Mining Bilingual Word Pairs from Comparable Corpus using Apache Spark Framework

Sanjanasri JP[1], Vijay Krishna Menon[2], Soman KP[1], and Krzysztof Wolk[3]

[1]Center for Computational Engineering and Networking, Amrita School of Engineering,
Coimbatore - 641112, India
[2]Gadgeon Systems Private Limited, Kochi, Kerala, India
[3] Department of Multimedia, Polish-Japanese Institute of Information Technology,
Warsaw 02-008, Poland
jp_sanjanasri@cb.amrita.edu,vijay.km@gadgeon.com,kwolk@pja.edu.pl

Abstract

Bilingual dictionaries are essential resources in many areas of natural language processing tasks, but resource-scarce and less popular language pairs rarely have such. Efficient automatic methods for inducting bilingual dictionaries are needed as manual resources and efforts are scarce for low-resourced languages. In this paper, we induce word translations using bilingual embedding. We use the Apache Spark® framework for parallel computation. Further, to validate the quality of the generated bilingual dictionary, we use it in a phrase-table aided Neural Machine Translation (NMT) system. The system can perform moderately well with a manual bilingual dictionary; we change this into our inducted dictionary. The corresponding translated outputs are compared using the Bilingual Evaluation Understudy (BLEU) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) metrics.

1 Introduction

Digitised bilingual dictionaries primarily exist for resource-rich language pairs, such as English-German, English- Chinese, English-Hindi, etc. (Lardilleux et al., 2010). Such dictionaries are helpful for many natural language processing (NLP) tasks such as Machine Translation (MT) for translating Out-Of-Vocabulary (OOV) words, cross-lingual information retrieval, cross-lingual word embedding and multilingual parts-of-speech tagging (Wołk, 2019; Ye et al., 2016; Sharma and Mittal, 2018). Creating a bilingual dictionary requires high-quality parallel corpora and expert linguists, both of which are scarce and costly in resource-poor languages (Hajnicz et al., 2016; Sarma, 2019).

Previous works focus on methods that were based on pivot languages (Tanaka and Umemura, 1994; István and Shoichi, 2009; Wushouer et al., 2015), aligning words (Daille and Morin, 2008; Tufiş and maria Barbu, 2002) or using dependency relations (Yu and Tsujii, 2009). The pivot-based dictionary induction is a contemporary method that uses only dictionaries to and from a pivot language (intermediate language) to generate a new dictionary. This method is not very effective for highly ambiguous languages as it yields highly noisy dictionaries because lexicons of a language do not exhibit transitive relationship (Wushouer et al., 2014). Word alignment systems identify the translation equivalence of lexical units between two sentences that are sentence aligned (Choueka et al., 2000; Och and Ney, 2003). Depending on the purpose, the system may focus on the specific lexical units, e.g. a single word or collocation (Tiedemann, 2004; Schreiner et al., 2011; Chen et al., 2009). The dependency relation method is based on the premise that related words in different languages have a similar dependency relationship. These methods require either excellent linguistic knowledge or linguistic resource. The research line has robust outcomes on bilingual lexicon induction with the evolution of word embedding either by independently aligning trained word embedding in two languages or using the bilingual embedding to induce word translation pairs through nearest-neighbour or similar retrieval methods. In the BDI task, given a list of ‘n’ source language words $w_{s_1}, w_{s_2}, ...w_{s_n}$, the goal is to determine the most appropriate translationw_{t_i}, for each query word w_{s_i}. Finding a target language word embedding wv_{t_i} is accomplished by computing the nearest neighbour to the source word embedding wv_{s_i} in the shared semantic space, where cosine similarity is a measure between the embedding (Artetxe et al., 2019). However, this creates a phenomenon called hubness. In high-dimensional spaces, some data points, called hubs, are extraordinarily close to many other data points (Huang et al., 2019); this results in inappropriate/noisy translation.

In this paper, a simple cartesian product of the bilingual/cross-lingual word embedding is used

Proceedings of 14th Workshop on Building and Using Comparable Corpora in conjunction with RANLP-2021, pages 2–7, held online, 6 September 2021.
https://doi.org/10.26615/978-954-452-076-2_002

and filters the product outcome based on some linguistic regularities and thresholds. The generated (inducted) bilingual dictionary is used as a separate phrase-table in an NMT system. The system produces translations for every word in the text; the translations are validated for quality using the Bilingual Evaluation Understudy (BLEU) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) metric.

2 Bilingual and Cross-lingual Word embedding

In this paper, the terms 'bilingual' and 'cross-lingual' for word embedding is used with varying notions. The bilingual embedding maps the source and target language embedding in the shared semantic space. In contrast, the cross-lingual embedding learns a transfer function to translate the embedding from the source language semantic space to target language space; this preserves the more actual semantics pertained to that language (Mikolov et al., 2013). Visualisation of the embeddings is shown in Figure 1 and Figure 2.

