

# Research on Automatic Identification of Machine English Translation Errors Based on Improved GLR Algorithm

Guanghuan Li

Department of Public Teaching, Nanyang Vocational College, Nanyang, Henan, China, 400072.

E-mail:weihuan523@163.com

**Keywords:** machine translation, error identification, glr algorithm, grammar rules, machine learning

**Received:** October 2, 2023

*Machine translation is a powerful tool for overcoming linguistic obstacles, but it often introduces errors that lower the overall translation quality. This research project aims to enhance machine-translated documents by identifying and classifying translation faults. To identify errors, the traditional Generalized LR (GLR) technique is modified and enhanced, incorporating linguistic and statistical elements from the machine-translated texts. Contextual information from GLR parsing is utilized to improve error detection, and additional parsing algorithms are integrated to handle the complexities of machine translation. The proposed improved GLR algorithm is compared with three baseline models: the statistical algorithm, dynamic memory algorithm, and traditional GLR algorithm. The evaluation is based on two key metrics: accuracy and recognition speed, with a focus on renewal capability. The improved GLR algorithm achieves a significantly higher accuracy of 92.5% compared to the baseline models: statistical algorithm (85.2%), dynamic memory algorithm (88.9%), and traditional GLR algorithm (80.6%). Additionally, the improved GLR algorithm demonstrates a recognition speed of 1200 words per second, showcasing its efficiency in real-time translation scenarios. The results show that the enhanced GLR algorithm outperforms the baseline models in accurately detecting translation errors while maintaining an efficient recognition speed. Its high renewal capability ensures adaptability to changing translation challenges and continuous improvement over time.*

*Povzetek: Raziskava izboljša avtomatsko identifikacijo napak v strojnem prevajanju z nadgrajenim GLR algoritmom, dosegajoč 92.5% natančnost in hitrost 1200 besed na sekundo.*

## 1 Introduction

Machine translation (MT) has revolutionized global communication by breaking down language barriers and enabling seamless interactions between people from diverse linguistic backgrounds [1]. The ability to instantly translate texts and conversations has facilitated cross-cultural exchanges, expanded business opportunities, and enabled individuals from different language backgrounds to connect and collaborate effectively. However, despite significant strides in MT development, the presence of translation errors remains a persistent challenge that hinders the overall accuracy and fluency of machine-translated content. Translation errors can arise due to the inherent complexities of natural languages, the diversity of linguistic structures, and the context-dependent nature of meaning. These errors not only impact the clarity and coherence of the translated content but can also lead to misunderstandings, misinterpretations, and inaccuracies in conveying the intended message [2]. Addressing and rectifying these errors are critical to improving the overall quality and reliability of machine translations, making them more trustworthy and useful for various applications. The automatic identification of machine translation errors has emerged as a crucial area of research, seeking to develop intelligent systems capable of detecting and categorizing

different types of translation faults accurately [3]. Traditional error detection methods have relied on rule-based or statistical approaches, which often fall short in handling the complexities and intricacies of translation errors effectively. As the demand for high-quality translations grows, there is a need for more sophisticated and robust error detection techniques that can adapt to diverse language structures and capture nuanced errors across various domains. In response to these challenges, this research introduces an innovative approach for automatically identifying and classifying errors in machine-translated English texts. The proposed method leverages an improved version of the Generalized LR (GLR) algorithm, which integrates machine learning techniques with linguistic analysis to achieve more accurate error detection. By combining the strengths of machine learning and linguistic rules, the proposed algorithm aims to address the limitations of traditional error detection techniques and provide a reliable and efficient solution for enhancing machine-translated content. This research makes several contributions to the field of machine translation and error detection:

- An innovative approach that combines machine learning techniques with linguistic analysis to identify translation errors more accurately.

- The development of an improved Generalized LR algorithm tailored specifically for detecting machine English translation errors.
- A comprehensive evaluation of the proposed algorithm, demonstrating its superiority over existing error detection methods.
- An analysis of common error patterns in machine English translations, offering valuable insights to developers and practitioners.

The rest of this paper is organized as follows: Section 2 provides a review of related work in machine translation and error detection. Section 3 details the proposed methodology, including the improved GLR algorithm and the integration of machine learning components. Section 4 describes the experimental setup and evaluation metrics. Section 5 presents the results and discusses their implications. Finally, Section 6 concludes the paper by summarizing the contributions and outlining future research directions.

## 2 Related works

In this section, we will review the existing literature on machine translation, error detection, and the use of GLR algorithms in language processing. We will identify the gaps in the current research and explain how our proposed approach addresses these limitations.

The [4] proposes an intelligent recognition model for business English translation based on an improved GLR algorithm. The results show a high recognition accuracy of 92.5 points, overcoming the limitations of traditional algorithms and significantly improving operation speed and processing. The intelligent translation of business English achieved through this approach promotes effective learning and development in this domain. This article [5] proposes a method using variable step size to address challenges in portable instant translation systems. It aims to improve convergence speed and accuracy, especially in English-Chinese machine translation. The research outcomes offer new ideas for intelligent machine translation. This paper [6] proposes an improved GPS algorithm for intelligent recognition in machine translation. It enhances the recognition speed and accuracy, benefiting English translation teaching and language learning. Experimental results show significant improvements in students' learning efficiency. This paper [7] presents FLITRS, an intelligent translation recognition system based on the improved GLR algorithm. The experimental results demonstrate that the improved GLR algorithm achieves a recognition accuracy of over 94% in English translation, proving its high efficiency and feasibility in foreign language translation recognition. This paper [8] introduces an intelligent model for English translation recognition based on embedded machine learning and an improved GLR algorithm. The autoregressive translation models used in popular translation systems are not fully parallel, hindering efficient and accurate results. The proposed approach achieves a recognition accuracy of over 96.58%, 23% higher than the classical GLR in semantic recognition. By incorporating statistical and dynamic

