FMDB *Transactions on Sustainable Computing Systems*

Context-Driven Document-Specific Word Translation for Empowering Low-Resource Neural Machine Translation

S. Senthamizh Selvi[1,*], R. Anitha[2], O. Jeba Singh[3], S. Rubin Bose[4], Mohammad Ayaz Ahmad[5]

[1,2]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai, Tamil Nadu, India.
[3]Centre for Academic Research, Alliance University, Bengaluru, Karnataka, India.
[4]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[5]Department of Mathematics, Physics and Statistics, University of Guyana, Georgetown, Guyana, South America.
senthamizhselvi@svce.ac.in[1], ranitha@svce.ac.in[2], jeba.singh@alliance.edu.in[3] , rubinbos@srmist.edu.in[4], mohammad.ahmad@uog.edu.gy[5]

Abstract: When translating specific words from one language to another, it is necessary to take into consideration several different variables to guarantee correctness, precision, quality, and contextual appropriateness. Context-aware translation systems face several issues due to the complexities of natural language. These challenges include contextual disambiguation, cultural and linguistic nuances, subject matter expertise, and maintaining translation consistency. In low-resource languages such as Tamil, which suffer from limited linguistic resources, high word-sense ambiguity, and flexible word order, the translation process is further complicated. These issues are magnified in low-resource languages. To address these concerns, the proposed study will provide linguistic resources for pre-training Neural Machine Translation (NMT) systems. This will be accomplished by developing a translation system that is language-independent, document-specific, and context-based, making it suitable for languages with limited resources. The research presents a novel word vector format known as Word2Line (W2L), which minimises the time required for the process. With the use of TF-IDF, the system can recognise document-specific words in the English language, which serves as the source language, and then predict the context-based translations of those words in Tamil, the target language.

Keywords: Neural Machine Translation; Data Mining Confidence; Context-Aware Translation; Natural Language; Maximum Likelihood Estimate; Document-Specific Translation.

Received on: 22/09/2024, **Revised on:** 06/12/2024, **Accepted on:** 11/01/2025, **Published on:** 05/06/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCS

DOI: https://doi.org/10.69888/FTSCS.2025.000434

Cite as: S. S. Selvi, R. Anitha, O. J. Singh, S. R. Bose, and M. A. Ahmad, "Context-Driven Document-Specific Word Translation for Empowering Low-Resource Neural Machine Translation," *FMDB Transactions on Sustainable Computing Systems*, vol. 3, no. 2, pp. 101–113, 2025.

1. Introduction

[*]Corresponding author.

Machine Translation (MT) is a subfield of computational linguistics that focuses on automating the translation of text or speech from one language to another. The goal of MT is to overcome language barriers in communication efficiently and accurately. Traditional MT systems have evolved significantly from rule-based methods to statistical methods and, more recently, to advanced neural network approaches. Rule-Based Machine Translation (RBMT) relies on predefined linguistic rules and dictionaries to translate text [1]. These systems struggled with flexibility and required extensive manual effort to support multiple language pairs and contexts. Statistical Machine Translation (SMT) leverages large parallel corpora to learn probabilistic models of translation. This approach improved translation quality but encountered challenges in contextual understanding and handling idiomatic expressions [2]. Neural Machine Translation (NMT) represents a significant advancement in the field of machine translation, leveraging deep learning techniques to provide more accurate, fluent and contextually aware translations. NMT models, particularly those based on transformer architectures, have shown remarkable improvements in handling context, a crucial factor for producing high-quality translations. The NMT employs an end-to-end learning approach, utilising an encoder-decoder architecture, attention mechanisms, transformer models, and contextual embeddings to enhance the system's ability for contextual translation [3]; [4]. The encoder processes the entire input sentence to capture its context, while the decoder generates the translation using that context. Attention mechanisms focus on relevant parts of the input sentence, especially in long sentences. The transformer models process input data in parallel to retain context over long sequences. While significant progress has been made in machine translation, achieving context-aware translations remains a challenging task.

Context-based machine translation faces challenges in understanding contextual information, such as polysemy and homonymy, maintaining coherence and cohesion during anaphora resolution and textual coherence, adapting to cultural and linguistic nuances, evaluating quality and accuracy, and addressing technical limitations, including scalability and real-time translation. A context-based document-specific translation system is a valuable tool that leverages various techniques from different research papers. For instance, the system can utilise Deep Neural Networks (DNNs), Hidden Markov Models (HMMs), and Gaussian Mixture Models (GMMs) for Automatic Speech Recognition (ASR) in spoken document retrieval [5]. Additionally, corpus-based methodologies can significantly enhance lexical competence and search strategies as demonstrated in public service interpreting and translation (PSIT) studies [6]. Furthermore, a context-based generic cross-lingual retrieval model can be employed to handle different language pairs effectively by considering context in query translation and document retrieval [7]. Moreover, in NMT, context gates can dynamically control the contributions of source and target contexts to improve translation adequacy and fluency, thereby addressing the limitations of conventional NMT systems [8]. Some of the approaches employed to enhance contextual NMT in specialised fields include pre-training and fine-tuning models, transfer learning by applying knowledge from one language pair to another, hybrid models that combine NMT with rule-based or statistical approaches, incorporating feedback from human translators, and integrating multimodal translation. Among these approaches, the hybrid model offers a powerful solution for enhancing contextual NMT in low-resource languages.

