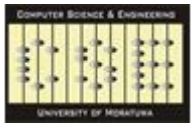
Sinhala-English Word Embedding Alignment:
Introducing Datasets and Benchmark for a Low Resource Language



Kasun Wickramasinghe and Nisansa de Silva

Department of Computer Science & Engineering, University of Moratuwa
Sri Lanka

Introduction

- Embeddings have become a primary ingredient in many flavours of Natural Language Processing (NLP) tasks.
- Multilingual embeddings share a common embedding space for many languages
- Due to the scarcity of parallel training data, low-resource languages such as Sinhala, still tend to focus more on monolingual embeddings instead of multilingual embeddings.
- Embedding alignment solves the problem of using monolingual embeddings for multilingual tasks.
- Even Though few previous research have been carried out for Sinhala word embedding alignment [1, 2], still we lack of a proper baseline research for that area.
- One major reason for less research for such areas is not having proper free and publicly available datasets for low-resource NLP tasks.

[1] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2016. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In International Conference on Learning Representations.
[2] Anushika Liyanage, Surangika Ranathunga, and Sanath Jayasena. 2021. Bilingual lexical induction for sinhala-english using cross lingual embedding spaces In 2021 Moratuwa Engineering Research Conference (MERCon), pages 579–584.

Our Contributions

- Align Sinhala and English word embedding spaces based on available alignment techniques.
- Introduce a benchmark for Sinhala language embedding alignment.
- Introduce Sinhala-English alignment datasets.
  - These datasets serve as the anchor datasets for supervised word embedding alignment.
- Simple statistical approach to create word-level alignment datasets using parallel corpora
- Make all our findings and resources open and publicly available for the community

https://github.com/kasunw22/sinhala-word-embedding-alignment

Alignment Dataset Creation

- Parallel aligned dataset is needed for the supervised word embedding alignment.
- For the Sinhala-English pair there is no MUSE[3]-like datasets available at the moment.
- We experimented two approaches to build a MUSE-like alignment datasets for the Sinhala-English pair.
    - Building a dataset from Si-En parallel corpora
    - Building a dataset using an available Si-En dictionary dataset

[3] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In ICLR.

Alignment Dataset Creation - approach 1

- This approach is building a word dictionary using available parallel corpora.
- Our assumption is,
  - "In a parallel corpus, the corresponding word translation pairs should co-occur."
  - In other words, "If two source and target language words co-occur more often, then there is a high chance for them to be a translation pair."
- When large enough parallel data points from parallel corpora are available, this measurement tends to be more accurate (statistical sampling)
- The optimization criterion is given in the next slide
- We performed this experiment to evaluate the feasibility of this new method and, not tried hard on creating a dataset using this method

Alignment Dataset Creation - approach 1 (cont.)

$$\max_{src,tgl} \left[P\left(src|tgl\right) P\left(tgl|src\right) \right]$$

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

$$\implies \max_{src,tgl} \left[\frac{count(src,tgl)^2}{count(src).count(tgl)} \right]$$

Where,

- `P(target|source)` - Finding the target word in the context of the source word (corresponding translation) given the source word
- `P(source|target)` - Finding the source word in the context of the target word (corresponding translation) given the target word

Alignment Dataset Creation - approach 2

- Take a subset of an available large-scale dictionary dataset [4] and form the alignment datasets
- We noticed that our first approach is a promising one and gave competitive results with the dataset created using the available dictionary dataset

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

[4] Kasun Wickramasinghe and Nisansa De Silva. 2023. Sinhala-english parallel word dictionary dataset. In 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), pages 61–66.

Sinhala-English Embedding Alignment

- We aligned the Sinhala and English Fasttext word embedding models using the available supervised alignment techniques
- Evaluated the alignment quality using the word translation precisions

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

Sinhala-English Embedding Alignment - (cont.)

- Here is a comparison of the top-1 word translation precision of different language pairs and our work.
- All the other pairs are high resource languages except Sinhala which is a low resource language.
- All the training sets are of 5000 unique source words and, test sets are of 1500 unique source words.