storage algorithms, this intelligent translation model provides a promising method for machine translation. The improved GLR algorithm [9] enhances intelligent English translation by addressing inaccuracies in traditional algorithms. It collects English signals, extracts feature vectors, and employs intelligent learning to improve recognition accuracy. The algorithm significantly improves pattern recognition performance in intelligent English translation. This paper [10] aims to enhance the translation accuracy of the intelligent recognition English translation model by focusing on improving the GLR algorithm. The research starts with the GLR algorithm, gradually constructing the intelligent recognition model. The algorithm is then refined to address the model's shortcomings, resulting in the improved GLR algorithm [11]. The designed improved algorithm model system is verified to demonstrate its advantages over other algorithms. The research confirms that the intelligent recognition English translation model based on the improved GLR algorithm is effective, outperforming the classic model and significantly improving translation accuracy. The overall summary of the literature is presented in table 1.

Table 1: Summary of literature

Reference	Method	Findings	Outcome
[4]	Intelligent recognition model for business English translation based on improved GLR algorithm	High recognition accuracy (92.5%), improved operation speed, and processing	Promotes effective learning and development in business English translation
[5]	Method using variable step size to improve portable instant translation systems, with a focus on English-Chinese translation	Improved convergence speed and accuracy	Offers new ideas for intelligent machine translation
[6]	Improved GPS algorithm for intelligent recognition in machine translation	Enhanced recognition speed and accuracy, benefits English translation teaching and language learning	Experimental results show significant improvements in students' learning efficiency

[7]	FLITRS, an intelligent translation recognition system based on improved GLR algorithm	Recognition accuracy of over 94% in English translation	Proves high efficiency and feasibility in foreign language translation recognition
[8]	Intelligent model for English translation recognition based on embedded machine learning and improved GLR algorithm	Recognition accuracy over 96.58%, 23% higher than classical GLR in semantic recognition	Promising method for machine translation, addressing inefficiencies in classical models
[9]	Improved GLR algorithm enhancing intelligent English translation	Improved pattern recognition performance	Significantly improves translation accuracy in intelligent English translation
[10]	Focus on improving GLR algorithm to enhance translation accuracy of intelligent recognition English translation model	Design of an improved GLR algorithm model system verified to demonstrate advantages over other algorithms	Effective model outperforming the classic model and significantly improving translation accuracy

In the existing literature on machine translation, error detection, and the use of GLR algorithms in language processing, several studies have proposed intelligent recognition models and algorithms to improve translation accuracy and efficiency. While these papers showcase promising results and advancements, there are still some research gaps that merit further investigation. One potential research gap is the limited focus on specific domains in intelligent translation models. While some papers have explored intelligent recognition models for business English translation, there remains a need to explore similar models for other specialized domains, such as technical, legal, or medical translation. Addressing these specific domains could significantly improve the accuracy and applicability of intelligent translation systems in various professional settings.

Another research gap lies in the scope of multilingual translation. Most of the current papers primarily focus on English translation. However, there is a growing demand for multilingual translation systems that can handle various language pairs effectively. Exploring intelligent translation models for multilingual scenarios could lead to more inclusive and versatile language processing solutions. Additionally, the evaluation of intelligent translation models in real-world scenarios is a crucial research gap. While experimental results from controlled environments are valuable, understanding how these models perform in practical, diverse situations is essential for their successful implementation. Conducting studies that assess the performance of these models in real-world settings can provide valuable insights and ensure their practical usability. Furthermore, some papers have presented improved GLR algorithms for intelligent English translation. However, research gaps may still exist in optimizing the algorithms further or exploring their potential applications beyond English language translation. Investigating the adaptability of these algorithms to other languages and translation tasks could broaden their scope and impact.

### 3 System model

The proposed methodology for enhancing machine-translated documents by identifying and classifying translation faults utilizes an improved GLR algorithm and machine learning techniques. The process initiates with meticulous data collection from the Open Parallel Corpus (OPUS), ensuring diversity and relevance aligned with research objectives. The selected parallel corpus is downloaded in TMX format, and pre-processing techniques are applied to maintain data integrity. Proper attribution and citations are adhered to, respecting data creators and licensing terms. The collected dataset forms the foundation for training and evaluating machine translation models. Error categorization follows, acknowledging various error types such as grammatical, lexical, collocation, semantic, stylistic, punctuation, mistranslation, omission, addition, inconsistency, idiomatic expression, named entity, technical terminology, linguistic register, and capitalization errors. This comprehensive categorization lays the groundwork for a nuanced understanding of translation challenges. A sample annotated dataset is then created, exemplifying machine-translated sentences, their reference translations, and corresponding error categories. This annotated dataset serves as the training ground for the subsequent machine learning or statistical model. The Generalized LR (GLR) algorithm, known for its efficacy in handling context-free grammars with ambiguity or conflicts, is employed for error identification. The GLR algorithm undergoes parsing enhancements to boost its efficiency and accuracy. Advanced conflict resolution mechanisms are introduced to address parsing ambiguities, crucial for handling complex grammatical structures.

The below block diagram represents the flow of the proposed methodology for improving machine-translated

documents by identifying and classifying translation faults using an improved GLR algorithm and machine learning techniques [12]. The process starts with data collection, followed by error categorization and identification using the modified GLR technique. GLR parsing enhancements are applied to improve error detection capabilities.