Enhancing context-based NMT for low-resource languages is a challenging task due to the scarcity of parallel corpora and linguistic resources [9]. Data augmentation and transfer learning, combining multiple approaches such as RBMT and SMT integration, Multilingual pre-training and cross-linguistic transfer, shared encoder-decoder frameworks, and zero-shot translation, can enhance translation quality and are among the strategies to improve contextual NMT for low-resource languages [10]; [11]; [12]. Engaging human translators for real-time feedback on translations can improve contextual understanding. Crowdsourcing annotated data can expand the training dataset for low-resource languages, thereby enhancing their capabilities. Linguistic resources such as morphological analysers, parsers, and bilingual dictionaries can enrich input data for translation models. Hybrid NMT systems can significantly enhance contextual translation quality in low-resource languages by combining rule-based and statistical approaches with neural methods and leveraging linguistic resources, data augmentation, transfer learning, and human feedback. The ongoing research, innovation and development in this area are essential to further improve the capabilities of NMT systems for low-resource languages. The proposed work leverages cross-lingual transfer learning to improve the translation performance of hybrid NMT models. It aims to create linguistic resources, such as context-based, document-specific bilingual dictionaries in English and Tamil, to pre-train the model with contextual information. The proposed work primarily makes the following contributions:

- A small volume of a comparable corpus is manually created from the English and Tamil versions of the VIII-grade school textbook.
- The proposed work introduced a novel Word2Line (W2L) vector representation of a word, which is more compatible with finding the contextual closeness between English and Tamil Words. The words that have similar meanings appear closer in the vector space.
- The proposed algorithm created context-based, document-specific linguistic resources using the TF-IDF algorithm with a modified association rule mining confidence approach to predict translation words. This algorithm is suitable for any low-resource language as the target language, as it does not require a large pre-annotated corpus or information about the surrounding words of source or target words.
- Finally, the proposed work was evaluated quantitatively and manually verified.

2. Related Work

Machine translation faces challenges in low-resource language scenarios due to data sparsity, which hinders the performance of NMT models [13]; [14]. To address this, researchers have proposed novel approaches such as Syntax-Graph guided Self-Attention (SGSA), which combines syntactic knowledge with multi-head self-attention to significantly improve NMT performance, especially in low-resource languages [15]. Additionally, the lack of robust evaluation benchmarks for low-resource languages has been identified as a major obstacle to assessing model quality, leading to the introduction of Flores-101, a benchmark that covers a wide range of topics and languages, enabling better evaluation of machine translation systems for low-resource languages [16]. Neural Machine Translation (NMT) faces challenges in low-resource languages, including limited datasets and high costs associated with data collection [17]. To address these challenges, researchers have proposed methods such as dual attention mechanisms and the use of large-scale paired datasets for training [17]. The size of parameters in attention-based NMT models is increased to ensure translation quality. This necessitates the use of compression methods such as linearization, weight compression, and near-memory hardware decoders to reduce latency and energy consumption [18]. The reliability and usability of automatically translated patient health information generated by NMT tools can pose safety risks to multicultural communities with limited bilingual skills, underscoring the importance of assessing the quality of NMT translations in clinical settings [19]. The NMT model is trained to predict target sentences as categorical outputs and embeddings, leveraging pre-trained embeddings to generalise beyond limited training data [13]. The new word embedding introduced by Chen [20] incorporates prior knowledge and shares training results across all embeddings, resulting in significant performance improvements in low-resource translation tasks.

Additionally, utilising monolingual corpora for training word embedding and language models enhances Multimodal Machine Translation (MMT) systems, thereby improving translation quality across various language pairs [21]. The challenges faced in machine translation for the Tamil language stem from its rich morphology, limited resources, high word-sense ambiguity, and unique word-order structure, as highlighted in multiple research papers, Prasanna and Latha [22], Gokila et al. [23]. To address these challenges, researchers have developed innovative solutions, including hybrid POS tagger algorithms, LSTM-based translation models, and NLP techniques integrated with ML and DL algorithms. The scarcity of annotated corpora and linguistic resources in Tamil adds complexity to the design of accurate translation systems, necessitating the use of cross-lingual transformation learning techniques. Context-based document-specific word translation for Tamil languages faces several limitations and challenges, particularly in Neural Machine Translation (NMT). These issues stem from the complexities of contextual integration and the unique linguistic features of Tamil. Traditional NMT systems often struggle with efficiently integrating extensive contextual information, leading to slower processing times [24]. The lack of effective mechanisms to balance source and target contexts can result in translations that are fluent but contextually inadequate. Tamil's complex structure, characterised by agglutination, complicates the modelling of contextual variations, making it difficult to achieve accurate translations [25]. The vast number of syllables in Tamil presents a challenge for consistent recognition and translation. Despite these challenges, ongoing research continues to explore innovative approaches to enhance context-based neural machine translation for low-resource languages by leveraging training methods and pre-existing linguistic knowledge, suggesting potential improvements for future NMT systems.

3. The Proposed Approach

The proposed architecture generates a context-based, document-specific bilingual dictionary (English-Tamil) system, as illustrated in Figure 1.

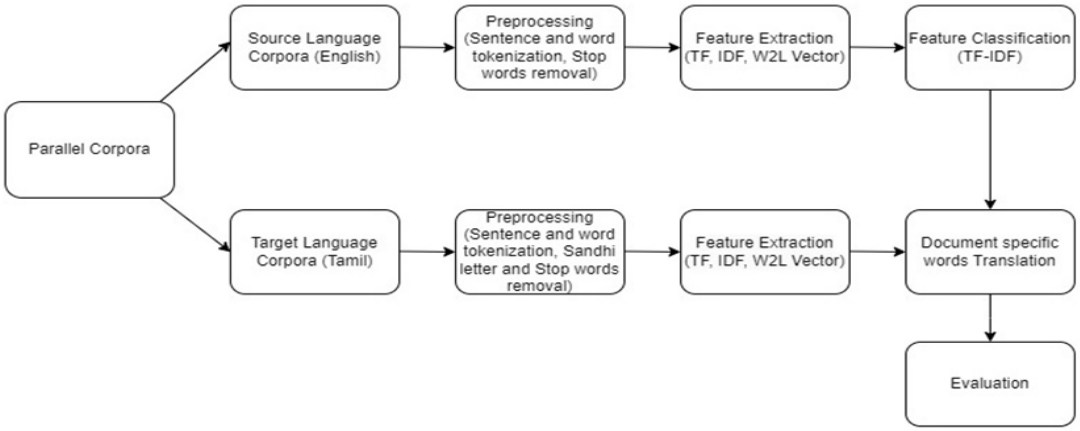

Figure 1: Architecture diagram of context-based document-specific bilingual dictionary (English-Tamil)

A small, manually created comparable bilingual corpus serves as input for the system. This corpus is pre-processed through sentence tokenisation, word tokenisation and stop-word removal. Key features, such as Term Frequency (TF), Inverse Document Frequency (IDF), and Word2Line (W2L), are extracted from the corpus. Document-specific words are identified using the TF-IDF method for the English corpus. For these document-specific words, the proposed algorithm applies W2L combined with association rule mining to predict their corresponding translations into Tamil. The resulting context-specific bilingual dictionary serves as a valuable linguistic resource for pre-training Neural Machine Translation (NMT) models. The performance of the proposed system is evaluated manually.

