International Journal of Multiphysics
Volume 18, No. 2, 2024
ISSN: 1750-9548

Cross-Language Translation Algorithm Based on Word Vector and Syntactic Analysis

Jiali Min*
School of Foreign Languages, Nanchang Institute of Technology Nanchang, 330044, Jiangxi, China;
minjiali2024@163.com

Abstract

Digital technology has created a necessity to protect intellectual uniqueness in translated works. Cross-lingual syntactic connection is critical for measuring the level of similarity between textual pairs produced in different languages for the purpose to detect plagiarism. This study addresses cross-lingual syntactic analysis and plagiarism detection and evaluation using the university student's dataset. The data is utilized for translation and plagiarism identification, and it is transferred into a research model using the Natural Language Process (NLP). The data is then extracted into features using a word vector. The research proposed that the word Embedded Fast Recurrent Network (WE-FRN) is used for syntactic analysis and plagiarism detection. To handle the plag detection issue, many neural network designs were proposed, including regime proposal (plagiarism or independently created) and binary classification (syntactic regression analysis of documents). Experimental results indicated that utilizing WE-FRN with rich syntactic characteristics produced better outcomes than baseline and the loss function of the classification is also analyzed by the regression of the source and suspicious documents.

Keywords: Cross-Language Translation, Natural Language Process, Word Vector, Plagiarism Detection, Syntactic Analysis

1. Introduction

Recent advances in artificial intelligence and machine learning technology were the principle of the excellent development in cross-language translation. The improvement of greater specific and effective translation systems that may translate textual content between numerous languages with previously unheard-of fluency is the result of those breakthroughs. Deep learning methods, such as transformer-based topologies and neural machine translation (NMT), that are able to grasp significant language structures and variations, were enormous in enhancing the quality of translations [1]. Furthermore, cross-language translation systems' effectiveness has been further stronger via the incorporation of enormous bilingual information and the application of techniques such as transfer learning. Moreover, the enhancement of pre-skilled language frameworkslike(Bidirectional Encoder Representations from Transformers BERT) and (Generative Pre-skilled Transformer GPT) has endedpossibility for more contextually sensiblepresentations [2].

Cross-language translation generation develops consequential an expanded style in the previous years, particularly inside the location of English grammar. The combined of (machine learning ML) and AI methods has been a major density inside the refunded of these developments. The usage of NMT patterns is a notable innovation that has significantly proceeded the precision and fluidity of translating from English to other languages. In addition, models collected with BERT and GPT that utilized transformer - based systems, had improved interpretations by utilizing capturing appropriate transformations [3]. Large- scale multilingual datasets have also made it less difficult to train those models, cultivating their ability and potential to offer grammatically correct translation in few linguistic contexts.

* Place the footnote text for the author (if applicable) here.

Such models delivered that closely resemble that language was utilized in natural language because they are accomplished at communicable complex grammatical structures, informal idioms, and linguistic intricacies. Furthermore, greater sophisticated cross-language translation capabilities are offered by transformer-based models as GPT and BERT [4]. These representations implement an incredible work of evaluating the word or segment's information, permitting them to offer analysis that devise applicable for the framework where they are used to grammatically correctness. The presentation of these translation structures has been suggestively improved by using methods such as ML and the availability of significant multilingual data collection.

These models are higher able to understand the subtle differences between languages and generate translations that are greater correct since they have been skilled on great volumes of text facts from many linguistic resources [5].

Plagiarism is using the work of someone else without credit, and it has grown to be more common as the information technology (IT) area has grown. Unauthorized textual content reuse is one of the primary reasons of plagiarism, which is a major concern in each academic and non-instructional context [6]. To evade detection, plagiarism uses of variability of unintelligible strategies, along with copying, pasting and paraphrasing.The size of the reference collection presents additional difficulties for the detection procedure. An expert system that can recognize plagiarism autonomously is required to solve this problem. By identifying copied portions of text and their source, the method gives human specialists the proof of unapproved text reuse so they can make knowledge judgments [7]. To identify plagiarism in cross-lingual texts, techniques include statistical machine translation, character N-gram syntactic alignment, dictionaries, and cross-lingual explicit semantic analysis. For translators, there is yet a considerable obstacle in overcoming this difficulty [8]. Although search engines try to decrease the amount of papers available, students, however, have to go through several pages to locate the information they need. Students gradually work on their search techniques and craft more focused quires by utilizing key phrases [9]. Students believe they entered the search keywords incorrectly if there weren't many result. A solid plan with an adequate number of key phrases is required. On the contrary, ineffective search techniques can result from things, including a lack of time, patience, or trouble finding pertinent material [10]. Translating materials from one language to another without giving due acknowledgment to their initial source is one kind of unintelligible. Other names for translated plagiarism include multilingual and cross-language plagiarism. The "Large Language Models (LLMs)" include the Bard, generative pre-training transformer (GPT)-3.5, Claude, GPT-4, and "Large Language Model Meta Artificial Intelligence (LLaMa-AI)", which are all significant breakthroughs in language dispensation.These representations are intended to evaluate and develop language that resembles the speech spoken by humans. They are frequently used in a variety of the fields because of their powerful capabilities. The most popular LLMs are trained on various language dataset, providing multilingual services to a worldwide user document. This research aims to enhance cross-lingual syntactic analysis and plagiarism detection using university student datasets. It employs a word embedding-driven fast recurrent network (WE-FRN) and syntactic feature analysis to improve the accuracy of identifying translational plagiarism.

