Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Automatic Bilingual Dictionary Construction for Tirukural

C.N Subalalitha & E. Poovammal

To cite this article: C.N Subalalitha & E. Poovammal (2018) Automatic Bilingual Dictionary Construction for Tirukural, Applied Artificial Intelligence, 32:6, 558-567, DOI: 10.1080/08839514.2018.1481590

To link to this article: https://doi.org/10.1080/08839514.2018.1481590

Published online: 18 Jun 2018.

APPLIED ARTIFICIAL INTELLIGENCE
2018, VOL. 32, NO. 6, 558–567
https://doi.org/10.1080/08839514.2018.1481590

Taylor & Francis
Taylor & Francis Group

Automatic Bilingual Dictionary Construction for Tirukural

C.N Subalalitha and E. Poovammal

Department of Computer Science and Engineering, SRM University, Kancheepuram, India

ABSTRACT

The Tirukural is a classic Tamil Sangam literature authored by Thiruvalluvar. Tirukural comprises of 1130 kuratpas. It has been translated into 37 world languages. This necessitates the cross-lingual access of Tirukural on the World Wide Web for which a bilingual dictionary is the primary knowledge base (KB). This KB needs to be constructed. This article puts forth a methodology for automatic construction of a bilingual dictionary for Tirukural in two different languages: Tamil and English. The proposed methodology makes use of the English and Tamil explanatory texts of Tirukural. Explanatory texts in English written by G.U. Pope and that in Tamil by Dr Varadharajan and Dr Solomon Pappaiya are considered in this work. A three-layered model is built using Tirukural and its explanations. Naive Bayes probabilistic learning is used to learn the best mappings between the Tamil and English words. The proposed methodology has been tested with all the 1330 Tamil kuratpas. An efficiency of 70% has been achieved and a performance comparison has been done by using different types of English and Tamil explanatory texts. This method can further be extended to build bilingual dictionaries for other Tamil literatures.

Introduction

Tirukural is dated to sometime between the third and first centuries BCE (Blackburn 2000). The Tirukural is a classic Tamil Sangam literature consisting of 1330 couplets or kurals or kuratpas authored by Thiruvalluvar. "Seiyul" or "pa" is one of the Tamil literature categories which should possess a predefined grammatical structure and should be precise in explaining the context. "Seiyul" or "pa" is of four types, namely "venpa," "aasiriyappa," "vangippa" and "kalippa" (BalaSundaraRaman, Ishwar, and Ravindranath 2003). Kuratpa is a sub category of "venpa" which explains the concept in two lines. The Tirukural is organized into 133 chapters, each containing 10 couplets, for a total of 1330 couplets. Each chapter has a theme that explains how a human should live a life in a righteous path. Tirukural has been translated into 37 world languages which necessitates the access of Tirukural in any language on the World Wide Web (Kumar 2010). In order to access the Tirukural on the web, a dictionary is the primary knowledge base that is required to process the Tirukural words. A bilingual dictionary

CONTACT C.N Subalalitha subalalitha@gmail.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

will aid in building a cross-lingual access for Tirukural, wherein a user can give a query in any language and retrieve the Tirukural in Tamil with its explanations in English or in any language. This article focuses on Tamil–English bilingual dictionary. Though many Tamil–English bilingual dictionaries exist, those dictionaries do not have meaning for the Tamil words used in ancient literatures. This article presents a methodology to automatically generate a bilingual dictionary for words used in Tirukural, which may further be extended for any ancient Tamil literatures, such as "aganaaru," "puranaaru," "tholkappiam," "nanool," etc.

Semantic processing of Tirukural becomes quintessential as it will aid in semantic indexing of Tirukural. Semantic indices can be used to build Natural Language Processing (NLP)-based web or mobile applications, such as Information Retrieval System, Summary Generation System, and Question Answering System. Semantic processing of any Tamil literature requires an explanatory text. Since the Tamil seiyuls are written in a precise manner, they often need an explanatory text to understand the meaning of it. Tirukural also has explanatory texts in all the languages in which it is translated. In this article, we make use of the Tirukural explanatory texts in Tamil and in English to construct the bilingual dictionary. The bilingual dictionary is constructed by using a three-layered model, wherein the first layer is the Tirukural in Tamil, the second layer is the explanatory text in Tamil, and the third layer is the explanatory text in English. Naive Bayes probabilistic learning model is incorporated between the second and the third layer to find the best match between the Tamil words and the English words.