3.1. Data Collection - Corpus Creation

The proposed work requires a bilingual corpus of various documents. For this purpose, the first three chapters from both the English and Tamil versions of the 8th-grade science textbook published by the Tamil Nadu state government in India were selected to manually create three small parallel corpora. Each chapter is treated as a separate document. Documents 1, 2, and 3 correspond to the lessons titled "Crop Production and Management," "Reaching the Age of Adolescence," and "Pictorial Features of the Plant Kingdom," respectively. A sample from Document 1 is shown in Figure 2.

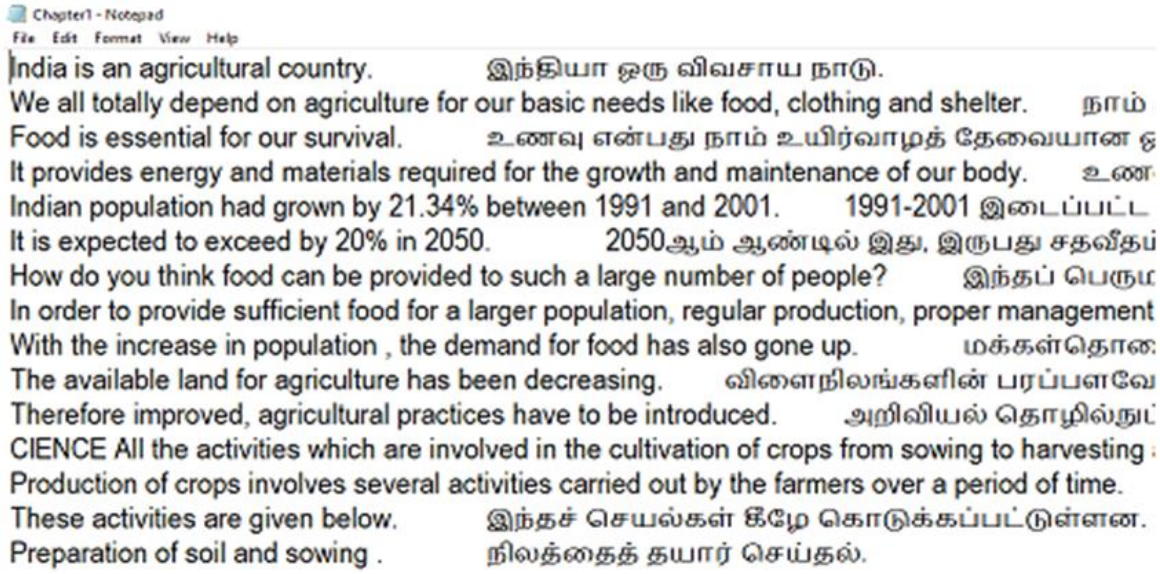

Figure 2: Sample of document 1

The feature of the three corpora is shown in Table 1. The proposed work uses English as the source language and Tamil as the target language. All three documents are very small in volume. The English corpus has unique words ranging from 149 to 550. The Tamil corpus has unique words ranging from 189 to 836. Although the corpus contains fewer than 1000 words, the proposed modified association rule mining algorithm can still accurately predict translated words.

Table 1: Features of input documents

Document	No. of Sentences	Language	No. of words	No. of words after stop words	No. of unique words after pre-processing
Document 1	124	English	1577	928	550
		Tamil	1205	1089	836
Document 2	57	English	739	433	265
		Tamil	464	424	307
Document 3	35	English	309	203	149
		Tamil	259	228	189

3.2. Corpus Pre-Processing

The bilingual documents are pre-processed using techniques for sentence tokenisation, word tokenisation, and stop-word removal.

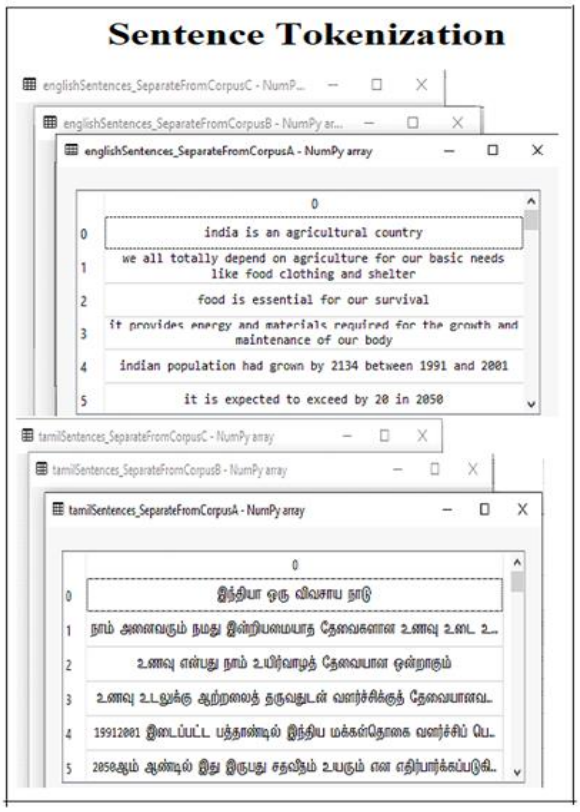

Figure 3: Sentence tokenisation

The sentence tokeniser splits large documents into collections of individual sentences. A word tokeniser splits a sentence into words.