Method	Joulin et al. (2018a)										Ours	
	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-si	si-en
Adv.+refine	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	-	-
Wass. Proc.+refine	82.8	84.1	82.6	82.9	75.4	73.3	43.7	59.1	-	-	-	-
Procrustes	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7	36.7	20.4	18.0
Procrustes+ refine	82.4	83.9	82.3	83.2	75.3	73.2	50.1	63.5	40.3	35.5	20.9	**21.7**
RCSLS + spectral	83.5	85.7	82.3	**84.1**	78.2	75.8	56.1	66.5	44.9	45.7	21.5	19.2
RCSLS	**84.1**	**86.3**	**83.3**	**84.1**	**79.1**	**76.3**	**57.9**	**67.2**	**45.9**	**46.4**	**22.6**	19.4

Comparison with Previous Related Work

- Here is a comparison between our work and Smith et al.[1]s' work
- Si→En direction only and also provided the alignment matrix associated with the alignment.
- Smith et al.s' alignment datasets are not available to public and therefore this comparison may not reflect a genuine comparison

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**

[1] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2016. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In International Conference on Learning Representations.

Impact of the Alignment Dataset Size

- We experimented the impact of the alignment dataset size
- For all the previous comparisons we used 5000 unique words for the alignment dataset (that was the size used for other language pairs [5])

Dataset	Unique Src within 200k	Retrieval					
		NN			CSLS		
		@1	@5	@10	@1	@5	@10
En-Si-para-wiki-5k	5000	11.4	26.4	33.2	14.8	31.5	39.8
En-Si-para-wiki-full	27846	**17.0**	**36.1**	**45.1**	**20.2**	**42.4**	**50.9**
En-Si-para-cc-5k	5000	16.4	35.7	43.6	20.4	39.9	49.1
En-Si-para-cc-full	27856	**17.4**	**37.9**	**45.5**	**20.9**	**42.4**	**50.8**

[5] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018a. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Discussion and Future Work

- We have set a baseline for the Sinhala word embedding alignment with this paper.
- We have experimented only with the available supervised alignment techniques here.
- The alignment dataset directly affects the quality of the alignment.
- Therefor, we are willing to extend our research towards unsupervised and deep learning based techniques to further improve the alignment quality of the embeddings

Thank You!

Questions?

Sinhala Word Frequency Analysis

- We used the following there Sinhala corpora for the frequency analysis
  - Corpus by Upeksha et al. [12, 13] which was created using web crawling [link]
  - The second one is a corpus based on Jathaka Stories [link]
  - The third one is based on web crawled news articles [link]
- We selected these corpora to cover a diverse range of domains so that the domain bias is minimised
- Word counts of the three corpora:
  - Total words - 251,621,888 (251.6M)
  - Unique words - 2,168,118 (2M)

[12] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. De Silva, and G. Dias, “Implementing a corpus for sinhala language,” in Symposium on Language Technology for South Asia 2015, 2015.
[13] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. de Silva, and G. Dias, “Comparison between performance of various database systems for implementing a language corpus,” in International Conference: Beyond Databases, Architectures and Structures. Springer, 2015, pp. 82–91.

Available En-Si Parallel Datasets

Sentence/Document Level

- FLORES [14]
- NLLB [15]
- Opus Parallel Corpus
- Other [16, 17]

Word/Token Level

- Subasa [18] - ~36000 entries
- https://github.com/Isurie/Text-Classification-Module/blob/master/Dataset/en-sinhala%20dictionary.csv - (Text-Classification-Module) - 36429 entries
- https://github.com/gdgsl/sid/tree/master/assets/dictionary (Dictionary App) - 133960 entries (85532 single word entries)
- https://github.com/sinhalatypography/English-Sinhala-Dictionary
- https://github.com/laknath/Sinhala-Dictionary (Sinhala only - not a parallel dictionary)

[14] Japan, 2008, pp. 20–23. F. Guzm´an, 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,” in EMNLPIJCNLP, Nov. 2019.
[15] M. R. Costa-juss`a, J. Cross, O. C¸ elebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard et al., “No language left behind: Scaling human-centered machine translation,” arXiv preprint arXiv:2207.04672, 2022. [16] 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
[16] M. Ba˜n´on, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Espl`a-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn et al., “Paracrawl: Web-scale acquisition of parallel corpora,” in ACL, 2020, pp. 4555–4567.
[17]] 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.
[18] 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, pp. 20–23.

Stopword Removal

- English - Spacy English stop-words list
- Sinhala - Work by Lakmal et al. [19]

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

Zoom

Display name: 0072_Kasun_Wickramasinghe

Attendance: 1570906653_Kasun Wickramasinghe