Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **4**, November **2024**

Research on Multimodal Interactive Machine Translation Based on Natural Language Processing Technology

Yuejun Zhang*

Department of Foreign Languages
Xi'an Jiaotong University City College, Xi'an 710000, China
yuejun.zhang@ucl.ac.uk

*Corresponding author: Yuejun Zhang

Received November 8, 2023, revised March 24, 2024, accepted June 28, 2024.

Abstract. *With the continuous maturity and wide application of artificial intelligence technology, people can use machines to express themselves in their study, work and life. Machine translation (MT) has gradually become a trend. However, traditional text translation systems cannot fully meet current needs, and how to effectively achieve human–computer interaction in translation systems is still a very thought-provoking issue. This study mainly explored the characteristics of MT and its semantic model. A complete and clear multimodal interaction system was established by analysing and extracting input information from multiple data sources, such as images, text and speech. A human–machine interface centred on user needs was designed, and relevant hierarchical interactions and training algorithms were combined to calculate MT automation evaluation indicators. This system was applied to MT and natural language processing (NLP) to complete the communication and interaction amongst different languages faster, better and more directly than before. This study compared the effectiveness of multimodal MT and traditional text translation from two aspects, automatic and manual evaluation, demonstrating the advantages of multimodal MT in terms of similarity and accuracy. Moreover, the language fluency of traditional pure text MT was only approximately 2.5288 and 2.6514. The multimodal MT increased to approximately 3.443, 3.4374 and 3.5032, demonstrating the enormous potential of multimodal translation for information exchange amongst different languages. MT based on multimodal and multilevel unified interaction improved translation quality and the understanding and use of various semantic information, bringing further seamless and effective integration to human–machine interaction. Moreover, MT was expected to make further breakthroughs in multi-format data fusion and NLP.*

Keywords: Multimodal Multilevel, Unified Interaction, Machine Translation, Natural Language Information Processing, Feature Extraction

1. **Introduction.** Language is the foundation for achieving intelligent information exchange and promoting knowledge sharing and social development. As an important component, translation plays an irreplaceable role in interpersonal communication. As the main development directions in the field of artificial intelligence, machine translation (MT) and natural language processing (NLP) are also the fastest-developing technologies in computer science and technology. However, traditional methods have certain limitations when dealing with multimodal data and multilevel interactions. These methods cannot meet the high concurrency problem between big data and complex semantics, thereby

proposing new requirements for machine learning ability and reasoning speed. This study aimed to explore and examine MT and NLP methods based on multimodal and multilevel unified interaction and to improve the effectiveness and accuracy of translation and processing by integrating multiple data and interaction methods.

In the context of globalisation, language barriers remain obstacles to obtaining information, and relying solely on manual translation cannot meet translation needs. Tools such as MT are becoming increasingly popular owing to their potential to overcome this problem [1]. Syntax knowledge can effectively improve the performance of neural machine translation (NMT). Source and target dependency structures can improve the quality of translation, and their effects can be accumulated [2]. NLP has gained considerable attention in expressing and analysing human language through computation, and its application has expanded to various fields, such as MT, spam detection, information extraction, abstracts, medicine and question answering [3]. End-to-end training and representation learning are key features of deep learning, making it a powerful tool for NLP. NLP has five main tasks, namely, classification, matching, translation, structured prediction and sequential decision-making processes. For the first four tasks, deep learning methods outperform or significantly outperform traditional methods [4]. The utilisation of a hierarchical to sequential attention NMT model to handle the optimal model parameters for learning long parallel sentences and effectively utilising different contexts can not only improve parameter learning but also allow well exploration of translation contexts of different ranges [5]. MT and NLP have extensive exploration and practice in the field of machine intelligence, particularly in semantic understanding and expression, promoting the development of translation towards intelligence.

Attention is an increasingly popular mechanism widely used in various neural architectures. The focus of attention structure models in NLP is to design models for vector representation of text data, providing performance gains [6]. Deep learning in computer vision and NLP provides state-of-the-art pre-training models, training scripts and training logs to facilitate rapid prototyping design and repeatable exploration, achieving efficient customisation [7]. The application of NPL-related technologies has improved the ability of machines to classify and recognise information during the translation process.