A performance evaluation and comparison have been done by testing the proposed technique on two different Tamil explanatory texts which forms the second layer. By doing so, the size of bilingual dictionary is enhanced as the two different explanations will have different word usages which will contribute in enhancing the semantics of the bilingual dictionary. Also, the accuracy of the translation is also cross-checked for the unique set of words present in the second layer.

To sum up, the main contributions of this article are twofold

(1) Bilingual dictionary construction for Tirukural which lacks explicit grammatical structure.
(2) Usage of Naive Bayes probabilistic model to find the Tamil–English word mappings.

The rest of the article is organized as follows. "Literature survey" section describes the existing works on bilingual dictionary construction. "Proposed work" section illustrates the proposed methodology, "Evaluation" section gives details on evaluation of the proposed system, and "Conclusion and future work" section gives the conclusion and future works.

Literature survey

The bilingual dictionary construction proposed by McEwan, Ounis, and Ruthven (2002) makes use of parallel documents in English and Spanish. Parallel sentences have been collected by using various filters, namely language filters, length filters, and structural filters. Language filters have been used to check if the language is English or Spanish. Length filters have been used to compare the length of the parallel sentences and structural filters have been used to compare the Hypertext Markup Language (HTML) tags. The parallel sentences are aligned using appropriate HTML tags. The bilingual dictionary is constructed by mapping the words that have more probability of associating together. This is calculated by forming a co-occurrence matrix and the Expected Mutual Information Measure (EMIM) has been used to find the words that co-occur together frequently.

Hiroyuki Kaji et al. (2008) have proposed a bilingual dictionary construction for Japanese and Chinese language. This method makes use of Japanese–English dictionary and Chinese–English dictionary. In order to overcome the spurious translation, the Japanese word associations and Chinese word associations have been extracted from the monolingual Japanese and Chinese corpora. A correlation matrix has been constructed for the word associations. Mutual Information (MI) metric has been used to calculate the words that have high probability of getting associated.

Smith, Quirk, and Toutanova (2010) have constructed many bilingual lexicons for many language pairs, namely Japanese–English, Japanese–Chinese, and Chinese–English. The lexicon is built from the parallel sentences which are in turn identified by using a topic model and context-based methods. These methods explore the fact that words expressing same topics and context tend to co-occur frequently. Also, such words will not occur in non-topic texts and in dissimilar contexts. Finally, the lexicon is identified from the parallel sentences by using several features, such as sentence length, alignment features such as length of the noncontiguous text span and sub string. Wikipedia data have been used for the experiment.

Dalianis, Xing, and Zhang (2010) have constructed a Chinese–English parallel corpus and from the parallel corpus, bilingual dictionary has been constructed. The parallel corpus is created manually and a tool called Uplug is utilized, which makes use of word alignment features to generate the bilingual word pairs.

Bilingual English–Hindi dictionary for the documents having same topics has been proposed (Dubey and Varma 2013). The idea behind this work is that such documents will have similar structural properties such as sections, subsections, etc. This work has focused on transliterated words which are primarily named entities.

It can be observed that the existing works on bilingual dictionary construction have widely made use of parallel corpora. The parallel sentences have been compared and the word mappings have been done by analyzing the probability of word associations using co-occurrence matrices. The metrics such as EMIM and MI have also been used to find the strongest word associations. In order to extract the parallel web documents, HTML tags have been used.

The proposed work constructs a bilingual dictionary for Tamil–English language by using parallel corpus. Instead of using MI and EMIM metrics, the proposed work has adapted Naive Bayes probabilistic model to find the word associations. The length of the text span in which the word associations to be performed is limited to seven words, since all the kurals are having seven words. In the proposed approach, it is preferred to use simple Naive Bayes probabilistic model because the usage of MI and EMIM may increase the complexity of the model.

The proposed work differs from the existing works by the fact that the existing works generate bilingual dictionary using expository texts. Expository texts or essay type texts have a structured format. In such case, syntactic analysis can be used to generate the bilingual dictionary, whereas the proposed work builds a bilingual dictionary for a literature text which lacks explicit grammatical structure. Hence, the proposed work lays the foundation which can be extended to generate bilingual dictionaries for other Tamil literatures.

Proposed work

Figure 1 shows the proposed three-layered model for bilingual dictionary construction for Tirukural. Layer-1 consists of Tirukural written by Thiruvalluvar in Tamil. Layer-2 consists of *standard Tamil* explanations written *for* Tirukural *by* Dr Varadharajan and Dr Solomon Pappaiya. The third layer consists *of English explanation of* Tirukural Layer-2. The bilingual dictionary construction is done in two steps.