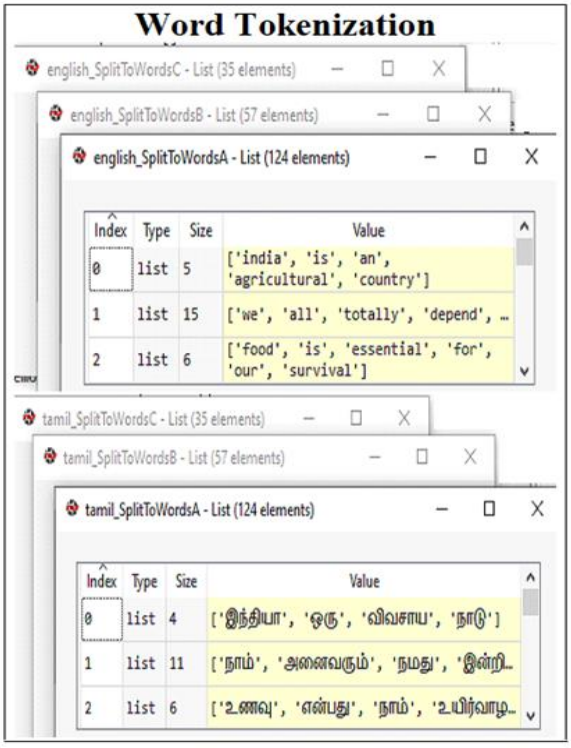

Figure 4: Word tokenisation

These tokenisers are required to predict the translated word in a sentence. Figure 3 and Figure 4 show the output after performing sentence tokenisation and word tokenisation on the corpus.

stop_words_tamil=["அங்கு", "அங்கே", "அடுத்த", "அதற்கு", "அதனால்", "அதன்", "அதிக", "அதில்", "அது", "அதே", "அதை", "அந்த", "அந்தக்", "அந்தப்", "அல்லது", "அவரது", "அவர்", "அவர்கள்", "அவள்", "அவன்", "அவை", "அன்று", "ஆகிய", "ஆகியோர்", "ஆகும்", "இங்கு", "இங்கே", "இடத்தில்", "இடம்", "இதற்கு", "இதனால்", "இதனை", "இதன்", "இதில்", "இது", "இதை", "இந்த", "இந்தக்", "இந்தத்", "இந்தப்", "இப்போது", "இரு", "இருக்கும்", "இருந்த", "இருந்தது", "இருந்து", "இவர்", "இவை", "இன்னும்", "உள்ள", "உள்ளது", "உள்ளன", "உன்", "எந்த", "எல்லாம்", "என", "எனக்", "எனக்கு", "எனப்படும்", "எனவும்", "எனவே", "எனினும்", "எனும்", "என்", "என்பது", "என்பதை", "என்ற", "என்று", "என்றும்", "என்ன", "என்னும்", "ஏன்", "ஒரு", "ஒரே", "ஓர்", "கொண்ட", "கொண்டு", "கொள்ள", "சற்று", "சில", "சிறு", "சேர்ந்த", "தவிர", "தனது", "தன்", "தான்", "நாம்", "நான்", "நீ", "பல", "பலரும்", "பல்வேறு", "பற்றி", "பற்றிய", "பிற", "பிறகு", "பின்", "பின்னர்", "பெரும்", "பேர்", "போது", "போல", "போல்", "போன்ற", "மட்டுமே", "மட்டும்", "மற்ற", "மற்றும்", "மிக", "மிகவும்", "மீது", "முதல்", "முறை", "மேலும்", "மேல்", "யார்", "வந்த", "வந்து", "வரும்", "வரை", "வரையில்", "விட", "விட்டு", "வேண்டும்", "வேறு"]

Figure 5: List of Tamil stop words

Words such as in, on, or, at, a, an, the, he, she, it, etc., are stop words. As these words do not play a significant role in the document-specific vocabulary, they are removed from the corpus. For the Tamil language, the finite set of stop words is identified as shown in Figure 5. The algorithm removes these words from the Tamil corpus, as shown in Figure 6.

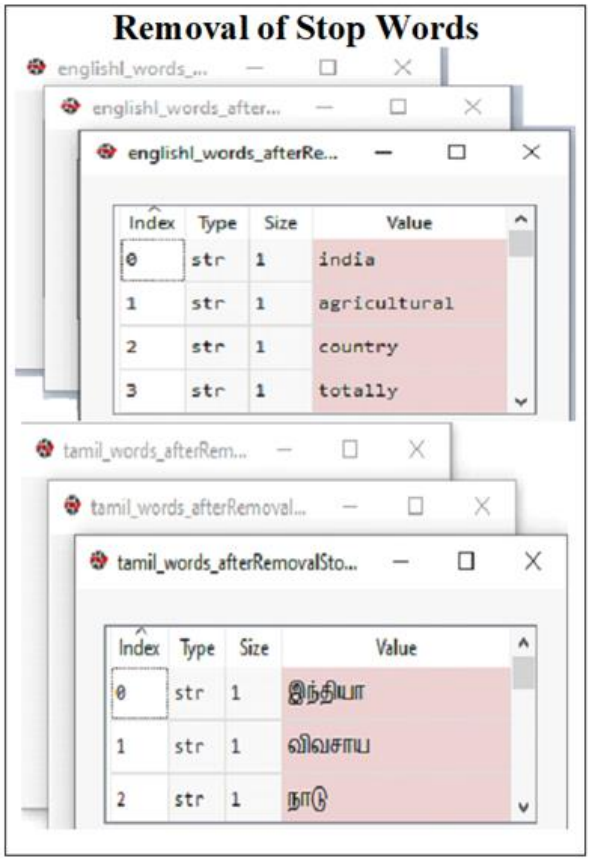

Figure 6: Stop word removal

3.3. Feature Extraction

Feature extraction is the process of converting raw data into numerical features. This text representation can be processed while preserving the information in the original dataset. Machine learning learns this data and yields better outcomes than learning directly from the raw data. The features of each word are represented as term frequency, inverse document frequency (IDF), TF-IDF, and Word2Vec vectors.