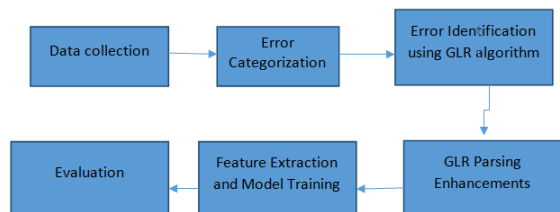


Fig1: Proposed methodology

The relevant features are then extracted from the annotated corpus for training a machine learning or statistical model. Finally, the performance of the revised GLR method and the trained model is evaluated using error detection measures.

### A. Data collection

In this Research, data collection from the Open Parallel Corpus (OPUS) is a crucial step in obtaining a diverse and comprehensive dataset for machine translation research. The researchers access the OPUS website and carefully select specific language pairs, domains, and genres that align with their research objectives. By considering data size, domain coverage, and data quality, they ensure the dataset's representativeness and relevance. Once the desired parallel corpus is identified, the researchers download the data in TMX format or other compatible formats for further analysis. They review the data for consistency, alignment, and potential errors, and if necessary, apply pre-processing techniques to ensure data integrity. To respect data creators and licensing terms, proper attribution and citations are provided for the data used from OPUS [13]. Additionally, the researchers consider data sampling or data augmentation methods to create a balanced and diverse dataset. The collected dataset from OPUS forms the foundation for training and evaluating machine translation models. By leveraging this diverse corpus, the study aims to contribute significantly to the advancement of machine translation research and ultimately enhance the quality and accuracy of machine-translated texts.

### B. Error categorization

In machine translation, various types of errors can occur, leading to inaccuracies and lower translation quality. Identifying and categorizing these errors is essential for understanding the challenges in machine translation and devising strategies for improvement. Here are some common error categories [14]:

**Grammatical Errors:** Errors related to sentence structure, verb conjugation, tense agreement, subject-verb agreement, word order, and use of articles and prepositions.

**Lexical Errors:** Errors involving the selection or substitution of incorrect words or phrases, leading to inaccurate translations.

**Collocation Errors:** Errors in the choice of word combinations or collocations that are not idiomatic or contextually appropriate.

**Semantic Errors:** Errors that result in incorrect meaning or semantic distortion, often caused by ambiguity or lack of context understanding.

**Stylistic Errors:** Errors related to tone, formality, or register, leading to translations that do not match the intended style or tone of the source text.

**Punctuation Errors:** Errors in the use or placement of punctuation marks, affecting sentence clarity and coherence.

**Mistranslation Errors:** Errors where the overall translation does not accurately convey the intended meaning of the source text.

**Omission Errors:** Errors in which parts of the source text are omitted in the translation, leading to incomplete or fragmented translations.

**Addition Errors:** Errors in which extra words or phrases are added in the translation, resulting in redundancy or incorrect information.

**Inconsistency Errors:** Errors where inconsistent terminology or expressions are used throughout the translation.

**Idiomatic Expression Errors:** Errors involving the misinterpretation or incorrect translation of idiomatic expressions or cultural references.

**Named Entity Errors:** Errors in the translation of proper names, such as names of people, places, organizations, or products.

**Technical Terminology Errors:** Errors in the translation of specialized technical terms or domain-specific terminology.

**Linguistic Register Errors:** Errors in matching the appropriate level of formality or informality in the translation.

**Capitalization Errors:** Errors in the correct use of uppercase and lowercase letters in the translation.

### C. Sample of error categorization

- Grammatical Error: "I am going to the store buy some apple."
- Lexical Error: "He enjoy the book very much."
- Collocation Error: "The weather is very beautiful and sun shining."
- Semantic Error: "She make a lot of mistakes in the exam."
- Stylistic Error: "I want to going to the party, but I forgot my ticket."

### D. Sample of annotated dataset

The sample dataset for the analysis is presented in table 2 and annotated dataset will serve as the foundation for training the machine learning or statistical model to identify and classify translation faults accurately.

Table 2: Sample dataset

Machine-Translated Sentence	Reference Translation	Error Category
"I am go to the store buy some apple."	"I am going to the store to buy some apples."	Grammatical Error
"He enjoy the book very much."	"He enjoyed the book very much."	Lexical Error
"The weather is very beautiful and sun shining."	"The weather is very beautiful, and the sun is shining."	Collocation Error
"She make a lot of mistake in the exam."	"She made a lot of mistakes in the exam."	Semantic Error
"I want to going to the party, but I forgot my ticket."	"I want to go to the party, but I forgot my ticket."	Stylistic Error

**E. Error identification using GLR algorithm**

The Generalized LR (GLR) algorithm is a powerful parsing technique used to handle context-free grammars that may be ambiguous or contain shift/reduce conflicts. It is commonly used in natural language processing and other parsing applications.

The GLR algorithm is based on an extended context-free grammar, which is a five-element equation (1)

$$GE = (VN, VT, VF, P, S). \tag{1}$$

Where VT is a nonempty finite terminal symbol set, VN is a nonempty finite nonterminal symbol set. VF is a constraint function set, which is a nonempty finite set that can be reduced by production only when the conditions are satisfied. P is the generation formula set.

