Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 61–64
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

cEnTam: Creation and Validation of a New English-Tamil Bilingual Corpus

Sanjanasri JP, Premjith B, Vijay Krishna Menon, Soman K P
Center for Computational Engineering and Networking (CEN), Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Coimbatore- 641112, India
{p_sanjanashree, b_premjith, m_vijaykrishna}@cb.amrita.edu, kp_soman@amrita.edu

Abstract

Natural Language Processing (NLP), is the field of artificial intelligence that gives the computer the ability to interpret, perceive and extract appropriate information from human languages. Contemporary NLP is predominantly a data-driven process. It employs machine learning and statistical algorithms to learn language structures from textual corpus. While applications of NLP in English, certain European languages such as Spanish, German, etc. have been tremendous, it is not so, in many Indian languages. There are obvious advantages in creating aligned bilingual and multilingual corpora. Machine translation, cross-lingual information retrieval, content availability and linguistic comparison are a few of the most sought after applications of such parallel corpora. This paper explains and validates a parallel corpus we created for English-Tamil bilingual pair.

1. Introduction

Accurately analyzing NLP tasks requires good quality corpus. However, creating such a corpus is a tedious and laborious task. There are only a few open-source bilingual corpora available for English-Tamil language pair. Existing corpora for English-Tamil language pair is listed in Table 1. EnTam (EnTam-v2) (Ramasamy et al., 2014) is an English-Tamil bilingual corpus crawled from the publicly available websites, especially form cinema, general news domain, and bible data. The author of this paper claimed that the corpus is plain raw data and requires some pre-processing before handling it for any NLP applications. Open subtitles (Lison and Tiedemann, 2016) is the corpus collected from the opus website. This corpus comprises bilingual movie subtitles that belong to the spoken language category. Tanzil (Tiedemann, 2012) is a collection of Quran translations compiled by the Tanzil project. OPUS website (Tiedemann, 2012) is a collection of English-Tamil bilingual localization files from open-source software projects like Ubuntu, KDE4, and GNOME. QED (QCRI Educational Domain) corpus (Abdelali et al., 2014) is again a data set belonging to the spoken language category. It includes bilingual subtitles of educational videos and lectures. The bilingual corpus is transcribed and translated using the AMARA web-based platform.
The following shortcomings were observed based on the information from these existing bilingual corpora:

- **Tanzil** is mostly translated poetry and **Bible** is non-contemporary prose. Hence, this cannot be utilised for generic NLP applications; specific dictionary has to be created.
- **EnTam** is a raw unstructured web corpus and contains a lot of noisy tokens such as image hyperlinks and other non-text web content. High-end pre-processing is required to make it usable. The sentences are aligned merely based on delimiter. The website data is crawled and is roughly comparable, which adversely effects bilingual embedding algorithms due to its high noise content.
- **Open subtitles** and **QED** are corpora belonging to spoken language style category, which might not help in efficient textual analysis.
- **Tatoeba** corpus has a minimal number of parallel sentences. Hence, it could not be used as standalone data for training machine learning models.

Although these existing corpora for English-Tamil language pair may still be useful in certain bilingual applications, we believe that these corpora still lack features that are strongly desirable for their use in word embedding context. Therefore, for justifiable analysis of semantic relatedness between language pairs using word embedding, a standard corpus has to be developed.

2. Data

Years back, creating bilingual corpus was an uphill task in NLP especially for Indian languages. Internet breaks the language barrier for both content and access today. Many literary works such as novels, short stories, plays, etc. are being translated among various languages and are made easily accessible mostly through crowd-sourcing. Having rich literature in a language doesn't imply that it is resource rich, at least in a bilingual context; creating parallel corpus is still a mammoth effort. The data provided is a collection of sentences taken from textbooks, bilingual novels, story books and bilingual websites that includes tourism, health and news domain. The source data are merely comparable. The sample data is shown in Table 2.

3. Experimental design

The methodology for acquisition of parallel corpus (cEnTam) from printed books and websites is shown in Fig. 1 and 2. In the pre-processing phase, the scanned images are cropped, skewed, rotated and even re-scanned wherever necessary to remove noise. The cleaned image is converted to text using Google OCR API. The text is further cleansed manually. It was necessary to ensure that the lines do not get blended with each other or that the font interferes with character recognition. The characters were at times not detected properly, which had to be typed manually.