Key Contributions

- Unique Cross-Linguistic Syntactic Analysis Technique has been implemented.
- Utilizes word embedding-driven fast recurrent network (WE-FRN) for plagiarism detection.
- Outperforms baseline methods on university student dataset.
- Increases plagiarism detection accuracy and document syntactic regression analysis.

Rest of the study is categorized into 5 portions, portion 2: literature reviews of the plagiarism detection based document translation, portion 3: methodology of the translation method using WE-FRN, portion 4: result of the translation method and portion 5: conclusion of the research.

2. Literature Review

Maqbool et al., [11] stated Machine Learning (ML) algorithms were used for cross-lingual plagiarism detection between Urdu and English texts, with successful results. Hong [12] developed a networked intelligent translation system for cross-language information extraction, with an emphasis on Chinese-to-English translations. Liu et al., [13] proposed a Genetic Algorithm (GA) and Cloud Computing (CC) strategy for machine translation in Low-Resource Languages (LRLs), which dramatically improved translation accuracy. Lan and Huang [14] suggested a hybrid partition-hierarchical cross-language query expansion approach was introduced for clustering analysis, which improved retrieval accuracy by eliminating theme shift and word mismatch. Dinhand Thanh [15] suggested an accurate detection of paraphrase occurrences between English and Vietnamese sentences, the study utilized a fuzzy-based technique and a Siamese recurrent model. Muttumana et al., [16] Deep Learning (DL) clustering was used with a synonym database to improve plagiarism detection in college assignments, resulting in improved academic evaluation integrity. Table 1 illustrates the overview of the related works.

Table 1. Related works

Reference	Algorithm and Optimization Technique	Dataset	Language	Result	Purpose
Maqbool et al., [11]	ML Algorithms of Majority Voting, Stacking, Averaging, Boosting, Bagging	Corpus of 2398 documents with source document (text-Urdu) and suspicious document (text-English) (CLPD-UE-19)	Urdu to English	Not Mentioned	Cross-lingual plagiarism detection
Hong [12]	Internet-Based Intelligent Translation Machine Cross-linguistic collection of data Multilingual word representation training.	Bilingual Chinese-English dataset	Chinese to English	Excellent data retention rate, correctness, security, and system adaptability.	Cross-language information extraction and translation,
Liu et al., [13]	GA and CC	Not Mentioned	LRLs	GA-optimized LRL-oriented machine translation significantly improves translation accuracy and CS is 94%.	To improve the quality and efficiency of machine translation.
Lan and Huang [14]	Hybrid Partition-Hierarchical Cross-Language Query Expansion algorithm for clustering analysis (HPH-CLQE)	NTCIR-5 CLIR Dataset	Translation between multiple languages	Effectively eliminates theme shift and word mismatch, increasing overall retrieval accuracy.	To reduce theme shift and word mismatch in the translation.
Dinh and Thanh [15]	Fuzzy-based method and Siamese recurrent	Not Mentioned	English and Vietnam	Accuracy: 87.4%	To identify paraphrasing instances

	model.		ese.		between English and Vietnamese phrases.
Muttumana et al., [16]	DL Clustering technique	Synonym database containing 100,000 words	Multiple languages	Uses deep features and clustering for accurate plagiarism identification. Improved academic assessment integrity.	To enhance plagiarism detection in college assignments

3. Methodology

The work explores cross-language contextual similarities and identification of the plagiarism model (Figure 1) in the text document. Study strives to determine whether a pair of texts is plagiarized and the degree of cross-language semantic text analogy among the two, such as the original language file and the language of the translation file. The method consists of two tasks: data text document (t, t') from two distinct languages (L and L'), and cross-language conceptual parallels in text challenge to assess the degree of similarities in syntactic and contextual terms between t and t'. The counterfeit identification task classifies t and t' as either plagiarized or not.