Step 1:

The words in Layer-1 (W_1) and Layer-2 (W_2) are matched completely or partially and paired up. The roots of the words in Layer-2 are identified using a Tamil Morphological Analyzer, which is an open source tool developed at Tamil Computing lab, Anna University. Tamil Morphological Analyzer identifies the grammatical components of the word (Anandan et al. 2002). A Tamil–Tamil dictionary is used to find the synonym pairs. For each $w_2 \in W_2$, the corresponding English meaning is identified using a Tamil–English dictionary. Hence, an initial list of Tamil–English word pairs is generated. This process is explained through the Example 1. The architecture of the three-layered model is depicted in Figure 2.

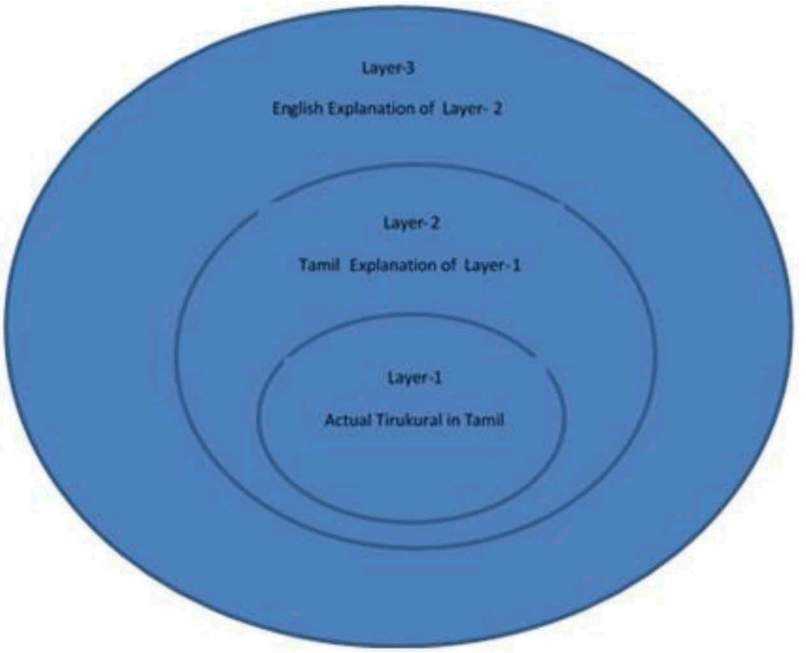

Figure 1. Three-layer model for Tirukural dictionary construction.

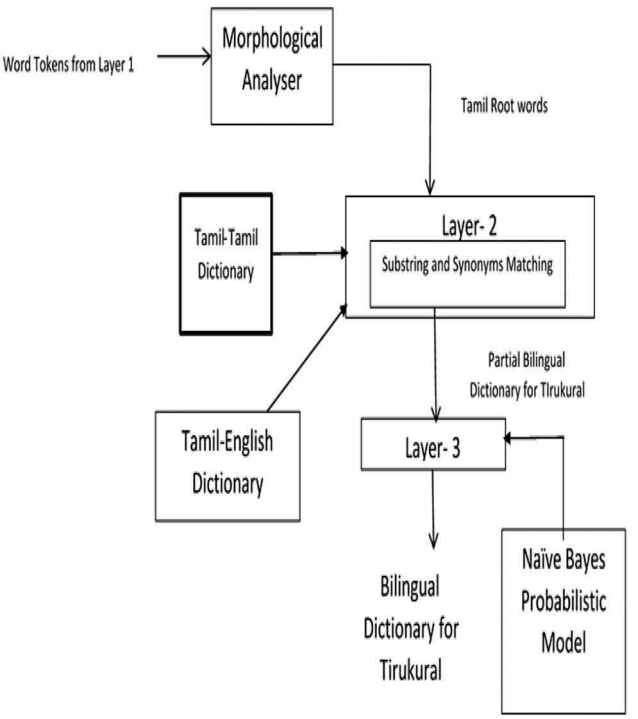

Figure 2. Architecture of the bilingual dictionary creation.

Example 1

Layer-1 – Tirukural in Tamil

அகர முதல எழுத்தெல்லாம் ஆதி

பகவன் முதற்றே உலகு

English transliteration: Akara mutala eḻuttellām āti
pakavaṉ mutaṟṟē ulaku

Layer-2 – Tamil explanation for Layer-1

எழுத்துக்கள் எல்லாம் அகரத்தை அடிப்படையாக கொண்டிருக்கின்றன.

அதுபோல உலகம் கடவுளை அடிப்படையாக கொண்டிருக்கிறது.