BilBOWA toolkit (Gouws et al., 2015) is used to generate bilingual word embedding. The embedding of source and target language are trained jointly so that related words of two languages are closer to each other in the shared space. Therefore, the translational equivalence has higher cosine similarity. The model is trained with minimal parallel corpus and large monolingual corpora. However, the cross-lingual embedding is learned with a very bare minimal resource as small as 5000 source-target word pairs. Global neighbourhood is estimated as cross-lingual entropy. The main advantage of this method over bilingual embedding is that it is possible to generate embedding in the target language semantic space instead of shared space. In shared semantic space, the most semantic information pertained to the language is lost and likely to infer word vectors for related languages.

3 Implementation

The embedding size of the English word list is $\in \mathbb{R}^{8994 \times 300}$ and Tamil is $\in \mathbb{R}^{10097 \times 300}$. Tamil has more number words compared to English because of the inflected forms. The dimension of the Cartesian product of the word pair list (English and Tamil) is 90812418×300; this takes months for a typical computer system to compute. This complex computation is deployed to the cluster using Apache Spark® Framework (Zaharia et al., 2016). The word pairs are filtered in two folds, cosine similarity and lemmatization (Kengatharaiyer et al., 2019), where the root word is extracted from the surface forms. In the case of cross-lingual embedding, cross-lingual entropy is used instead of the cosine similarity measure. Figure 3 shows the architecture.

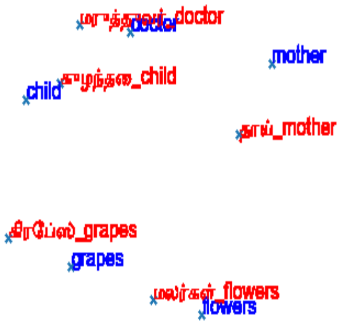


Figure 1: Visualization of Bilingual Embedding using T-SNE plot

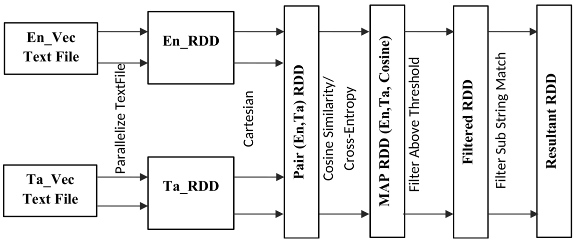
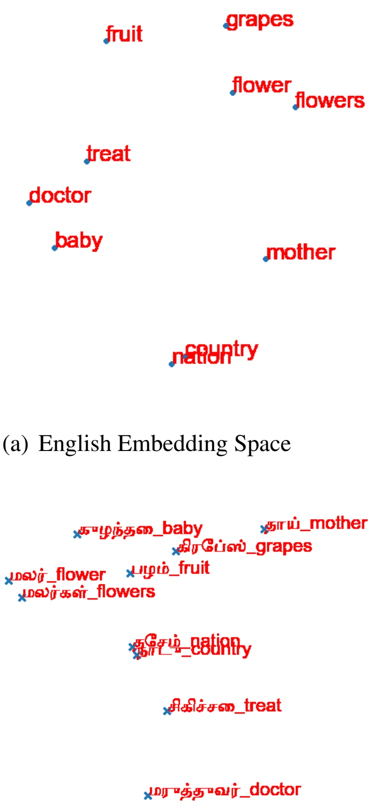


Figure 3: Apache Spark Implementation for Bilingual Dictionary Induction

The word embedding of Source and Target Language is mapped to a key-value pair Resilient Distributed Datasets (RDDs), a fundamental data structure of Spark; the word being a key and 300-dimensional representation as values. The Cartesian Product of two RDDs (En_RDD and Ta_RDD) generates the Pair RDD. On the Pair RDD, cosine similarity or cross-lingual entropy is applied to filter top similar words. Filtered RDD is further refined using a lemmatizer to avoid the inflected terms. The resultant RDD is saved as text file; this has the most similar source and target word, a bilingual dictionary.

The OpenNMT framework (Klein et al., 2017) is used for training an NMT system with the training parameter as shown in Table 1. The inducted lexi-



(a) English Embedding Space

(b) Tamil Embedding Space

Figure 2: Visualization of Cross-lingual embedding using T-SNE plot

cons are used as a phrase-table in NMT for translating Out-Of-Vocabulary (OOV) words. Training is done on Google Colab with GPU at backend.