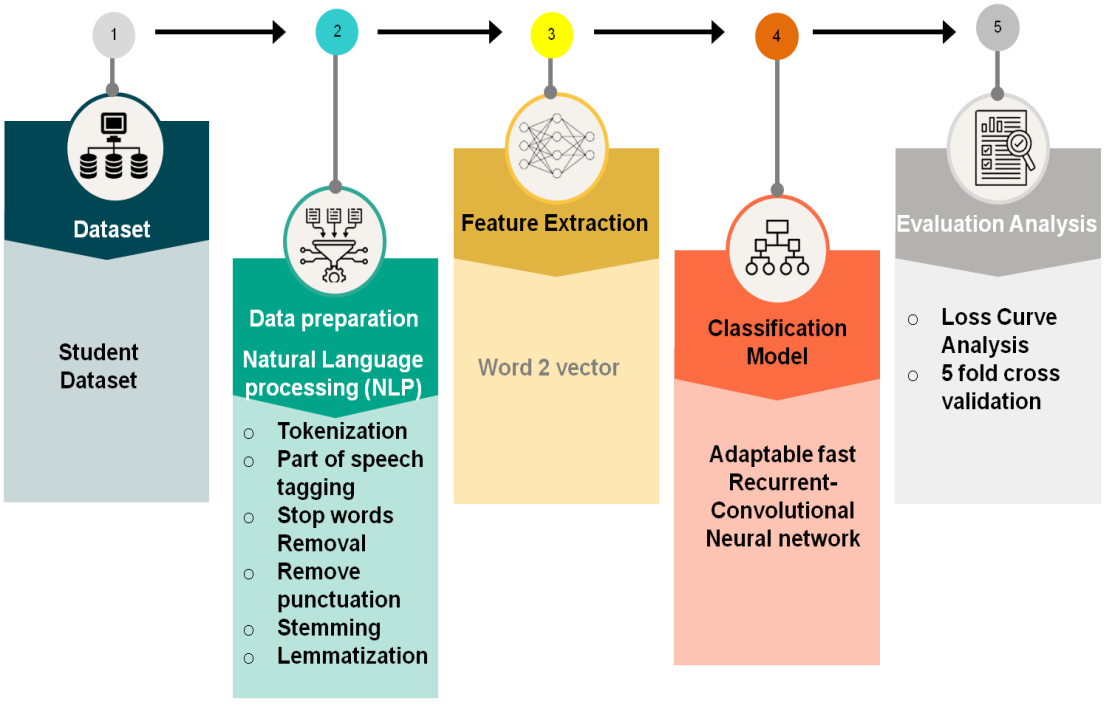

Figure 1. Research Flow Model

3.1 Dataset

A university student dataset was generated to examine the accuracy of translations from English to other languages such as Chinese Japanese and Korean. The resultant benchmark cross-language English-Chinese dataset was developed by a study of 500 students aged 18 to 24 who translated English texts into Chinese Japanese and Korean. Each student translated three paragraphs from the Physical Electronics (120) textbook into their own language. The data is separated into 80% (100 textbook documents) for training and 20% (20 textbook documents) for testing. The study's goal was to guarantee that translations were free of presumptions.

3.2 Data Preparation Using Natural Language Process (NLP)

NLP leverages word vectors are used to represent meanings across languages and syntactic evaluation to understand grammatical structures. This combination permits algorithms to appropriately translate textual content by mapping semantic similarities and preserving syntactic coherence, enhancing translation quality between special languages.

Text cleansing and preprocessing were carried out using the Natural Language Process (NLP), which included punctuation removal, tokenization, parts of speech tagging, stop words removal, regular expression, stemming and lemmatization. Natural Language Tool Kit (NLTK) permits tagging and lemmatization of English text. More components were employed for English and Chinese texts, such as the cardinal English processor for word lemmatization (Green and Manning) and the cardinal portions of speech tagger.Tokenization is a method of breaking down phrases into words, characters, and punctuation to filter out unwanted terms in processing. Parts of speech tagging categorizes words into fundamental grammatical classes like nouns, pronouns, verbs, adjectives, and prepositions. Negation words like "No" and "cannot" are crucial for determining sentence context and purpose. Regular expressions (Regex) are used for pattern search, while stemming is a violent word-shortening strategy that reduces a term to its basic root. Lemmatization is a text preparation step that eliminates or changes a word's suffix to return it to its base, resulting in intelligible words. The sequence of these procedures is critical to avoid incorrect results due to improper removal of stop words. Figure 2 depicts the data preparation method of the research.