3.3.1. TF-IDF Vectorisation

Term Frequency-Inverse Document Frequency (TF-IDF) measures the importance of a word depending on how frequently it occurs within a document and a collection of documents. TF-IDF comprises Term frequency (TF) and Inverse Document Frequency (IDF). TF denotes the frequency of a word in a document. For a specified word, it is defined as the ratio of the number of times a word appears in a document to the total number of words in the document.

- TF (term) = Number of times term appears in a document / Total Number of terms in the document.

IDF measures the importance of the word in the corpus. It measures the frequency of a particular word across all documents in the corpus. It is the logarithmic ratio of the total number of documents to the number of documents with a particular word. If a word appears multiple times across many documents, then the denominator will increase, reducing the value of the second term. Thus, common words would have lesser importance.

- IDF (term) = log (Total amount of documents / Number of documents having term).
- The TF-IDF formula is the product of TF and IDF.
- TF-IDF (term) = TF (term) * IDF (term).

3.3.2. Word2Line (W2L) Vector Representation

The Word2Line representation of words is a novel approach proposed to simplify the calculation of confidence based on support in data mining. The Word2Line vector represents each word by the line numbers in which it occurs within the corpus. This applies to both the source and target languages. Unlike traditional NLP vector representations, such as one-hot encoding and bag-of-words, which capture only the presence or absence of words in observations and are useful for certain machine learning tasks, they fail to encode a word's meaning or contextual relationships. The proposed method captures potential relationships between source and target words by leveraging their proximity in sentence line numbers across both languages. As illustrated in the pictorial representation of the Word2Line vector for a sample set of words, the probability of correctly predicting target words like "பயிர்" (payir), "விவசாயிகள்" (vivasayikal) and "உணவு" (unavu) as translations for the source words "crop," "farmers," and "food," respectively, is higher due to their greater contextual proximity in the vector space (Figure 7).

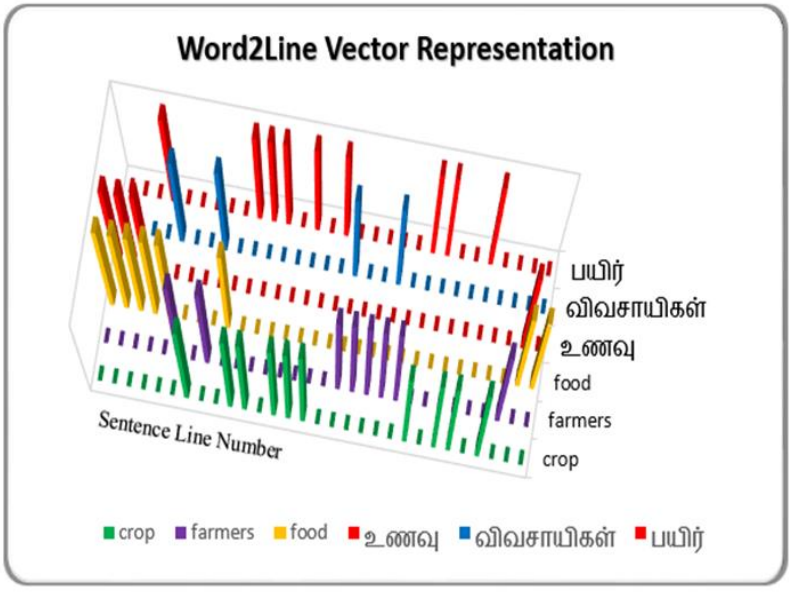

Figure 7: Pictorial representation of word2line vector for the sample of words

The number representation of the Word2Line vector is shown in Figure 8. The value shows that the first column is the word itself, the second column is the number of times the word appears in the document, and the remaining columns are the line numbers in which the word appears.

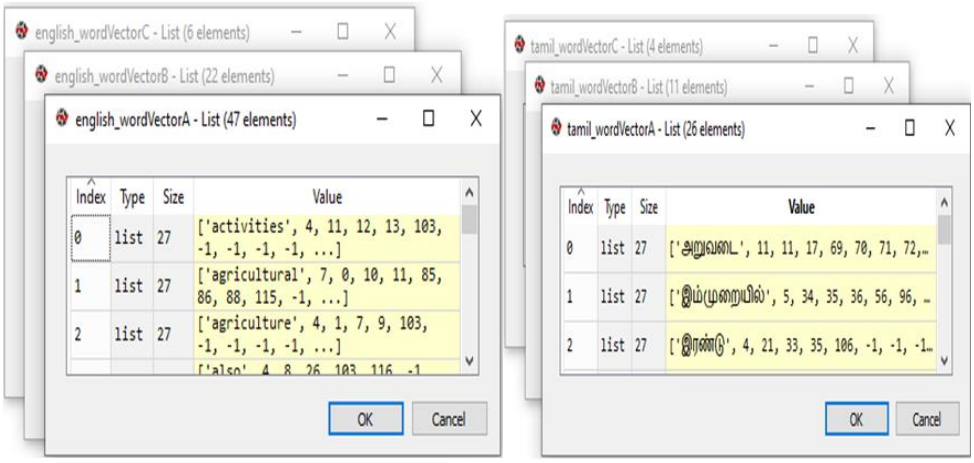

Figure 8: Word2Line vector number representation

3.4. Feature Classification

The feature classification is performed in the following stages, namely, the Classification of Chapter-specific words and the prediction of Translation words

3.4.1. Classification of Document-Specific Words

The TF-IDF technique is used to eliminate the most common terms and extract only the most relevant terms from the corpus. The TF-IDF algorithm was applied to 45 documents to extract term features [26]. An approach proposed by Dadgar et al. [27] involved classifying news texts using a support vector machine, with features extracted using the TF-IDF method. The TF-IDF algorithm was used to classify Bahasa Indonesia news articles [28]. A high precision has been achieved in this classification. This classification technique is used to identify chapter-specific words in the source language, English, for which the corresponding Tamil translation is predicted. The sample document-specific words are shown in Figure 9.