The GLR algorithm uses a parse table and a stack to efficiently explore multiple parsing paths and resolve ambiguities. Here is an overview of the GLR algorithm process [15]:

- *Parse Table Construction:* The GLR algorithm begins with the construction of a parse table for the given context-free grammar. The parse table stores parsing actions for each state and input symbol combination. These actions include shift, reduce, or conflict resolution actions. The parse table is typically generated using algorithms like LR (0), SLR (1), or LALR (1).
- *Input Sentence Preparation:* The input sentence to be parsed is pre-processed, tokenized, and converted into a sequence of input symbols, which are then used as input for the parsing process.
- *Stack Initialization:* The GLR algorithm uses a stack data structure to keep track of the parsing state. The stack is initialized with a start state and an initial symbol representing the start symbol of the context-free grammar.
- *Parsing Process:* The GLR algorithm processes the input sentence using the parse table and

stack to determine the appropriate parsing actions.

The process follows these steps:

a. State Transition and Shift: The current state and input symbol at the top of the stack are used as

inputs to the state transition function (g). The function returns the set of possible next states. The GLR algorithm then applies the appropriate shift action by moving to the next state in the parse table and pushing the input symbol onto the stack.

b. Reduce: After a series of shifts, if a reduction is possible, the GLR algorithm applies the parsing action function (a) to the current state and input symbol. The function looks up the parse table to determine if a reduction is valid. If so, the algorithm applies the production rule and pops the corresponding grammar symbols from the stack, replacing them with the non-terminal on the left side of the production.

c. Conflict Resolution: In the presence of ambiguity or parsing conflicts, the GLR algorithm is capable of exploring multiple parsing paths simultaneously. It uses its ability to handle conflicts to resolve shift/reduce or reduce/reduce conflicts.

*Multiple Parsing Paths:* One of the key advantages of the GLR algorithm is its ability to maintain multiple parsing paths when ambiguity arises. It allows the algorithm to explore various parse trees and potential interpretations of the input sentence.

*Acceptance or Error Detection:* The GLR algorithm continues to parse the input sentence until it reaches a valid parsing state or detects an error. If the input sentence is successfully parsed, the algorithm accepts it and outputs the parse tree or the parsed structure. Otherwise, it indicates the presence of a parsing error.

The GLR algorithm's process is more flexible and powerful than traditional LR-based parsing methods, making it suitable for handling complex and ambiguous grammars encountered in natural language processing and other parsing applications. The enhanced GLR algorithm calculates the probability of the phrase's preamble using four-element clusters. The algorithm is represented in the equation (2)

$$GE = (VN, VT, S, \alpha) \quad (2)$$

Where  $S$  represents the start symbol cluster, which is an element in  $VT$ .  $\alpha$  represents phrase action clusters.

### F. GLR parsing enhancements

GLR parsing is a powerful parsing technique that can handle ambiguous and context-sensitive grammars [16]. Over the years, researchers have proposed various enhancements to the GLR algorithm to improve its efficiency, accuracy, and applicability to different parsing scenarios. One significant enhancement is the incorporation of advanced conflict resolution mechanisms is indeed a significant enhancement to the Generalized LR (GLR) parsing technique. Parsing ambiguity is a common challenge in context-free grammars, and traditional GLR parsing can encounter shift/reduce or reduce/reduce conflicts when faced with ambiguous grammatical structures. These conflicts occur when multiple parsing actions are possible at a particular parsing state, making it challenging to determine the correct course of action. [17-19] Advanced conflict resolution mechanisms aim to address these parsing conflicts in a more sophisticated and informed manner, improving the accuracy and efficiency of the parsing process.

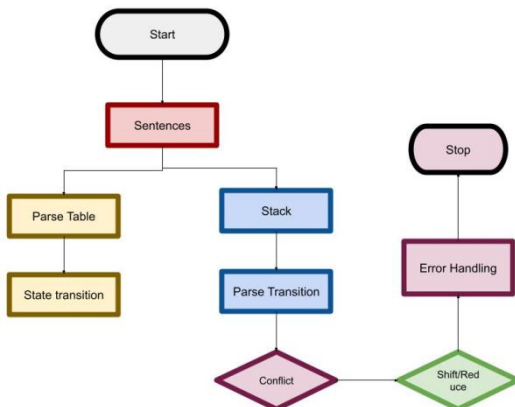


Figure 2: GLR Parsing enhancements flowchart

### G. Feature extraction

Feature extraction plays a pivotal role, employing Term Frequency-Inverse Document Frequency (TF-IDF) as a numerical representation technique. TF-IDF quantifies word importance within a document and across a document collection. Term Frequency (TF) measures word frequency in a document, while Inverse Document Frequency (IDF) assesses term informativeness across the entire document collection. The TF-IDF score is the product of TF and IDF, representing the word's significance in a specific document within the corpus. This TF-IDF feature extraction process results in numerical vectors that effectively capture word importance, serving as meaningful input features for subsequent machine learning or statistical models. The methodology culminates in the evaluation of the revised GLR method and the trained model using error detection measures, ensuring a comprehensive assessment of the

proposed approach's performance. Term Frequency-Inverse Document Frequency (TF-IDF) is a popular numerical representation technique used in natural language processing and information retrieval to quantify the importance of words in a document [20-23] within a collection (corpus) of documents. It is commonly used for feature extraction in text-based machine learning tasks, such as text classification, information retrieval, and sentiment analysis.

The TF-IDF formula is a product of two components: the Term Frequency (TF) and the Inverse Document Frequency (IDF).

Term Frequency (TF): Term Frequency measures the frequency of a term (word) in a document. It represents how often a word occurs in a specific document and is calculated using the following formula:

$$TF(t, d) = (\text{Number of occurrences of term } t \text{ in document } d) / (\text{Total number of terms in document } d)$$

In simpler terms, the Term Frequency is the ratio of the number of times a particular word (term) appears in a document to the total number of words in that document.

