Sinhala-English Parallel Word Dictionary Dataset

Kasun Wickramasinghe
Nisansa de Silva

Introduction

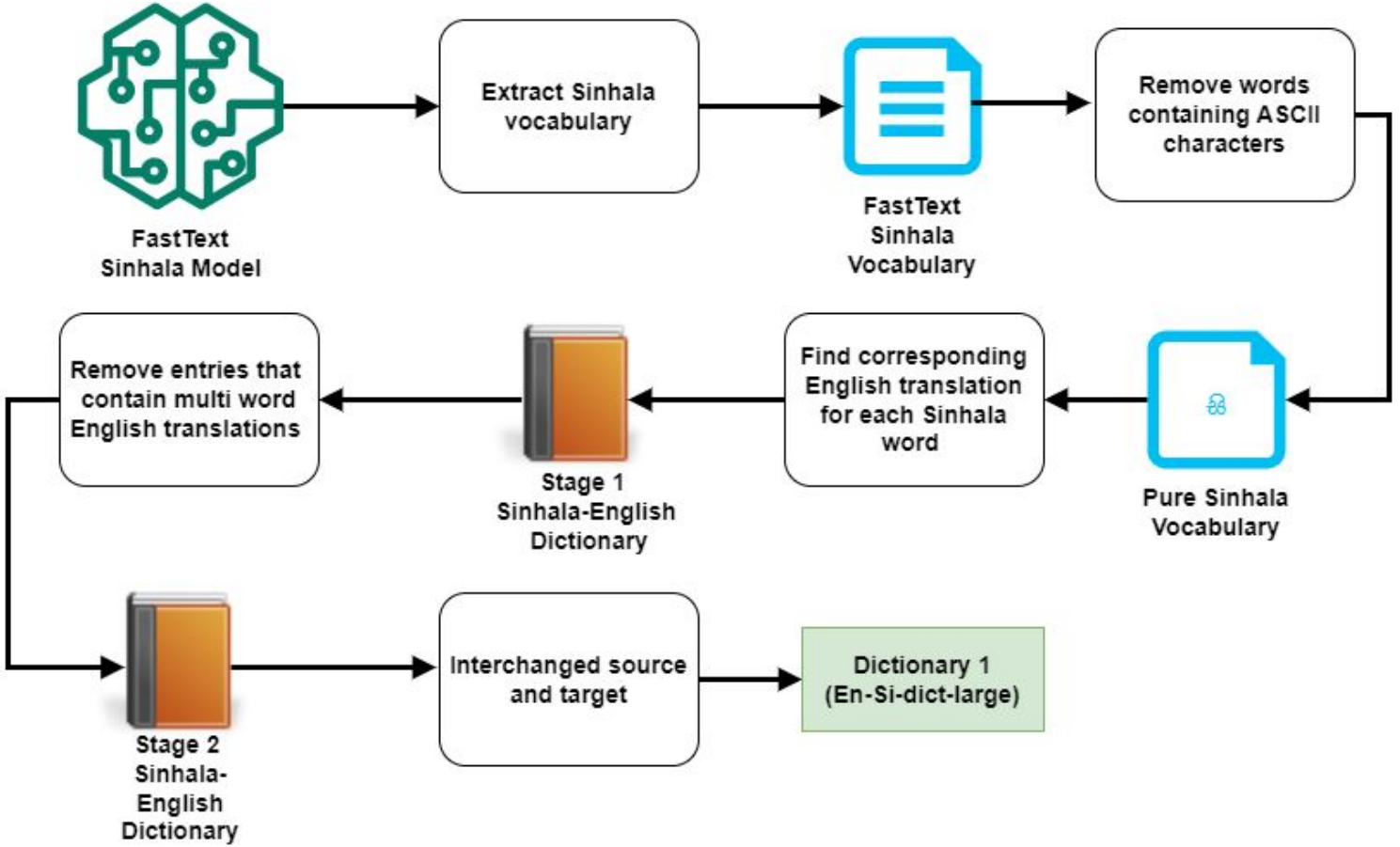
- Parallel datasets are vital for performing and evaluating any kind of multilingual task.
- Low resource languages like Sinhala are lacking from such datasets [1, 2, 3].
- Several free and publicly available parallel (sentence and paragraph) corpora are available for Sinhala [4].
- Large scale word/ token level parallel datasets (word dictionaries) are not found free and publicly available at the moment for Sinhala (few such dictionaries are there but they are very small in size. i.e. < 100k data-points).
- Such datasets are useful for bottom-up (word/ token level → sentence / paragraph level) multilingual NLP tasks such as,
    - Lexicon induction tasks [4]
    - Lexicon and synonym dictionary development [5]
    - Dictionary induction [6]
    - Word embedding alignment [7, 8, 9]

