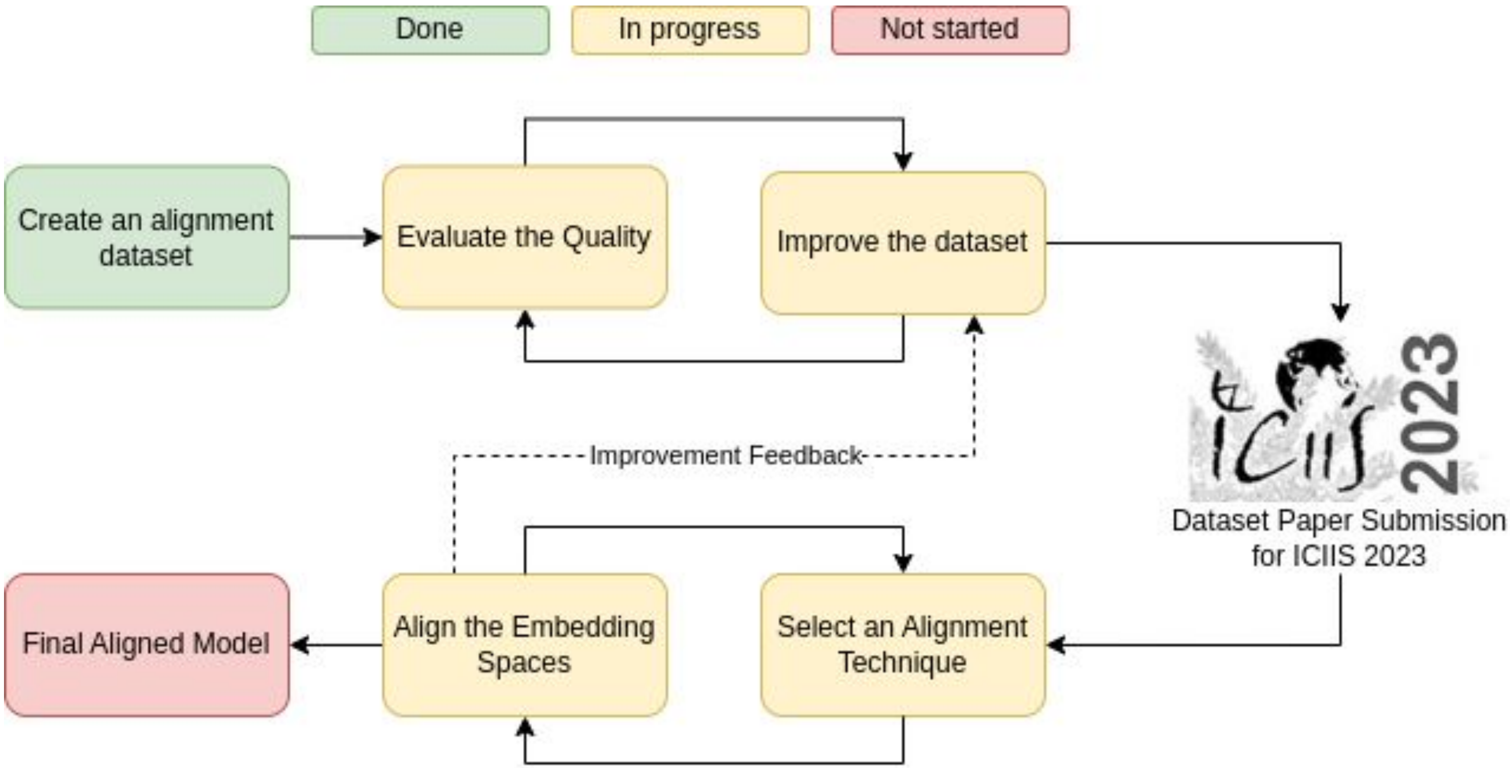
- We created 3 large-scale Sinhala-English parallel word datasets that facilitate word-level NLP tasks such as lexicon induction and supervised word embedding alignment
- We have created a data generation pipeline for the dataset creation process
- We used these datasets to create the alignment dataset for Sinhala-English word embedding alignment tasks we carried out **(animate properly to fit the figures to height)**


FastText
Sinhala Model
Extract Sinhala vocabulary
FastText
Sinhala
Vocabulary
Remove words containing ASCII characters
Pure Sinhala
Vocabulary
Find corresponding English translation for each Sinhala word
Stage 1
Sinhala-English
Dictionary
Remove entries that contain multi word English translations
Stage 2
Sinhala-
English
Dictionary
Interchanged source and target
Dictionary 1
(En-Si-dict-large)


Cluster Sinhala words which have the same English translation
Word
Clusters
Find Sinhala word frequency analysis using three Sinhala corpus.
Sinhala Word
Frequencies
Sort the words in each cluster in the descending order according to the Sinhala word frequencies.
Sorted Word
Clusters
Extract the top 95% Sinhala words according to the cumulative frequency of the words in each cluster.
Pruned
Sorted Word
Clusters
Decompose the clusters by making each Sinhala word a separate entry with the corresponding English word
Dictionary 2
(En-Si-dict-
filtered)
Keep the entries where both Sinhala and English words are present in the FastText vocabulary.
Dictionary 3
(En-Si-dict-
FastText)

[43] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, Aug. 2023

Progress



Done

In progress

Not started

Create an alignment dataset

Evaluate the Quality

Improve the dataset

Dataset Paper Submission for ICIIS 2023

Improvement Feedback

Final Aligned Model

Align the Embedding Spaces

Select an Alignment Technique

ICIIS Paper Submission

- Conference: ICIIS Conference 2023
- A dataset paper
- Presents three Sinhala-English parallel word datasets
- Auxiliary task of the main research - Creating an alignment dataset for supervised word embedding alignment
- Notification of outcome: 7th June 2023

Limitations