Table 1: Training Parameters for English-Tamil OpenNMT Framework

Hyper Parameters	Values
Layers	3
Rnn_size	512
Embedding_size	512
Encoder/Decoder Type	Transformers
Train_steps / Validation_steps	3000/ 5000
Positional Encoding	True
Heads	8
Dropout	0.3
Learning rate	3
Batch size	4096
Optimiser	ADAM

4 Corpora Description

For the training language model, the monolingual Tamil corpus from the cEnTam dataset (P. et al., 2020) is used. Likewise, for training the machine translation systems, the English-Tamil parallel corpus from the cEnTam dataset is utilised. The specifics of the cEnTam corpus used is reported in Table 2.

Table 2: Specification of cEnTam Corpus

Corpus Type	English	Tamil
	(No. of sentences)	*(No. of sentences)*
Monolingual	589856	563568
Parallel	56495	56495

5 Results and Discussion

Table 3 and 4 show a sample of bilingually similar words above the cosine distance of 0.90 and 0.95. The correct translations are given in bold letters in Table 3 and 4. It can be inferred that the much more words that are not semantically similar (translational equivalent) but related crowds the search space, which might result in noisy word inductions (into the dictionary) and ambiguity. Hence the search space was shrunk above the cosine distance of 0.98 as shown in Table 5. It is observed that the inflected forms (surface forms) are closer than the related words in the embedding space to the query word. Unlike English, Tamil has no prepositions. Instead, it has case inflected nouns, for example, the translation of the prepositional phrase “in minutes” in English is equivalent to “*NimidanGkaLil*”, a case inflected noun(*NimidanGkaL + il* = minutes + in) in Tamil. Likewise, various sandhi inflected form of the noun “*kuzhanthai*” are *kuzhanthaip*, *kuzhanthaith*, etc. The chances of getting associated or related words in such a small space is negligible. The inflections are removed, and the root forms are inducted at the second stage of filtering, lemmatizer. The inducted dictionary is added as a lookup table in the NMT system.

Table 3: Sample output of bilingual words extracted above cosine similarity (threshold) 0.90

English	Tamil	Cosine Similarity
go	avaL	0.92
go	ennai	0.90
go	evvaLavu	0.90
go	anGkae	0.92
go	poaka	0.92
go	enGkae	0.90
go	un	0.90
good	**chariyaana**	**0.92**
good	aen	0.91
good	avaL	0.90
good	**nanRaaka**	**0.94**
good	evvaLavu	0.91

Table 4: Sample output of bilingual words extracted above cosine similarity 0.95

English	Tamil	Cosine Similarity
forests	pachumaiyaana	0.92
forests	adarNtha	0.95
forests	**kaadukaL**	**0.98**
flowers	malar	0.95
flowers	**malarkaL**	**0.97**
flowers	pookkaL	0.96

Table 5: Sample output of bilingual words extracted above cosine similarity 0.98. The exact translation of the query word is annotated with double raised asterisk ** and their inflected forms are annotated with single raised asterisk*.

English	Tamil	Cosine Similarity
minutes	NimidanGkaL **	0.98
minutes	NimidanGkaLil*	0.99
minutes	Nimidaththil *	0.97
minutes	NimidanGkaLaaka*	0.98

The accuracy of the translated sentence of the NMT system before and after appending the dictionary as a phrase table is shown in Table 6. The induced translation is evaluated based on both the Bilingual Evaluation Understudy (BLEU) (Koehn, 2010) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010) metrics. BLEU is the oldest and most adopted metrics to evaluate Mt system. It rewards systems for n-grams that have exact matches in the reference system. The longer n-gram scores account for the fluency of the translation in BLEU metric. In contrast, RIBES is sensitive towards word reordering, works well for language pairs having very different grammar and word order. It uses rank correlation coefficients based on word order to compare hypothesis and reference translations.

Table 6: Precision of NMT system

NMT System	BLEU	RIBES
Reference-Baseline	0.31	0.61
Reference-ManDic	0.33	0.66
Reference-InDic	0.34	0.71
ManDic-InDic	0.89	0.95

Although BLEU is a standard metric for the evaluation of MT system, RIBES is better suited for distant language pairs like English and Tamil (Callison-Burch et al., 2006). Hence, both measures are used for validating the NMT system developed. In the Table 6, the score is computed by comparing the reference translations with the translations of the NMT system after appending the manual and inducted dictionary (ManDic & IndDic). The ManDic and InDic systems are compared to showcase that the hypothesis translation of InDic is highly correlated with ManDic, though InDic has comparatively better score than ManDic when validated against Reference translation.