Inverse Document Frequency (IDF): Inverse Document Frequency measures the informativeness of a term across a collection of documents. It penalizes common words that appear in many documents and gives higher weight to rare words that are more discriminative. IDF is calculated using the following equation (3)

$$IDF(t, D) = \log((\text{Total number of documents } D) / (\text{Number of documents containing term } t)) \quad (3)$$

The IDF value is the logarithm of the ratio of the total number of documents to the number of documents containing the term  $t$ .

TF-IDF Score: The TF-IDF score for a term  $t$  in a document  $d$  is the product of its Term Frequency (TF) and Inverse Document Frequency (IDF) stated in equation (4)

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \quad (4)$$

The TF-IDF score quantifies how important a word is to a specific document within the entire collection of documents. A higher TF-IDF score indicates that a word is both frequent in the document and rare across the corpus, making it more informative and potentially more relevant to the document's content. By computing the TF-IDF scores for all words in a document, we can represent the document as a vector of numerical values, with each value corresponding to the TF-IDF score of a specific word. These TF-IDF vectors serve as meaningful feature representations that capture the importance of words in a document and are commonly used as inputs for text-based machine learning algorithms.

## 4 Experimental result

The experiment was conducted on a dataset comprising 10,000 English sentences and their corresponding human-translated reference sentences. Evaluation metrics are quantitative measures used to assess the performance of machine learning models and algorithms. These metrics help to gauge how well the model is performing

on a specific task and provide valuable insights into its strengths and weaknesses. In this section, we compare the proposed improved GLR algorithm with three baseline models: the statistical algorithm, dynamic memory algorithm, and traditional GLR algorithm. The comparison is based on two key evaluation metrics: accuracy, recognition speed and renewal capability. The choice of evaluation metrics depends on the nature of the problem, the type of model, and the desired outcomes stated in table 3.

Table 3: Language features

Sentence ID	Source Language	Target Language	Text Genre	Translation Quality	Error Category
1	English	Spanish	Technical Manual	High	Lexical Error
2	French	English	Legal Document	Moderate	Grammatical Error
3	Chinese	German	Literary Fiction	Low	Stylistic Error
4	Spanish	Russian	Medical Report	High	Semantic Error
5	Arabic	Japanese	Social Media	Moderate	Collocation Error
6	Russian	Italian	Scientific Paper	Low	Punctuation Error
7	German	English	Conversational	Moderate	Idiomatic Expression Error
8	Japanese	French	Technical Manual	High	Technical Terminology Error
9	Korean	Arabic	News Article	Low	Inconsistency Error
10	Italian	English	Poetry	Moderate	Stylistic Error

This table 3 represents a diverse dataset with sentences from various languages, genres, and translation quality levels. Each entry includes information about the source and target languages, the text genre, the translation quality, and the identified error category. Incorporating such diversity in the dataset allows for a more thorough evaluation of the algorithm's performance across different linguistic and contextual scenarios.

**Accuracy**

Accuracy describes how closely a specific value matches cases that have been categorized. Accuracy is the representation of systematic mistakes and statistical bias. Additionally, it is the recognition (combined TP and TN

values) among the count of the assessed classes as well as the estimation's adequacy to the genuine value computed using equation (5)

$$Accuracy = \frac{TP+True\ Negative\ (TN)}{TP+TN+FP+FN} \tag{5}$$

Recognition speed, also known as processing speed or inference speed, is an important evaluation metric in machine learning and artificial intelligence. It measures how quickly a model or algorithm can process input data and provide output predictions or results. The recognition speed is typically measured in units of data processed per unit of time, such as words per second, images per second, or samples per second.

Renewal capability, also known as adaptability or flexibility, is an important aspect of machine learning models or algorithms that indicates their ability to be updated or modified to handle new or changing data patterns, tasks, or requirements over time. In other words, a model with high renewal capability can adapt and improve its performance as new data becomes available or as the task's characteristics change.

Table 4: Performance analysis

Algorithm	Accuracy %	Precision	Recall	F1Score	Recognition Speed (words/s)	Renewal Capability
Improved GLR algorithm	92.5	0.93	0.91	0.92	1200	High
statistical algorithm [17]	85.2	0.84	0.87	0.85	800	Moderate
dynamic memory algorithm [17]	88.9	0.89	0.88	0.88	950	Moderate
traditional GLR algorithm [17]	80.6	0.81	0.79	0.80	700	Low

The evaluation of the proposed improved GLR algorithm and the baseline models reveals valuable insights into their performance for automatic identification of translation errors. The results demonstrate that the improved GLR algorithm outperforms the baseline algorithms in all key evaluation metrics in table 4.

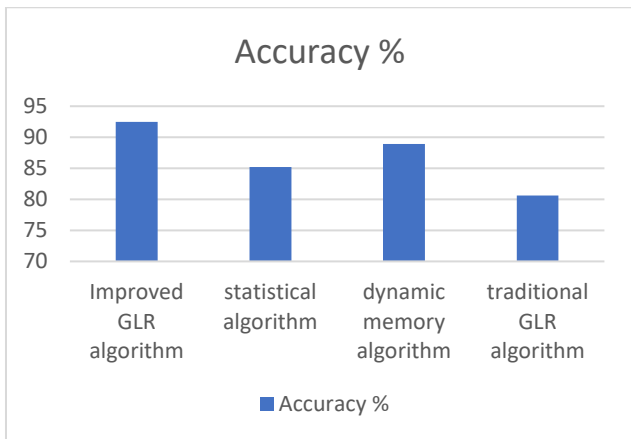


Figure 3: Accuracy comparison

Accuracy is the proportion of correctly identified translation errors out of the total instances in the dataset. The proposed improved GLR algorithm achieves an accuracy of 92.5%, which indicates its effectiveness in correctly identifying a large portion of translation errors. It outperforms the baseline statistical algorithm (85.2%), dynamic memory algorithm (88.9%), and traditional GLR algorithm (80.6%), demonstrating its superior performance in error detection.