[1] A. Magueresse, V. Carles, and E. Heetderks, “Lowresource languages: A review of past work and future challenges,” arXiv preprint arXiv:2006.07264, 2020.
[2] S. Ranathunga and N. de Silva, “Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world,” arXiv preprint arXiv:2210.08523, 2022.
[3] N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,” arXiv preprint arXiv:1906.02358, 2019.
[4] X. Saralegi, I. Manterola, and I. San Vicente, “Analyzing methods for improving precision of pivot based bilingual dictionaries,” in EMNLP, 2011, pp. 846–856.
[5] A. H. Nasution, Y. Murakami, and T. Ishida, “Constraintbased bilingual lexicon induction for closely related languages,” in LREC, 2016, pp. 3291–3298.
[6] M. Wushouer, D. Lin, T. Ishida, and K. Hirayama, “A constraint approach to pivot-based bilingual dictionary induction,” TALLIP, vol. 15, no. 1, pp. 1–26, 2015.
[7] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” arXiv preprint arXiv:1309.4168, 2013.
[8] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in NAACL:HLT, 2015, pp. 1006–1011.
[9] A. Joulin, P. Bojanowski, T. Mikolov, H. Jegou, and ´ E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” arXiv preprint arXiv:1804.07745, 2018.

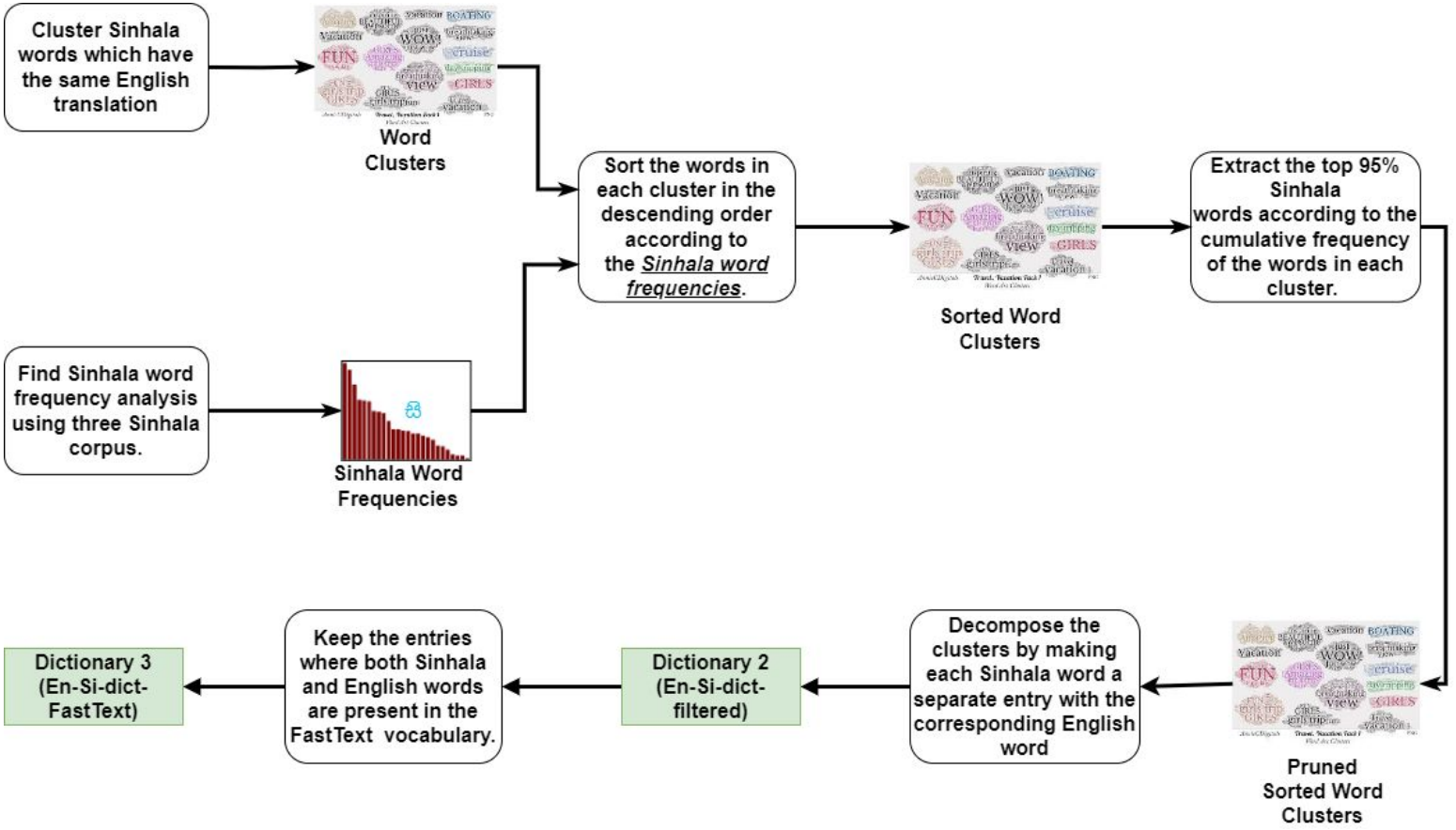
Our Contribution

- We introduce ***three*** English-Sinhala parallel dictionaries which will help in word and token-level Sinhala-English multilingual NLP tasks.
  - We have released two versions of the datasets
  - Those are ***word*** dictionaries (i.e. consists of word pairs)
- A pipeline to created datasets

Dataset Creation Pipeline - Dataset 1



Dataset Creation Pipeline - Dataset 2,3



Statistics

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

- Stop-word Removal:
  - **English:** Spacy stop-word list was used
  - **Sinhala:** stop-words list proposed by Lakmal et al. [10] was used