6 Conclusion and Summary

In this paper, we generated an English-Tamil bilingual dictionary using both bilingual (vectors in the same space) and cross-lingual (vectors in separate space, mapped) word embedding. In order to validate this induced dictionary, we have employed a table driven Neural Machine Translation (NMT) system. The goal was to measure the quality of the translated output (Tamil as the target language) when the original manual dictionary (ManDic) is replaced with the induced dictionary (InDic). The Baseline NMT system was trained on English-Tamil parallel corpus with over 56000 entries. A testset with 700 aligned sentences was used for validation. The translation quality is measured over the reference translations which are available (aligned Tamil sentences). Eventually, we will have three categories of translated output, namely, Baseline, ManDic and InDic. We compare each of them with the reference translation using the RIBES and BLEU metric (Isozaki et al., 2010; Koehn, 2010) to ascertain their quality. It is important to note that the quality of the translations is not of our interest but the change in performance when using different dictionaries. RIBES is used as the scoring model as it is invariant to word order and morphology (Tan et al., 2015).

Our results suggest that the induced dictionary performs at par or better than the original manual dictionary. This is also due to the fact that the lexicons are rendered in a context-sensitive manner from word embedding. The lookup process is implemented using Apache Spark® Framework in Scala language. Induction is a simple reverse lookup using the Cartesian product of all bilingual embedding. The size of this Cartesian product matrix is $1 \times 10^7 \times 300$ values which makes it highly computational. Apache Spark can run in parallel, hence, accelerate time and optimise memory. In this paper, bilingual embedding generated by BilBOWA (Gouws et al., 2015) is mainly used, but this methodology is also tested with cross-lingual embedding and found equally effective (JP et al., 2020). The differences between them are: bilingual embeddings are generated from parallel and good quality comparable bilingual corpus, whereas cross-lingual embedding can be learned from minimal bilingual data. Learning such cross-lingual embedding for resource-poor languages can help to generate induced dictionary resources of even unknown words with a fair amount of accuracy.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256.

Yidong Chen, Xiaodong Shi, Changle Zhou, and Qingyang Hong. 2009. A word alignment model based on multiobjective evolutionary algorithms. *Computers & Mathematics with Applications*, 57(11):1724 – 1729. Proceedings of the International Conference.

Yaacov Choueka, Ehud S. Conley, and Ido Dagan. 2000. *A comprehensive bilingual word alignment system*, pages 69–96. Springer Netherlands, Dordrecht.

Béatrice Daille and Emmanuel Morin. 2008. An effective compositional model for lexical alignment. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 748–756.

Elżbieta Hajnicz, Anna Andrzejczuk, and Tomasz Bartosiak. 2016. Semantic layer of the valence dictionary of Polish walenty. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2625–2632, Portorož, Slovenia. European Language Resources Association (ELRA).

Jiaji Huang, Qiang Qiu, and Kenneth Church. 2019. Hubless nearest neighbor search for bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4072–4080, Florence, Italy. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 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. Association for Computational Linguistics.

Varga István and Yokoyama Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 862–870, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sanjanasri JP, Vijay Krishna Menon, and Soman KP. 2020. BUCC2020: Bilingual dictionary induction using cross-lingual embedding. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 65–68, Marseille, France. European Language Resources Association.

Sarveswaran Kengatharaiyer, Gihan Dias, and Miriam Butt. 2019. Thamizhifst: A morphological analyser and generator for tamil verbs.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, New York, NY, USA.

Adrien Lardilleux, Julien Gosme, and Yves Lepage. 2010. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sanjanasri J. P., B. Premjith, Vijay Krishna Menon, and K. P. Soman. 2020. centam: Creation and validation of a new english-tamil bilingual corpus. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora, BUCC@LREC 2020, Marseille, France, May, 2020*, pages 61–64. European Language Resources Association.

Prof. Shikhar Kr. Sarma. 2019. Assamese-english bilingual dictionary. CLARIN-PL digital repository.

Paulo Schreiner, Aline Villavicencio, Leonardo Zilio, and Helena M. Caseli. 2011. Improving lexical alignment using hybrid discriminative and post-processing techniques. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Vijay Sharma and Namita Mittal. 2018. *Cross-Lingual Information Retrieval: A Dictionary-Based Query Translation Approach*, pages 611–618.

Li Ling Tan, Jonathan Dehdari, and Josef van Genabith. 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.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 297–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2004. Word to word alignment strategies. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Tufiş and Ana maria Barbu. 2002. Lexical token alignment: experiments, results and applications. In *In Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas*, pages 458–465.

Krzysztof Wołk. 2019. *Machine Learning in Translation Corpora Processing*.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. 2014. Pivot-based bilingual dictionary extraction from multiple dictionary resources. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 221–234, Cham. Springer International Publishing.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. 2015. A constraint approach to pivot-based bilingual dictionary induction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1):4:1–4:26.

Zhonglin Ye, Zhen Jia, Junfu Huang, and Hongfeng Yin. 2016. Part-of-speech tagging based on dictionary and statistical machine learning. In *2016 35th Chinese Control Conference (CCC)*, pages 6993–6998.

Kun Yu and Junichi Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 121–124, Boulder, Colorado. Association for Computational Linguistics.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65.