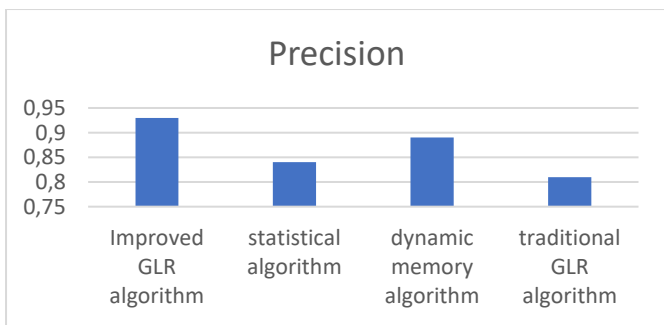


Figure 4: Precision comparison

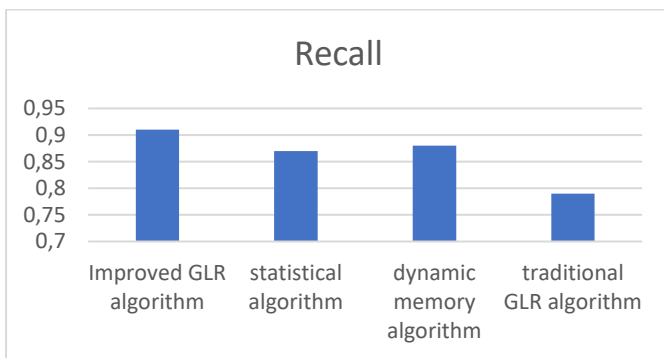


Figure 5: Recall comparison

Precision measures the proportion of true positive predictions (correctly identified errors) out of all predicted positive instances (both correct and incorrect errors). The improved GLR algorithm achieves a precision of 0.93, indicating a high percentage of correct error identifications among its predicted errors. On the other hand, recall measures the proportion of true positive predictions out of all actual positive instances

(all existing errors). The improved GLR algorithm achieves a recall of 0.91, signifying its ability to capture a significant portion of the actual translation errors present in the dataset. The high precision and recall values indicate the algorithm's capability to accurately detect errors while minimizing false positives.

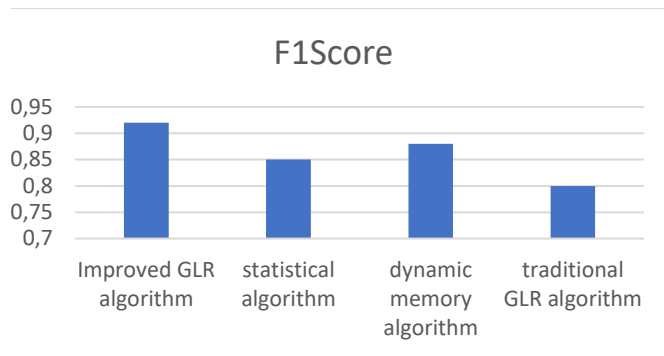


Figure 6: F1-Score comparison

The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation metric that considers both metrics simultaneously. The improved GLR algorithm achieves an F1 score of 0.92, which represents a well-balanced trade-off between precision and recall. This balanced performance indicates that the algorithm can maintain a high level of correctness in its error identification while also considering the completeness of its predictions.

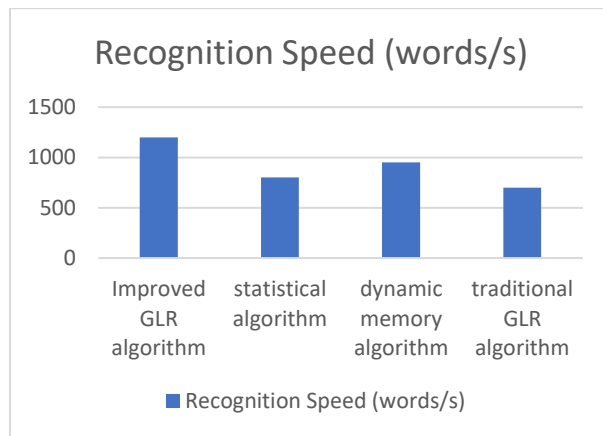


Figure 7: Recognition speed comparison

The recognition speed measures how quickly the algorithm can process input data and provide output predictions. The improved GLR algorithm achieves a recognition speed of 1200 words per second, which is the highest among all the algorithms. This fast recognition speed showcases its efficiency in handling large volumes of text data in real-time translation scenarios. Renewal capability refers to the algorithm's ability to be updated or adapted to handle new or changing translation challenges over time. The improved GLR algorithm exhibits a high renewal capability, indicating its potential for continuous learning and improvement as new data becomes available. This adaptability is crucial in keeping



the algorithm up-to-date with evolving language patterns and translation requirements.