Table 1: Details of existing corpora for English-Tamil language pair

Source	Domain	Sentences	English Tokens	Tamil Tokens
EnTam	Generic (bible, cinema, news)	169.8k	3.9M	2.7M
Open subtitles	Movie Subtitles	32.4k	0.2M	0.2M
OPUS website	Ubuntu,KDE4, GNOME	111.1k	3.2M	1.0M
Tateoba	Simple Sentences	0.3k	2.1k	1.6k
Tanzil	Quran Data	93.5k	2.8M	7.0M
QED	Subtitles of Educational Videos	0.7k	1.0M	0.5M

Table 2: Sample data for cEnTam

English	Tamil
kerala express connects daily to delhi	*thinamum kaeraLa viraivu rayil thilliyOtu in-NaikkiRathu*
i was at the cinema yesterday	*Naan NaeRRu thirai aranGkaththil iruNthaen*
thambidurai unanimously elected to lok sabha deputy speaker	*makkaLavai thunNai chapaaNaayakaraaka athimukavil thampithurai orumanathaaka thaervu cheyyappattaar*
this medicine will protect children from fever	*iNtha maruNthu kuzhaNthaikaLai kaay-chchalil iruNthu kaakkum*)

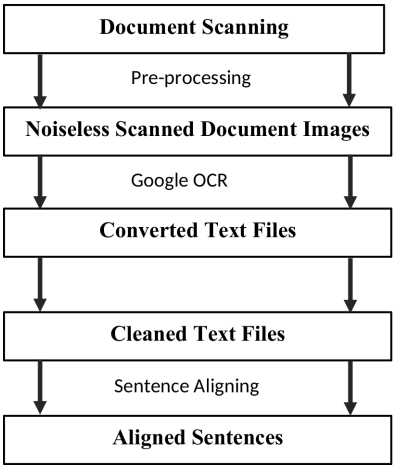

Figure 1: Block diagram for creation of parallel corpus (cEnTam) - printed books

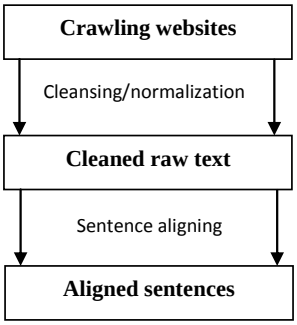

Figure 2: Block diagram for creation of parallel corpus (cEnTam) - website data

Table 3: Details of cEnTam Corpus.

Corpus Type	English (#. of sentences)	Tamil (#. of sentences)
Monolingual	457396	563568
Bilingual	56495	56495

In case of website data, the selective bilingual/monolingual websites are crawled using python library *"Scrapy"* to extract the main text from the web pages. Headline, hyperlinks, images, name(s) of author(s), publication date are all ignored. The extracted raw text is cleansed and normalized to remove punctuation, quotations, brackets, currency chars and digits. Since bilingual websites are already parallel, the sentences are aligned based on delimiter. Aligned sentences are checked manually for corrections. Lengthy sentences are split into shorter ones, to maintain consistency in data. The shorter sentence (less than six tokens/sentence) are less likely to contain any of the linguistic rule patterns, hence, the sentences vary from six to thirty tokens in length, with a corpus average of fifteen tokens per sentence including functional words. . Please find the specifics about the corpus in Table 3.

4. Comparative Analysis of corpora

The bilingual corpora are assessed based on *coherence*. In a coherent text, there are logical links between the words, sentences, and paragraphs of the text. Coherence can

be quantified by measuring similarity between sentences and/or documents. We use simple cosine similarity measure using appropriate embeddings, called the neighbourhood method. This approach assesses the translation quality of words using the bilingual embeddings trained on the aforementioned corpora. It measures the accuracy of the translation for the given source word. The evaluation is based on a test dictionary (AI, 2020).

For computing coherence between the sentences, we need to use pre-trained monolingual embeddings in English and Tamil separately from each corpora (Table 1). Using MUSE (Conneau et al., 2017), we can generate bilingual embeddings of all the pairs of words in the vocabulary, in an unsupervised manner. We then use these bilingual word embeddings to generate bilingual sentence embeddings. This embeds sentences of source and target language in a shared vector space. Average cosine similarity of the sentences is used as an accuracy metric.

5. Neural Machine Translation

This section discusses the comparative study of various corpora, using Neural Machine Translation (NMT) using the corpus created in-house (cEnTam) and EnTam. The process of translating lots of sentences is very complex and we chose to do it only on two main data sets. The quality of translation is directly assessed using a BLEU and RIBES scoring.A simple NMT architecture is used, to keep the training easy and fast which is shown in Fig. 3. The induced translation is evaluated based on both Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) metric. However, BLEU is known to be a standard metric for Machine Translation (MT) evaluation, RIBES is best suited for distant pair languages like English and Tamil (Tan et al., 2015). The accuracy can be improved further when used with attention mechanism (Bahdanau et al., 2014). This evaluation can demonstrate the better coherence of our Corpus.