Language	Segmented, Part of Speech tagged Text
Chinese	<s>如果_CS 您_PN 在_P 新加坡_NR 只_AD 能_VV 前往_VV 一_CD 间_M 俱乐部_NN ，_PU 祖卡_NN 酒吧_NN 必然_AD 是_VC 您_PN 的_DEG 不二_JJ 选择_NN 。_PU</s>
English	<s>If_IN you_PRP only_RB have_VBP time_NN for_IN one_CD club_NN in_IN Singapore_NN ,_, then_RB it_PRP simply_RB has_VBZ to_TO be_VB zouk_JJ ._.</s>
Indonesian	<s>Jika_nn Anda_nn hanya_rb memiliki_vbt waktu_nnc untuk_in satu_cdp klub_nnc di_in Singapura_nn ,_, pergilah_nn ke_in Zouk_nn ,_, mungkin_rb satu-satunya_jj klub_nnc malam_nn di_in Singapura_nn yang_sc bereputasi_nn internasional_jj ._.</s>
Japanese	<s>シンガポール_名詞-固有名詞-地域-国 で_助詞-格助詞-一般 一つ_名詞-一般 の_助詞-連体化 クラブ_名詞-一般 に_助詞-格助詞-一般 しか_助詞-係助詞 行く_動詞-自立 時間_名詞-副詞可能 が_助詞-格助詞-一般 なかっ_形容詞-自立 た_助動詞 と_助詞-格助詞-引用 し_動詞-自立 たら_助動詞 、_記号-読点 間違い_名詞-ナイ形容詞語幹 なく_助動詞 、_記号-読点 この_連体詞 ズーク_名詞-一般 に_助詞-格助詞-一般 行く_動詞-自立 べき_助動詞 です_助動詞 。_記号-句点 </s>
Korean	<s>싱가포르_NNP 에서_JKB 클럽_NNP 한_NNP 군데_NNB 밖에_JX 가_VV 르_ETM 시간_NNG 이_JKS 없_VA 다면_EC ,_SP Zouk_SL 를_JKO 선택_NNG 하_XSV 시_EP 어요_EF ._SF</s>
Vietnamese	<s>Nếu_C bạn_N chỉ_R có_V thời_gian_N ghé_V thăm_V một_M câu_lạc_bộ_N ở_E Singapore_Np ,_, hãy_R đến_V Zouk_Np ._.</s>

Figure 2. Data Preparation Techniques

Note:"*DT: Determiner. NN: noun, single or plural. JJ: an adjective, WDT: Wh-determiner, TO: "to"; NNS: noun, plural,VB: Verb; basic form, VBZ: Verb, third person singular present. IN: A preposition or subordinating conjunction.CC: Coordinating conjunction. VBG: Verb, gerund, or present participle. NNP: proper noun, singular. Characters (\w), whitespace (\s), commas (\,) and full stop (\.)"*

3.3 Feature extraction and classification using Word Embedding-Driven Fast Recurrent Network (WE-FRN)

To create a Word Embedding-Driven Fast Recurrent Network (WE-FRN) system for detecting cross-language plagiarism by comparing syntactic and contextual similarities between text pieces in two different languages to determine the level of syntactic and contextual similarity between the original and translated writings, as well as whether they were plagiarized or generated independently.

3.3.1 Word 2 Vector (W2V)

W2V enhanced cross-language translation are used by generating phrase vectors those detention semantic meanings will be considered. Combined with syntactic assessment, these vectors contribution align and translate phrases by identifying similar contexts and structures, improving translation accuracy and linguistic rationality.

A sophisticated natural language processing (NLP) model called W2V builds word vectors by gathering syntactic and grammatical data from models known as skip-gram and (continuous bag of words CBOW), as shown in Figure 3.