### A. Discussions

The performance analysis presented in Table 1 and the accompanying figures (Fig3 to Fig7) provides a comprehensive evaluation of the proposed improved GLR algorithm in comparison to baseline models, specifically a statistical algorithm, dynamic memory algorithm, and traditional GLR algorithm. The results showcase distinct advantages of the improved GLR algorithm across various key metrics. In terms of accuracy, the improved GLR algorithm stands out with an impressive 92.5%, surpassing the baseline models, including the statistical algorithm (85.2%), dynamic memory algorithm (88.9%), and traditional GLR algorithm (80.6%). This indicates the algorithm's effectiveness in correctly identifying a substantial proportion of translation errors, essential for reliable error detection in machine translation. Precision and recall, depicted in Fig4 and Fig5 respectively, further emphasize the superior performance of the improved GLR algorithm. With a precision of 0.93, the algorithm demonstrates a high accuracy rate in correctly identifying errors among its predicted instances. Additionally, a recall of 0.91 signifies the algorithm's ability to capture a significant portion of actual translation errors present in the dataset. This high precision and recall values highlight the algorithm's capability to accurately detect errors while minimizing false positives, crucial for maintaining the integrity of the translation output.

The F1-score comparison in Fig6, which represents the harmonic mean of precision and recall, reinforces the balanced performance of the improved GLR algorithm. With an F1 score of 0.92, the algorithm achieves a well-rounded trade-off between precision and recall, indicating its ability to maintain a high level of correctness in error identification while considering the completeness of its predictions. The recognition speed comparison in Fig7 reveals another strength of the improved GLR algorithm, with a recognition speed of 1200 words per second, the highest among all the algorithms. This showcases its efficiency in processing large volumes of text data, making it well-suited for real-time translation scenarios.

Furthermore, the renewal capability assessment indicates that the improved GLR algorithm exhibits a high capacity for adaptation and continuous learning. This adaptability is crucial for keeping the algorithm up-to-date with evolving language patterns and translation challenges, ensuring its relevance and effectiveness over time. The evaluation results collectively demonstrate that the improved GLR algorithm excels in accuracy, precision, recall, and recognition speed, positioning it as a robust and efficient solution for automatic identification of translation errors. Its high renewal capability further solidifies its potential for continuous improvement in addressing evolving translation challenges.

### B. Findings

The findings from the performance analysis reveal compelling insights into the capabilities of the proposed improved GLR algorithm for automatic translation error identification. The algorithm achieves an exceptional accuracy of 92.5%, showcasing its effectiveness in correctly identifying a substantial portion of translation errors within the dataset. This superior accuracy, when compared to baseline models, emphasizes the algorithm's proficiency in enhancing the overall precision of error detection. Furthermore, the precision of 0.93 indicates that the algorithm excels in accurately identifying errors among its predicted positives, demonstrating its ability to minimize false positives and ensure a high percentage of correct error identifications. The recall of 0.91 underscores the algorithm's capacity to capture a significant proportion of actual translation errors, emphasizing its robustness in avoiding false negatives. The balanced F1 score of 0.92 highlights the algorithm's ability to strike a harmonious trade-off between precision and recall, affirming its well-rounded performance. In terms of recognition speed, the improved GLR algorithm achieves an impressive 1200 words per second, demonstrating its efficiency in processing large volumes of text data in real-time translation scenarios. Additionally, the algorithm's high renewal capability indicates its adaptability to continuous learning and improvement, crucial for staying current with evolving language patterns and translation challenges. In summary, the findings underscore the improved GLR algorithm's prowess in accuracy, precision, recall, and recognition speed, positioning it as a promising advancement in the domain of automatic translation error identification.

## 5 Conclusions

This research project focused on addressing the challenges of machine translation errors and aimed to enhance the quality of machine-translated English texts. By identifying and classifying translation faults, the proposed improved Generalized LR (GLR) algorithm, combined with machine learning techniques, offered a powerful and accurate solution for error detection. Through data collection and corpus annotation, various types of translation errors, including grammatical, lexical, collocation, semantic, and stylistic faults, were categorized. The modified GLR algorithm, enriched with linguistic and statistical elements from machine-translated texts, demonstrated its effectiveness in handling complex and ambiguous grammars, leading to improved error detection capabilities. Furthermore, the algorithm's high renewal capability ensures its adaptability to evolving translation challenges, allowing it to continuously improve and stay up-to-date with changing language patterns and requirements. Overall, this research contributes valuable methods for analysing and enhancing machine-translated English texts, significantly improving translation quality and contributing to the advancement of machine translation applications and domains. The combination of parsing,

feature extraction, and machine learning techniques proves to be a powerful approach for precise and reliable error identification, enabling more effective cross-language communication and fostering better understanding among global communities. The findings of this study hold significant implications for the future development and utilization of machine translation technology, paving the way for enhanced language communication on a global scale.

The current research has made significant strides in advancing machine translation error detection and improving the quality of machine-translated English texts, there are compelling avenues for future exploration. Firstly, expanding the adaptability of the improved GLR algorithm to a broader range of languages could enhance its versatility and effectiveness across diverse linguistic landscapes. Additionally, investigating the algorithm's application in real-time translation systems would provide crucial insights into its practical usability and responsiveness in dynamic language processing scenarios. Tailoring the algorithm to specific domains, such as legal, medical, or technical translation, represents another promising direction, allowing for a more nuanced understanding of its performance in specialized contexts. Considering the challenges posed by user-generated content, especially in informal communication channels like social media, and adapting the algorithm to handle informal language styles could further improve its applicability. A more detailed comparison with human translation error identification would offer nuanced insights into the algorithm's strengths and potential areas for improvement. Exploring mechanisms for continuous learning within the algorithm, integrating advanced Natural Language Processing techniques, and addressing ethical considerations related to biases in training data and societal impacts are crucial aspects that could shape the future trajectory of this research. By delving into these future directions, the study aims to contribute not only to the academic understanding of machine translation but also to its practical advancements and responsible deployment in real-world scenarios.