English Transliteration: Eḻuttukkaḷ ellām akarattai aṭippaṭaiyāka koṇṭirukkiṉṟaṉa. Atupōla ulakam kaṭavuḷai aṭippaṭaiyāka koṇṭirukkiṟatu.

Layer-3 – English explanation for Layer-2

As all letters have the letter A for their first, so the world has the Eternal God for its first.

As mentioned earlier, the following words in Layer-1 and Layer-2 gets paired up through complete/partial word match.

அகர- *அகரத்தை* (Akara- akarattai)

எழுத்தெல்லாம்- *எழுத்துக்கள்* (Eḻuttellām- eḻuttukkaḷ)

பகவன்- *கடவுளை* (Pakavaṉ- kaṭavuḷai)

உலகு- *உலகம்* (Ulaku- ulakam)

The word pair "*பகவன்- கடவுளை*" are synonyms and are paired up using the using Tamil–Tamil dictionary. Then, by using the Tamil–English dictionary, the following words from Layer-3 are matched.

அகர (Akara)- A

எழுத்தெல்லாம் (Eḻuttellām)- letters

பகவன்(Pakavaṉ)-God

உலகு (Ulaku)- world

Step 2:

The words left in both Layer-1 and Layer-3 in Step 1 are processed in this step by using a statistical model. The statistical model is built using Naive Bayes probabilistic learning. This statistical model is built by exploring the following assumption.

The words left in Layer-1 and Layer-3 are definitely related as Layer-3 is the explanatory translation of the words present in Layer-1. The Tamil word and its

corresponding English word can be mapped by finding their co-occurrence probability in Tirukural.

Naive Bayes probabilistic model

The Model uses the words left after Step 2 of bilingual dictionary construction as training set. K is a set of kuratpas present in the training set, W_{Tamil} is a set of words present in the training set in Layer-1, and W_{English} is a set of words present in Layer-3.

(a) For each Kuratpa $k_i \in K$, the probabilities of the word $w_{\text{Tamil}} \in W_{\text{Tamil}}$ present in Layer-1 occurring with a word $w_{\text{English}} \in W_{\text{English}}$ present in Layer-3 are calculated using the training set.
(b) The maximum probability of each word in w_{Tamil} and its corresponding w_{English} is identified using the Naive Bayes probability given in Equations (1) and (2).
(c)

$$P\left(\frac{w_{\text{English}}}{w_{\text{Tamil}}}\right) = \frac{\left(P\left(w_{\text{English}}\right) * P\left(\frac{w_{\text{Tamil}}}{w_{\text{English}}}\right)\right)}{P\left(w_{\text{Tamil}}\right)} \tag{1}$$

which can be simplified as

$$P(w_{\text{English}}) = \text{Argmax}\left(P\left(w_{\text{English}}\right)\prod_{i=0}^{i=n} * P\left(\frac{w_{\text{Tamil}}}{w_{\text{English}}}\right)\right. \tag{2}$$

where n is the number of words present in Layer-3.
Hence, for each w_{Tamil}, the most probable w_{English} is identified by the Naive Bayes probabilistic model. The procedure is explained using Example 2.

Example 2: In the example considered, Example 1, the stop words are removed and the Tamil words left in Layer-1 are given below:

முதல ஆதி முதற்றே

The English words left in Layer-3 will be as follows:

First eternal first.

By using the Naive Bayes probabilistic model, the following word mappings are identified which has the maximum probability.

முதல (Mutala) -First
ஆதி (Āti) - eternal
முதற்றே (Mutarrē) – First.

Hence, Tamil–English bilingual dictionary for the Tirukural words is generated.

Evaluation

The proposed methodology has been tested using all 1330 kuratpas and its Tamil and English explanations. Since the proposed methodology is the first of its kind, a comparison with the state-of-art system has not been made, but a performance comparison has been done by testing the proposed methodology by using two different standard Tamil explanations authored by Dr Varatharajan and Dr Solomon Pappaiya. The performance is evaluated using the metric "*Precision*," which is defined as per Equation (3).

$$\text{Precision} = \frac{\text{Number of correct Tamil} - \text{English word mappings present in the bilingual dictionary}}{\text{Total number of word mappings present in the bilingual dictionary}} \tag{3}$$

Out of 9310 words present in the Tirukural, the proposed approach was able to find 1390 unique set of Tamil–English word mappings. Out of 1390 word mappings, 980 word mappings were obtained without using the Naive Bayes probabilistic model. The precision of the Naive Bayes probabilistic model is 68%. Table 1 shows the precision values of the proposed approach using two Tamil explanatory texts. The precision values are calculated by using human judgment involving three human judges. Table 1 shows the average precision scores of three human judges.