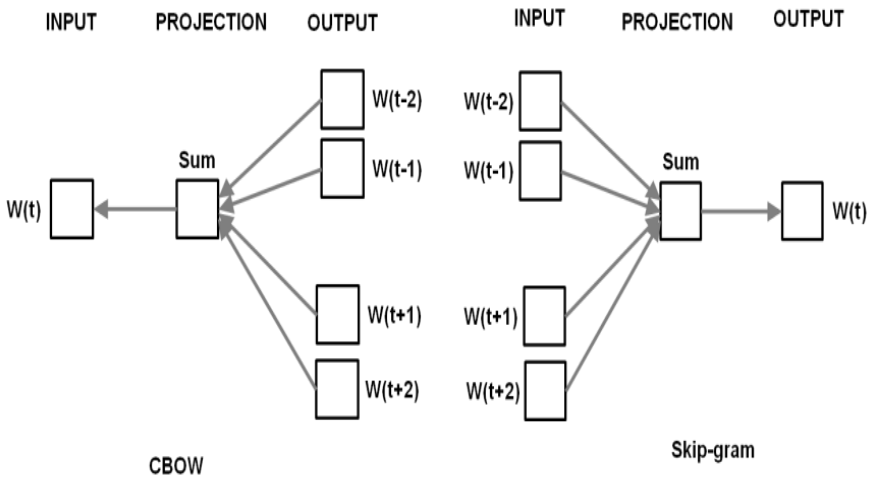

Figure 3: Word 2 Vector Process

For outcome connections in a quantity and paralleling tokens, this feature of the presentation technique utilities efficiently. It is suggested for training on large datasets and workings well in a variety of NLP applications. Context, hidden, and output layers are used by CBOW to identify words constructed on framework, while skip-gram forecasts the framework of a word. Negative sampling and hierarchical softmax, both based on Huffman trees, are used for performing optimization.

3.3.2 Adaptable Fast Recurrent Convolutional Neural Network

The AF-RCNN enhances cross-language translation by integrating word vectors and syntactic analysis. It captures complex linguistic styles, permitting efficient and accurate translation throughout languages by leveraging both semantic and syntactic information for improved contextual information. The AF-RCNN model (see Figure 4) is an excellent network model for target recognition algorithms, providing high accuracy and real-time performance.

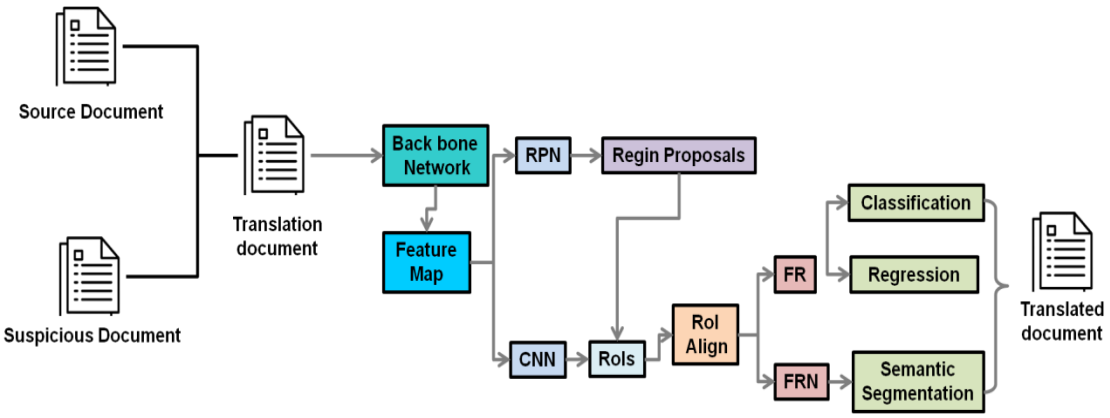

Figure 4. FRN Model

It was used to create a multi-task syntactic segmentation algorithm called AF-RCNN, which involves information fusion as opposed to earlier techniques. The approach combines source and suspicious documents, recognizes plagiarism, and uses a fully convolutional neural network for syntactic segmentation. The fusion of the basis file and the suspect file reduces the effect of lighting elements while increasing the detection speed of the translational target. The fused document has both source and suspicious document information, which improves its feature expression ability over techniques that just employ source and suspicious document training. To increase target identification accuracy, a novel approach to screening target text document frames was developed that takes into account the overlap degree of the document plag region as well as the number of papers around it. The sequence-to-sequence approach analyzes the sequence distribution of source and suspect documents to improve feature extraction and document training.

$$C_i = \sum a_{ij} h_j \quad (1)$$

Here a_{ij} is the word of the source and suspicious documents h_j height of the suspicious document, $C_i$word sequential count of the source document. The sequence-to-sequence method aids in understanding the word sequence distribution of semantic, syntactic, and contextual channels. The sequence distribution across three channels is comparable, showing significant word characteristics. However, a channel in a sequential document lacks distinguishing qualities, prompting its replacement with a relevant document from the channel documents. This fusion produces the source document, which combines horizontal parallax, height to the ground, and word count sequence information while efficiently integrating color and depth elements. The backbone network is a Convolutional Neural Network (CNN), which is used as a tool for extracting features to improve mapping identification capabilities. ResNet50 or ResNet101 is used to introduce residual learning with connections skipped. ResNet improves similarity identity mapping capabilities of the cross-language translation by inventing the skip connection framework, which enables network extension and increased performance. The underlying mapping uses a superimposed nonlinear layer to satisfy another mapping, resulting in the original feature that is mapped to$G(w_{input})$. The process is written as:

$$z_{output} = G\left(w_{input}, \{\omega_j\}\right) + w_{input} \quad (2)$$

Where w_{input} and z_{output} represent input and output, respectively, and $G\left(W_{input}, \{\omega_j\}\right)$ is a learned residual mapping. The Region Proposal Network (RPN) scans areas and outputs anchor boxes using sliding windows. The RPN shares features with the backbone network and generates around 17,000 students regions. The Non-Maximum Suppression (NMS) algorithm refines bounding boxes. The introduction of Region of Interest (RoI) Align based on fully convolutional improved the syntactic segmentation branch by removing quantization operations and addressing misalignment issues in RoI pooling, ensuring full alignment of the original and translated output documents while avoiding plagiarism errors.The Intersection-over-Union (IoU) approach is

used to discover targets by calculating the overlap rate between the translated content and the similarity model. The AF-RCNN algorithm calculates the overlap between these source documents and authentic artificially labeled plagiarismof the source file frames using a 0.5 threshold.

$$U_{\substack{(o,r) \\ o=[1,m-\alpha] \\ r=[o+1,n]}} = \frac{inter_{(o,r)}}{file_{(o)}+file_{(r)}-inter_{(o,r)}} \quad (3)$$

Where o and$r$ represent the originality and plagiarism rating of the content. The updated NMS method treats each frame as a five-tuple and calculates the overlap ratio$U_{(o,r)}$ between frames. Implementing this technique can minimize the number of frames to 2000 while swiftly reducing plagiarism, resulting in successful syntactic detection and classification in complicated sentences. This method effectively addresses regime proposal (plagiarism or independently created) and binary classification (syntactic regression analysis of documents) in source document scenes by combining semantic, syntactic, and contextual channels, enriching feature extraction, and enhancing training and experimental results.

The foundation of most computational operations is the requirement to transform obscure symbols into spatial vectors that the machine can understand. To begin, the textual data must have a mathematical and structured representation; for example, sequence information can be fully represented using distributed vectors. When dealing with text in English, words are often employed as processing units before being transformed into word vectors. This is in contrast to the flexible use of the English representation, where a word often contains numerous meanings, each of which may indicate a distinct meaning.

$$a_j = \int (U^{u-1} \times C^j) \quad (4)$$

BILSTM

The first stage in implementing BILSTM is to use the "forgetting gate's" Sigmoid function layer to calculate whether or not a piece of data should be kept or destroyed.

$$a_j = tanh\frac{\sqrt{z_{d-1}+y_d}}{U_{V-1}} - p_v \quad (5)$$

CRF

The CRF layer refines output by modeling dependencies between translated words, ensuring syntactic and contextual coherence. It enhances the translation through leveraging phrase vectors and syntactic evaluation to produce grammatically and contextually correct sequences.

4. Result of the Translation

To develop a machine learning model, install translation libraries and frameworks, create bilingual corpora, define training parameters, and assess the model using Bilingual Evaluation Understudy (BLEU) results. Install a high-performance Graphics Processing Unit (GPU) with the appropriate drivers, as well as high-capacity data storage and model checkpoints, to achieve rapid read/write speeds. The loss function computes the distinction between the model's predicted value and the training data. A lower number suggests higher model robustness, whereas a larger difference indicates stronger model robustness. When the overall loss of a multi-task syntactic sequence based on source document training reaches 15 epochs, plagiarism detection occurs, and both classification and syntactic sequential loss converge. The model training impact is positive, as seen in Figure 5. However, when the number of iterations exceeds 16w, the regression of the source translation similarity suspicious index loss function value falls to around 0.002, suggesting fundamental convergence. The research proposed a WE-FRN classification, detection, and segmentation task, dividing the loss function into three parts: classification loss, regression loss, and syntactic segmentation branch loss function. The loss function is defined as:

$$L_{total} = L_{cls} + L_{box} + L_{seg} \quad (6)$$

Where $L_{cls}$and L_{box} represent classification loss (Figure 5(a)) and regression loss (Figure 5(b)), respectively. The segmentation branch output dimension is km^2, representing n binary syntactic sequence loss (Figure 5(c)) masks with resolution of$M \times M$.