6. Results

Efficacy of the bilingual embeddings trained over the various corpora are assessed using word level and sentence level neighbourhood. This method is inspired from (Mikolov et al., 2013). In this approach, we test whether the bilingual embedding is able to generate an appropriate target word for the given source word within the confining window of top similar words. Table 4 shows the performance of Nearest Neighbourhood word tasks.

The percentage accuracy of how likely the target words appear as nearest neighbour to the source word within K (words) window size, is measured. We see the value for K=1 itself is very high for our corpus compared to other corpora. This proves that the parallel sentences in our corpus are more coherent compared to others. Table 5 shows the performance of sentence similarity task on various corpora. Considering the performance of the all other corpora in the aforementioned tasks, cEnTam shows considerably better results; EnTam shows the next best results. Henceforth, for comparative study using NMT, cEnTam and EnTam corpora were used. The results are shown in

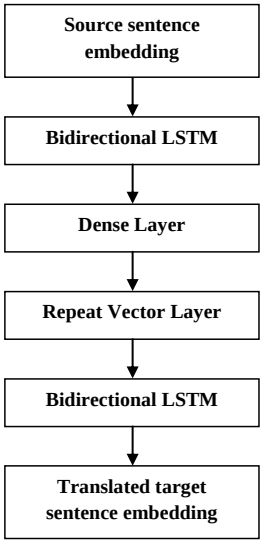

Figure 3: Neural Machine Translation Deep network used for testing corpora performances.

Table 4: Accuracy of the Nearest Neighbour analysis of word translation task using various window sizes in different corpora. The value represents the relative frequency of finding the target translation for a source word amongst the paired sentences expressed

Corpora	**Window size (Number of target words / 100 source words)**		
	K=1	**K=5**	**K=10**
EnTam	11.83	18.58	21.7
Open subtitles	11.61	18.37	20.53
OPUS website	4.91	7.06	7.8
Tanzil	0.47	0.95	1.05
QED	0.06	0.13	0.15
cEnTam	***27.08***	***35.15***	***39.36***

Table 6. Both the BLEU and RIBES metric yield better scores over translations created using cEnTam corpus over EnTam. This further proves the quality of cEnTam over EnTam in a real machine translation system.

Table 5: Average cosine sentence similarity of various corpora. A highest average and a lower deviation of cosine relations between sentence indicate coherence of the corpus.

Corpora	Avg. Cosine Similarity	Std.Dev
EnTam	0.12	0.09
Open subtitles	0.06	0.07
OPUS website	0.07	0.10
Tanzil	0.03	0.13
QED	0.04	0.21
cEnTam	***0.32***	***0.04***

Table 6: Results of Neural Machine Translation system performance with EnTam and cEnTam corpora

Corpora	BLEU	RIBES
EnTam	0.12	0.52
cEnTam	***0.39***	***0.74***

7. Conclusion

Non-existence of standard bilingual corpora is a major obstruction in effectively utilizing NLP technologies in many languages. Whether it is explainable (AI) analysis of semantic relatedness between language pairs or end-to-end deep learning models, it is necessary to have a standard bilingual corpus. Here, we have effectively demonstrated and implemented a methodology to create bilingual corpora, those are comparatively fast and requires less human effort. The corpus created is sentence aligned, hence it can be used for implementing NLP applications such as machine translation, cross-lingual information retrieval, semantic comparison and bilingual dictionary induction. The validations using nearest neighbourhood approach, sentence similarity and neural machine translation.

Acknowledgement

We would like to thank Dr.Rajendran S, Professor & Head (Retd.), Department of Linguistics, Tamil University, Thanjavur, India currently serving as adjunct professor in Centre for Computational Engineering and Networking (CEN), Amrita University, India, Dr. A.G. Menon, associate professor (Retd.), Department of Indian Studies and Department of Comparative Linguistics, Leiden University and Dr. Loganathan Ramaswamy, Machine Learning Engineer at MSD and main author of EnTam corpus, Prague, Czech Republic for their immense suggestion on creating bilingual corpus.

Bibliographical References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

AI, F. (2016). Pretrained vectors fasttext. https://fasttext.cc/docs/en/pretrained-vectors.html.

AI, F. (2020). Tamil-english dictionary. https://dl.fbaipublicfiles.com/arrival/dictionaries/ta-en.txt, April.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 748–756.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA, October. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ramasamy, L., Bojar, O., and Žabokrtský, Z. (2014). EnTam: An english-tamil parallel corpus (EnTam v2.0). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tan, L. L., Dehdari, J., and van Genabith, J. (2015). An awkward disparity between bleu / ribes scores and human judgements in machine translation. In *Proceedings of the Workshop on Asian Translation (WAT-2015)*, pages 74–81. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).