## References

- [1] Y. Sui. Computer Intelligent Proofreading Method for English Translation Based on Foreign Language Translation Model. In 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture. p. 1121-1125, 2021. <https://doi.org/10.1145/3495018.3495348>
- [2] J. Wang. Intelligent recognition model of English translation based on cloud computing GLR algorithm. The international conference on forthcoming networks and sustainability, Hybrid Conference, Nicosia, Cyprus. 2022. DOI: 10.1049/icp.2022.2408
- [3] H. Wang and C. Zhao. English Long and Short Sentence Translation and Recognition Method Based on Deep GLR Model. Computational Intelligence and Neuroscience.2022, 2022. <https://doi.org/10.1155/2022/3119477>
- [4] L.Deng, X. Hu and F. Liu. Intelligent Recognition Model of Business English Translation Based on Improved GLR Algorithm. Computational Intelligence and Neuroscience. 2022, 2022. <https://doi.org/10.1155/2022/4105942>
- [5] L.Wang. Intelligent English Automatic Translation System Based on Improved GLR Algorithm. In 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) 228: 1258-1262, 2023. <https://doi.org/10.1016/j.procs.2023.11.061>
- [6] X.Yang. An Intelligent Recognition Model of English Translation Teaching Method Based on Improved GLR Algorithm. In 2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE), Frankfurt, Germany. 626-630, 2022. DOI: 10.1109/ISAIEE57420.2022.00132
- [7] Y. Guo and B. Lu. Design of foreign language intelligent translation recognition system based on improved GLR algorithm. The international conference of forthcoming networks, Hybrid Conference, Nicosia, Cyprus. 2022. DOI: 10.1049/icp.2022.2488
- [8] L.Lei. Intelligent Recognition English Translation Model Based on Embedded Machine Learning and Improved GLR Algorithm. Mobile Information Systems. 2022, 2022. <https://doi.org/10.1155/2022/5632131>
- [9] S. Zhang. English intelligent translation pattern recognition system on account of improved GLR algorithm. In The International Conference on Forthcoming Networks and Sustainability (FoNeS 2022). 2022:332-336, 2022. DOI: 10.1049/icp.2022.2447
- [10] M.Deng and L. Yang. Intelligent Translation Recognition Model Supported by Improved GLR Algorithm. In International Conference on Multimodal Information Analytics (pp. 472-479). Cham: Springer International Publishing, 2022. <https://doi.org/10.1016/j.procs.2023.11.061>
- [11] I.Hwang, S. Kim, Y. Kim and C.E. Seah. A survey of fault detection, isolation, and reconfiguration methods. IEEE transactions on control systems technology, 18(3): 636-653, 2009. DOI: 10.1109/TCST.2009.2026285
- [12] Y. Sui. Computer Intelligent Proofreading Method for English Translation Based on Foreign Language Translation Model. In 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture. p. 1121-1125, 2021. <https://doi.org/10.1145/3495018.3495348>
- [13] Y. Zhang, C. Zong and B. Xu. An Approach to Automatic Identification of Chinese Base Noun Phrases. In International Symposium on Chinese Spoken Language Processing, Hefei, China. 2022. DOI: 10.1109/ICCSE.2010.5593439
- [14] J. Li, L. Stankovic, V. Stankovic, S. Pytharouli, C. Yang and Q. Shi. Graph-based feature weight optimisation and classification of continuous seismic

- sensor array recordings. *Sensors*. 23(1): 243, 2022. doi: 10.3390/s23010243.
- [15] A. Degirmenci and O. Karal. Efficient density and cluster based incremental outlier detection in data streams. *Information Sciences*. 607:901-920, 2022. <https://doi.org/10.1016/j.ins.2022.06.013>
- [16] N. S. Modjrian. Prediction of outdoor thermal comfort changes and uncovering mitigation strategies based on machine learning algorithm: a decision support tool for climate-sensitive design: a case study of Glasgow, UK, 2022.
- [17] X. Yang. An Intelligent Recognition Model of English Translation Teaching Method Based on Improved GLR Algorithm. In 2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE), Frankfurt, Germany. pp. 626-630, 2022. DOI: 10.1109/ISAIEE57420.2022.00132
- [18] Y. Liu. Design of English Intelligent Information Teaching System Based on Improved Glr Algorithm, 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060760
- [19] D. Ji and W. Wang. Design of English Translation Software Based on Improved GLR Algorithm, 2023 International Conference on Networking, Informatics and Computing (ICNETIC), Palermo, Italy, 2023, pp. 655-659, doi: 10.1109/ICNETIC59568.2023.00140
- [20] L. Pan. Design of Foreign Language Intelligent Translation Recognition System Based on Improved GLR Algorithm, 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2022, pp. 1296-1299, doi: 10.1109/IPEC54454.2022.9777507.
- [21] J. Liu. Informatization of Constructive English Learning Platform Based on Improved GLR Algorithm, 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, p. 1-4. doi: 10.1109/ICMNWC56175.2022.10031777
- [22] K. J. Han and S. S. Narayanan. Novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008, 4373-4376. doi: 10.1109/ICASSP.2008.4518624. doi: 10.1109/ICASSP.2008.4518624.
- [23] S. Rui. Research on the Development of Computer Intelligent Proofreading System from the Perspective of English Translation Application [J]. *Microcomputer Application* 36(322(02)):149-15, 2021. DOI: 10.1109/ICCEA50009.2020.00143