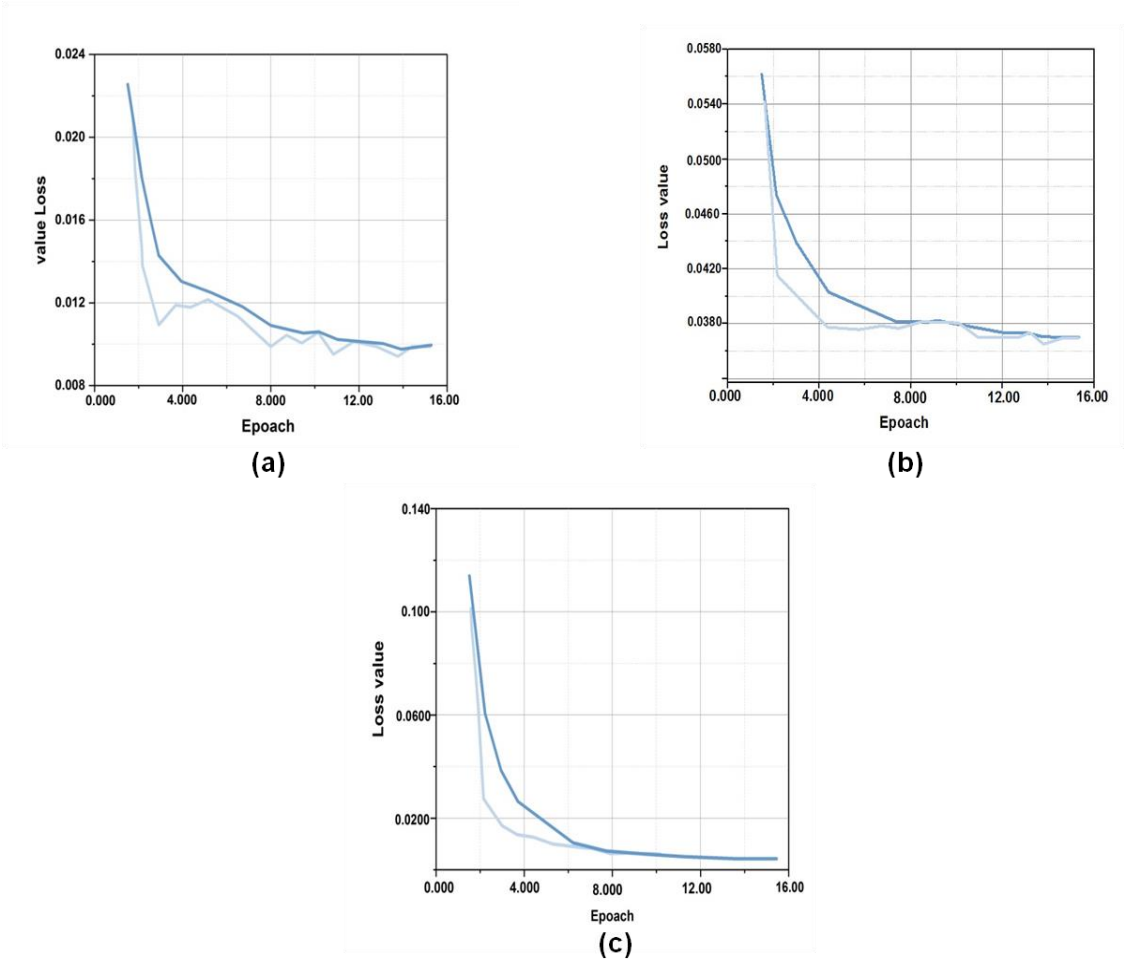

Figure 5: Loss Function (a) Classification loss, (b) Regression loss and (c)Syntactic sequence loss

Table 2 illustrates the accuracy of the W2V+IFRCNN model for cross-language translation detection across four language combinations (English, Chinese, Japanese, Korean) using a 5-fold cross-validation with 100 training files separated into 5 folds and each fold with 20 files.

Table 2. Accuracy of the Translation WE-FRN Model

Training Dataset (100 Files)	Accuracy of WE-FRN (%)			
	English	Chinese	Japanese	Korean
Fold 1	92.06	87.48	86.74	89.63
Fold 2	90.06	89.84	88.63	86.43

Fold 3	92.08	90.00	87.89	86.48
Fold 4	90.09	87.96	88.21	86.03
Fold 5	90.22	89.08	86.03	87.74
Average	**90.9**	**88.87**	**87.49**	**87.26**

Table 3 illustrates the translation knowledge gained by 500 students on the topic of electronics.Data were gathered via questionnaires and a web link. After cleaning, 80,932 passage pairs were extracted from 100 documents. Approximately 71.2% of the students had previous expertise in translating English into other languages such as Chinese, Japanese and Korean.