- Lookup-precision (P_L) is calculated using the following formula (the vocabulary is the FastText vocabulary)

$$P_L = Pr\left(\frac{word\ present\ in\ the\ vocabulary}{word\ present\ in\ the\ dictionary}\right)$$

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

Evaluation

- We have defined two scoring criteria (simple lookup score and nearest neighbour lookup score - NN-lookup) inspired by the widely used ROUGE-1 [11]
- The score is calculated using translation pairs of parallel corpora and our dataset

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

- 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 neighbours of each target word for Nearest Neighbor Lookup score) of above N_t source words.
- We evaluated our datasets using English-Sinhala versions of two parallel corpora *WikiMedia* (~7.9k aligned pairs) and *TED2020*(~4k aligned pairs).

[11] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81.

Example

En → Si direction

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

- everyone - {එක්කෙනා, එකිනෙක, එකිනෙකා සෑවෙන්}
- has - {තියෙනවා ඇත}
- Done - {කලා}
- something - {කිසිවක්, කිසිවෙක්, ටිකක්, මොකවත්, යමකිසි, යමක්}
- for - {අතරතුර, උදෙසා ගණ, දෙසට, නිසා පිණිස, වෙනුවෙන්, සඳහ}

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

Results

WikiMedia Evaluation

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**

TED-2020 Evaluation

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

Discussion and Future Work

- We have removed all the ASCII entries from the Sinhala vocabulary and therefore we do not have identical entries in both languages
- We have only single words, no any n-grams (n > 1)
- Above two are main reasons for having bit low evaluation scores
- Look for better matrices to evaluate a dictionary-type datasets
- Extend the dataset with more monolingual vocabularies following the same procedures
- Release another version of the dataset with n-grams (n > 1)

Thank You!

Quentions?