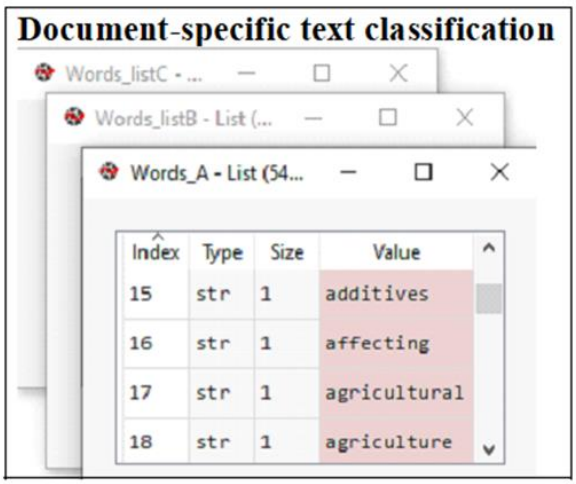

Figure 9: Document-specific words

3.4.2. Prediction of Translation Words Using a Mining Approach

Association Rule mining was originally designed to examine customer behaviour by identifying the items they purchase together in a single transaction [29]. Rules find relationships between sets of elements of every distinct transaction. Support

and Confidence metrics are used to understand the strength of association between sets of elements. Support indicates the frequency with which an itemset appears across all transactions. Consider X and Y to be the itemsets itemset1 and itemset2, respectively; the support is the fraction of transactions in which the itemset occurs.

- Support (X -> Y) = Transactions containing both X and Y / Total number of transactions.

Confidence indicates the likelihood that item Y will be purchased when item X is purchased, expressed as {X -> Y}. The proportion of transactions with item X, in which item Y also appears, measures this.

- Confidence (X -> Y) = Transactions containing both X and Y / Transactions containing X.

Association Rule Mining is a machine learning technique commonly used in Market Basket Analysis to identify frequent itemsets based on metrics such as support and Confidence. The proposed algorithm leverages this approach to predict the most associated translated target words for chapter-specific source words. Confidence represents the likelihood that a target word in a sentence will be predicted as the translation of a source word. It is calculated as the ratio of the support of both the source and target words to the support of the source word. The support of a source and target word is defined as the number of lines containing both words, divided by the total number of lines in the corpus. The support of the source word is the number of lines containing the source word, divided by the total number of lines in the corpus. To simplify computation, the Word2Line (W2L) vector for each word is represented as a Set. The confidence can then be expressed as:

$$\text{Confidence}(\text{Source word} \rightarrow \text{Target word}) = \frac{\text{No. of } \{ \text{ Set of line numbers of the source word} \} \text{ n } \{\text{Set of line numbers of the target word}\}}{\text{No. of } \{\text{Set of line numbers of the source word}\}}$$

The prediction algorithm for the translated word is shown in Algorithm I.

3.5. Algorithm I: Translation Text Prediction

- **Input:** Bilingual (English-Tamil) Parallel Corpus.
- **Output:** A context-based, document-specific bilingual dictionary with English as the source language and a list of its Tamil translations.
- After preprocessing the corpus, perform steps 2-4 separately for both the English and Tamil corpora.
- **Word Frequency Vector:** Calculate the frequency of each word and store it in a dictionary, where the key is a sorted list of unique words and the value is their respective frequency.
- **W2L Feature Extraction:** Extract features by identifying the line numbers where each word appears in the corpus.
- **Feature Vector Representation (FV):** For each word, create a list containing the word itself, its frequency and the line numbers where it occurs in the corpus.

3.5.1. Predicting English-to-Tamil Translations

For each word in the English feature vector (FVe), do the following:

- Let A be the set of line numbers where the English word Ei appears in sentence Sl, as indicated by its feature vector.
- For each word in the Tamil feature vector (FVt) in the same sentence Sl, do the following:
    - Let B be the set of line numbers where the Tamil word Tjappears, as indicated by its feature vector.
    - Find the intersection of A and B. If not null, add this to set C.
    - Calculate the length of C (denoted as X), which represents the number of times Ei and Tj appear together in the same line of the corpus.
    - Compute the confidence value by dividing the support (number of times Ei and Tj appear together) by the frequency of Ei. Store this confidence value in Y.
    - Add an entry to the list data structure, Count, in the format (Ti, Ej, X, Y).
- Sort the Count list by Y in descending order, with the most probable Tamil translation appearing first.

The predicted translated words for each document, along with their confidence scores, are shown in Figure 10.

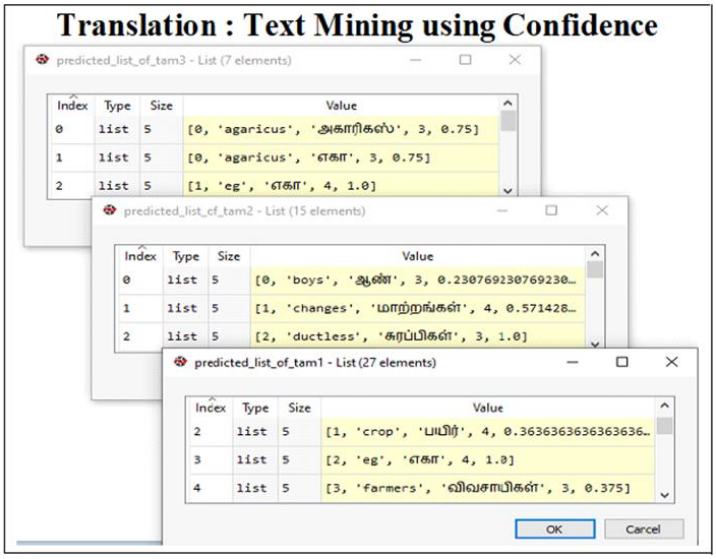

Figure 10: Predicted translated word

4. Experimental Setup

The proposed model was evaluated through a series of experiments. These experiments were conducted on a computer system equipped with an Intel Core i5-7200U processor, running at 2.50 GHz (with a maximum frequency of 2.71 GHz), 8 GB of memory, and a 1 TB hard drive. The system was configured with NetBeans IDE 8.2 (Java 1.8.0_171) and Spyder 5.0.1 (Python 3.7.9, 64-bit) on Windows 10.