Table 3. Translation Experience Outcomes

Experience of the Translation	Number of Students	Field of Study
No	72	Electronics
English to Chinese	78	Electronics
English to Japanese	58	Electronics
English to Korean	42	Electronics
Total Students	250	
Total Students with Experience in Translation	178 (71.2%)	

5. Conclusion

The proposed WE-FRN was used to detect the plagiarism of the cross-language translation and extract the syntactic and contextual elements from texts in two distinct languages. The features include topic similarity, punctuation, tokenization, parts of speech tagging, regular expression, stemming, lemmatization, and stop words. WE-FRN novel approach efficiently distinguishes between plagiarized and independently produced cross-lingual patterns. It can also do syntactic regression analysis on documents. Combining syntactic features can be valuable, despite the consequences of whether the texts have been composed in two distinct languages. Future research will examine more syntactic aspects using a variety of language resources and document lengths, as well as sophisticated deep learning techniques such as recurrent neural networks.

Reference

1. Roostaee, M., Sadreddini, M.H. and Fakhrahmad, S.M., 2020. An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes. Information Processing & Management, 57(2), p.102150. https://doi.org/10.1016/j.ipm.2019.102150
2. Nagy, I., Rácz, A. and Vincze, V., 2020. Detecting light verb constructions across languages. Natural Language Engineering, 26(3), pp.319-348. https://doi.org/10.1017/S1351324919000330

3. Aljuaid, H., 2020. Cross-Language Plagiarism Detection using Word Embedding and Inverse Document Frequency (IDF). International Journal of Advanced Computer Science and Applications, 11(2).
4. Alotaibi, N. and Joy, M., 2020. Using Sentence Embedding for Cross-Language Plagiarism Detection. In Artificial Intelligence XXXVII: 40th SGAI International Conference on Artificial Intelligence, AI 2020, Cambridge, UK, December 15–17, 2020, Proceedings 40 (pp. 373-379). Springer International Publishing. https://doi.org/10.1007/978-3-030-63799-6_28
5. Esmailpour, R., Ebrahimy, S., Fakhrahmad, S.M., Mohammadi, M. and Abbaspour, J., 2020. Developing an effective scheme for translation and expansion of Persian user queries. Digital Scholarship in the Humanities, 35(3), pp.493-506. https://doi.org/10.1093/llc/fqz041
6. Alzahrani, S. and Aljuaid, H., 2022. Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases. Journal of King Saud University-Computer and Information Sciences, 34(4), pp.1110-1123.https://doi.org/10.1016/j.jksuci.2020.04.009
7. Li, J., Liu, Y., Liu, C., Shi, L., Ren, X., Zheng, Y., Liu, Y. and Xue, Y., 2024. A Cross-Language Investigation into Jailbreak Attacks in Large Language Models. arXiv preprint arXiv:2401.16765.https://doi.org/10.48550/arXiv.2401.16765
8. Son, J. and Kim, B., 2023. Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. Information, 14(10), p.574. https://www.mdpi.com/2078-2489/14/10/574#
9. Liu, Y., Lin, J. and Cleland-Huang, J. 2020. Traceability Support for Multi-Lingual Software Projects. arXiv (Cornell University), 5(7).https://doi.org/10.1145/3379597.3387440.
10. Roostaee, M., Fakhrahmad, S.M. and Sadreddini, M.H. 2020. Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection. Expert Systems with Applications, 160(5), p.113718.https://doi.org/10.1016/j.eswa.2020.113718.
11. Maqbool, M.S., Hanif, I., Iqbal, S., Basit, A. and Shabbir, A., 2023. Optimized Feature Extraction and Cross-Lingual Text Reuse Detection using Ensemble Machine Learning Models. Journal of Computing & Biomedical Informatics, 5(01), pp.26-40.
12. Hong, L., 2021, July. Design of Networked Intelligent Translation System Based on Machine Learning Algorithm. In Journal of Physics: Conference Series (Vol. 1982, No. 1, p. 012126). IOP Publishing. https://doi.org/10.1088/1742-6596/1982/1/012126
13. Liu, X., Chen, J., Qi, D. and Zhang, T., 2024. Exploration of low-resource language-oriented machine translation system of genetic algorithm-optimized hyper-task network under cloud platform technology. The Journal of Supercomputing, 80(3), pp.3310-3333. https://doi.org/10.1007/s11227-023-05604-6
14. Lan, H. and Huang, J., 2020. The Cross-Language Query Expansion Algorithm Based on Hybrid Clustering. In The 8th International Conference on Computer Engineering and Networks (CENet2018) (pp. 279-286). Springer International Publishing.https://doi.org/10.1007/978-3-030-14680-1_31
15. Dinh, D. and Le Thanh, N., 2022. English–Vietnamese cross-language paraphrase identification using hybrid feature classes. Journal of Heuristics, 28(2), pp.193-209. https://doi.org/10.1007/s10732-019-09411-2
16. Muttumana, A.V., Goel, H., Teotia, Y. and Bhardwaj, P., 2021. Plagiarism Detection Using Deep Based Feature Combined with SynmDict. In Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020 (pp. 45-52). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_5