It was observed that by using two different Tamil explanatory texts, the final bilingual dictionary contained different Tamil words mapped to the same English words. This was due to the difference in word usage in the two different Tamil explanatory texts. Seventy-five such different Tamil words were mapped to the same English word.

The proposed methodology makes use of the maximum probability of the Tamil and English words occurring together to find the Tamil–English word mappings. The accuracy can further be increased by enhancing the features set used for the word mappings. Semantic relations can be identified in the Tirukural in Tamil and English. These semantic relations can in turn be

Table 1. Evaluation of the proposed approach.

Factors	Precision values using Dr Varatharajan Explanatory Text	Precision values using Dr Solomon Pappaiya Explanatory Text
Number of total Tamil–English word mappings	1390	1275
Number of total correct Tamil–English word mappings	980	899
Number of total Tamil–English word mappings identified by Naive Bayes probabilistic model	410	398
Number of total correct Tamil–English word mappings identified by Naive Bayes probabilistic model	279	258
Precision percentage of the probabilistic model	68	54.8
Precision percentage of the proposed approach	70	70.5

used as the feature set to find the word mappings. For instance, semantic relations such as Universal Networking Language (UNL) relations and Rhetorical Structure (RS) relations can be identified in the Tirukural which will increase the accuracy of the bilingual dictionary (Mann and Thompson 1988; Uchida, Zhu, and Della Senta 1999). The accuracy can also be enhanced by using a pure statistical word alignment model between the layers by exploring the structure of the Tirukural, which has four words in the first sentence and three words in the second sentence. The proposed methodology suits Tirukural which has limitation of seven words in each kural. This methodology needs enhancement to construct bilingual dictionary for other literatures. For the enhancement, UNL and RS relations can be used to find the word mappings.

Conclusion and future work

Bilingual dictionary construction for Tirukural has been proposed in this article. The Tamil words are mapped with the English words by using a three-layered model in which Layer-1 is the Tirukural in Tamil, Layer-2 is the Tamil explanation of Tirukural, and Layer-3 is the English explanation of Tirukural explanation in Layer-2. Naive Bayes probabilistic model has been incorporated to identify the maximum probabilities of co-occurring Tamil and English words.

In order to construct a bilingual dictionary for other Tamil literatures, a better probabilistic model to find the mappings need to be designed. By doing so, the hidden information in Tamil literature can be extracted by building useful NLP applications. Also, semantic features need to be incorporated to analyze literature which have lengthy sentence patterns. The words not handled by the Morphological Analyzer need to be analyzed and a methodology to handle them need to be identified which will eventually increase the efficiency and precision of the proposed methodology.

References

Anandan, P., K. Saravanan, R. Parthasarathi, and T. V. Geetha. 2002. Morphological analyzer for Tamil. Proceedings of International Conference of Natural Language Processing, Mumbai, December.

BalaSundaraRaman, L., S. Ishwar, and S. K. Ravindranath. 2003. Context Free Grammar for Natural Language Constructs - An implementation for Venpa Class of Tamil Poetry". Proceedings of Tamil Internet, Chennai, International Forum for Information Technology in Internet. 128–136.

Blackburn, C. 2000. Corruption and redemption: The legend of Valluvar and Tamil literary history (PDF). *Modern Aian Studies* 34(2):449–82. doi:10.1017/S0026749X00003632.

Dalianis, H., H. Xing, and X. Zhang. 2010. Creating a reusable English-Chinese parallel corpus for bilingual dictionary construction. Proceedings of the International Conference on Language Resources and Evaluation, Valletta, Malta, 1–4.

Dubey, A., and V. Varma. 2013. Generation of bilingual dictionaries using structural properties. *Computacion Y ´ Sistemas* 17(2):161–168.

Kaji, H., Shin'ichi Tamamura and Erdenebat, D., 2008. Automatic Construction of a Japanese-Chinese Dictionary via English. In LREC. 699–706.

Kumar, R. 2010. *Morality and ethics in public life*, 92. New Delhi, India: Mittal Publications.

Mann, W. C., and S. A. Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text* 8(3):243–81. doi:10.1515/text.1.1988.8.3.243.

McEwan, C. J. A., I. Ounis, and I. Ruthven. 2002. Building bilingual dictionaries from parallel web documents. Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, London, UK, Springer-Verlag. 303–323.

Smith, J. R., C. Quirk, and K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 403–11, Los Angeles, California, June.

Uchida, H., M. Zhu, and T. Della Senta. 1999. *A gift for a millennium*. Tokyo: IAS/UNU.