4.1. Document-Specific Bilingual Dictionary

The feature extraction algorithm was applied to three documents to generate document-specific bilingual dictionaries. A sample of the generated dictionaries is shown in Figure 11.

4.2. Performance Evaluation

Once the list of chapter-specific source words was identified using the TF-IDF algorithm, the translation prediction algorithm was applied to the corpus to predict the corresponding target language words.

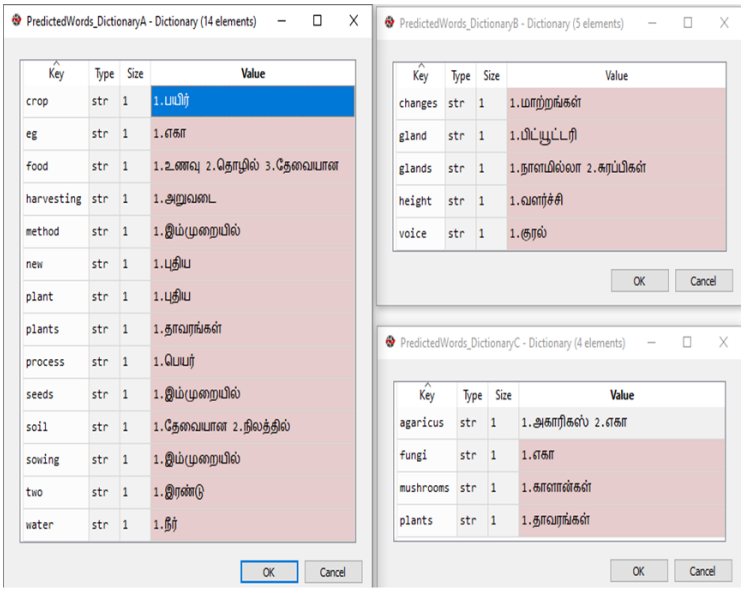

Figure 11: Context-based document-specific bilingual dictionary

The experiments were conducted by varying the minimum number of occurrences of source and target words per sentence line—specifically, 2, 3, and 4 times, as shown in Table 2. The evaluation was performed manually. A predicted translation was considered correct if any word from the predicted list matched the actual translation. For instance, the predicted translations for the word "food" included "unavu", "thozhil" and "thevaiyana". Since "unavu" is the correct translation, the prediction was deemed accurate.

Table 2: Evaluation of the developed system

Minimum no. of times the source and target words occur in the same line	Document	No. of document-specific words using TF-IDF	No. of Predicted words that are Correct	Number of Predicted words that are Wrong	Accuracy (%)
2	Document 1	78	56	22	71.79
	Document 2	40	27	13	67.5
	Document 3	13	12	1	92.31
3	Document 1	17	14	3	82.35
	Document 2	10	7	3	70
	Document 3	4	3	1	75
4	Document 1	5	5	0	100
	Document 2	2	2	0	100
	Document 3	2	2	0	100

Results indicated improved translation accuracy when the source and target words appeared at least four times in the sentence lines, as illustrated in Figure 12.

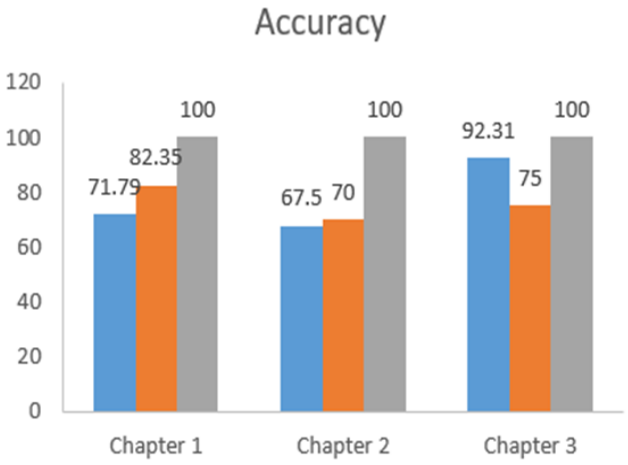

Figure 12: Accuracy

5. Conclusion

The proposed algorithm constructs a context-based, document-specific bilingual dictionary by integrating the existing TF-IDF method with a modified association rule mining confidence approach to predict translation words. The introduction of the Word2Line vector effectively represents both source and target-language words, simplifying translation prediction by calculating confidence based on support. Since the algorithm is language-independent, it is well-suited for use with any low-resource target language. Additionally, it does not rely on large volumes of pre-annotated corpora or contextual information about surrounding words, making it simpler. The method achieved 100% accuracy when the minimum co-occurrence of source and target words per sentence was set to 4, demonstrating its strong potential for practical applications.

Acknowledgment: N/A

Data Availability Statement: The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Funding Statement: This research was conducted and prepared independently by the authors without any form of financial assistance or external support.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest related to the content of this manuscript. All references and citations have been duly acknowledged based on the sources utilised.

Ethics and Consent Statement: The study followed all ethical research standards, and informed consent was obtained from all participants prior to data collection and analysis.

References

1. A. Hautli-Janisz, "Pushpak Bhattacharyya: Machine translation" *Mach. Transl*., vol. 29, no. 3–4, pp. 285–289, 2015.
2. J. Zhang and C. Zong, "Neural machine translation: Challenges, progress and future," *Sci. China Technol. Sci*., vol. 63, no. 10, pp. 2028–2050, 2020.
3. K. Chen, R. Wang, M. Utiyama, and E. Sumita, "Integrating prior translation knowledge into neural machine translation," *IEEE ACM Trans. Audio Speech Lang. Process*., vol. 30, no. 12, pp. 330–339, 2022.
4. Z. Liu, J. Li, and M. Zhu, "Alleviating exposure bias for neural machine translation via contextual augmentation and self-distillation," *IEEE ACM Trans. Audio Speech Lang. Process*., vol. 31, no. 5, pp. 2079–2089, 2023.
5. A. Gupta and D. Yadav, "A novel approach to perform context-based automatic spoken document retrieval of political speeches based on wavelet tree indexing," *Multimed. Tools Appl*., vol. 80, no. 14, pp. 22209–22229, 2021.
6. M. del Mar Sánchez Ramos, "Documentation in specialised contexts: A quasi-experimental corpus-based study in public service interpreting and translation studies," *Transl. Translanguaging Multiling. Contexts*, vol. 8, no. 1, pp. 30–48, 2022.
7. W. Lam, K. Chan, D. Radev, H. Saggion, and S. Teufel, "Context-based generic cross-lingual retrieval of documents and automated summaries," *J. Am. Soc. Inf. Sci. Technol*., vol. 56, no. 2, pp. 129–139, 2005.
8. Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context gates for neural machine translation," *Trans. Assoc. Comput. Linguis*t., vol. 5, no. 3, pp. 87–99, 2017.
9. A. Kumar, A. Pratap, A. K. Singh, and S. Saha, "Addressing domain shift in neural machine translation via reinforcement learning," *Expert Syst. Appl*., vol. 201, no. 9, p. 117039, 2022.
10. M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *arXiv: Computation and Language*, 2017. Available: https://arxiv.org/abs/1705.00440 [Accessed by 02/06/2024].
11. M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized data augmentation for low-resource translation," *arXiv: Computation and Language*, 2019. Available: https://arxiv.org/abs/1906.03785 [Accessed by 11/07/2024].
12. S. M. Lakew, Q. F. Lotito, M. Negri, M. Turchi, and M. Federico, "Improving zero-shot translation of low-resource languages," *arXiv: Computation and Language*, 2018. Available: https://arxiv.org/abs/1811.01389 [Accessed by 30/07/2024].
13. I. J. Unanue, E. Z. Borzeshi, and M. Piccardi, "Regressing word and sentence embeddings for low-resource neural machine translation," *IEEE Trans. Artif. Intell*., vol. 4, no. 3, pp. 450–463, 2023.
14. R. Haque, C.-H. Liu, and A. Way, "Recent advances of low-resource neural machine translation," *Mach. Transl*., vol. 35, no. 4, pp. 451–474, 2021.
15. L. Gong, Y. Li, J. Guo, Z. Yu, and S. Gao, "Enhancing low-resource neural machine translation with syntax-graph guided self-attention," *Knowl. Based Syst*., vol. 246, no. 6, p. 108615, 2022.
16. N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, no. 5, pp. 522–538, 2022.
17. Y. Wu, S. Zhao, Y. Zhang, X. Yuan, and Z. Su, "When pairs meet triplets: Improving low-resource captioning via multi-objective optimization," *ACM Trans. Multimed. Comput. Commun. Appl.,* vol. 18, no. 3, pp. 1–20, 2022.
18. M. Go, J. Kong, and A. Munir, "Linearization weight compression and in-situ hardware-based decompression for attention-based neural machine translation," *IEEE Access*, vol. 11, no. 5, pp. 42751–42763, 2023.
19. W. Xie, M. Ji, M. Zhao, T. Zhou, F. Yang, X. Qian, and T. Hao, "Detecting symptom errors in neural machine translation of patient health information on depressive disorders: developing interpretable Bayesian machine learning classifiers," *Frontiers in Psychiatry*, vol. 12, no. 10, p. 771562, 2021.
20. Q. Chen, "A smaller and better word embedding for neural machine translation," *IEEE Access*, vol. 11, no. 4, pp. 40770–40778, 2023.
21. T. Hirasawa, M. Kaneko, A. Imankulova, and M. Komachi, "Pre-trained word embedding and language model improve multimodal machine translation: A case study in Multi30K," *IEEE Access*, vol. 10, no. 6, pp. 67653–67668, 2022.

22. A. S. D. Prasanna and C. B. C. Latha, “Bi-lingual machine translation approach using long short–term memory model for Asian languages,” *Indian J. Sci. Technol*., vol. 16, no. 18, pp. 1357–1364, 2023.
23. S. Gokila, S. Rajeswari, and S. Deepa, “Tamil-NLP: Roles and impact of machine learning and deep learning with natural language processing for Tamil,” *in 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM),* Chennai, Tamil Nadu, India, 2023.
24. Y. Zhao and H. Liu, “Document-level neural machine translation with recurrent context states,” *IEEE Access*, vol. 11, no. 2, pp. 27519–27526, 2023.
25. R. Thangarajan, A. M. Natarajan, and M. Selvam, “Syllable modeling in continuous speech recognition for Tamil language,” *Int. J. Speech Technol*., vol. 12, no. 1, pp. 47–57, 2009.
26. P. Bafna, D. Pramod, and A. Vaidya, “Document clustering: TF-IDF approach,” *in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),* Chennai, Tamil Nadu, India, 2016.
27. S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, “A novel text mining approach based on TF-IDF and Support Vector Machine for news classification,” *in 2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, Tamil Nadu, India, 2016.
28. A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, “Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach,” *in 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia, 2014.
29. R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *in Proc. 1993 ACM SIGMOD Int. Conf. on Management of Data*, San Jose, California, United States of America, 1993.