The free and available online data construction baseline system is used to filter the six languages of European NMT models developed by the corpus. The results generated by the system are compared with those generated by Google Translate, which proves that the method can effectively improve the system's performance. In addition, the generation speed is faster, and the quality is higher in terms of facilitating multilingual review in a secure environment [8]. Artificial intelligence is the imitation and learning of humans, and humans are intelligent agents that work together in multiple modes of seeing, listening and speaking. Multimodal technology is the key to future applications of artificial intelligence. Multimodal machine learning faces broad challenges, including representation, translation, alignment, fusion and collaborative learning. This new classification method would enable technicians to greatly understand its current situation and determine future directions [9]. Multimodal MT enables mutual translation amongst multiple languages, solves the problem of information asymmetry amongst different languages and provides high-quality language communication and communication services.

Traditional MT systems have many shortcomings in speech recognition, vocabulary input and speech speed calculation, particularly for machine learning models that cannot effectively fuse multi-source information. This study utilised multimodal language analysis technology to partition corpus in sentence structure and contextual environment, extracted semantic features through natural language modelling and established MT models in combination with existing machine learning algorithms. This approach can effectively improve machine learning algorithm performance and improve MT efficiency and accuracy whilst reducing the labour intensity of manual translators. Compared with conventional methods, MT based on multimodal interaction improves the correlation amongst texts, reduces incorrect output caused by incorrect annotations during the translation process and reduces error rates. MT achieves efficient and accurate translation of target documents and improves user experience.

2. Processing Method and Application of MT Multilevel Interaction.

2.1. **Development of MT and NLP Methods.** With the continuous maturity of computer science and technology, humans are no longer limited to language or text in social life. Machines can process, transform and store various pieces of information and output corresponding results according to different requirements, providing people with many new functions and experiences. MT and NLP are products of the development of this technology to a certain extent, and they would become an important means to solve current complex problems [10].

MT is an information processing mode based on computer technology, which has the advantages of automated reasoning, accuracy and speed. Owing to its involvement in several complex logical operations, MT has a wide range of applications. Many scholars have proposed some system models based on MT, but most of these models only consider MT as a simple linear decision-making problem, ignoring the impact of multimodal feature information on the performance of machine learning algorithms.

As a new type of artificial intelligence theory, NLP explored in this article has been widely applied in recent years. NLP transforms knowledge that people need into text form, enabling it to be understood, accepted and transmitted, thereby achieving the goal of solving practical problems, such as speech recognition and image editing. However, natural language is a complex system with strong nonlinear characteristics, containing rich and abstract information that is difficult to obtain using traditional manual methods. Therefore, using advanced computer vision and intelligent information processing technology is particularly important to achieve this goal.

Based on this background, this study mainly analysed the translation task of multimodal machines, using the description of the source language and its corresponding images as input and the sentences of the target language as output. Combined with the theory of multimodal systems in machine vision technology, the goal of establishing a target corpus is achieved, which can quickly and accurately complete multilingual MT without changing the existing vocabulary structure. The difficulty of this task lies in how to effectively integrate two or more modal information, construct corresponding templates by extracting various modal features and design patterns that can meet specific needs according to the semantic rules to be determined, thereby achieving auxiliary translation.

2.2. **Progress in Multimodal Data Processing Technology.** Multimodal interaction is composed of visual, auditory, olfactory, tactile and taste sensory interactions, completed by the touch of the eyes, ears, nose, mouth and skin, respectively. This technology is applied in reality, which revolves around these senses and integrates multiple sensory interaction technologies to form a multimodal interaction form.

Amongst the existing common multi-mode data preprocessing methods, the improved algorithm based on the combination of least squares support vector machine and Markov chain model has a better classification effect. However, its training time is long, and it is not suitable for large-scale real-time applications. In addition, wavelet transform denoising is a fast and simple denoising algorithm that can effectively remove noise and retain useful signal information, thereby enhancing recognition accuracy. However, wavelet transform denoising cannot extract multiple features simultaneously and does not have an adaptive filtering function, which cannot achieve automatic recognition of targets in complex backgrounds.

In summary, in MT systems, multiple levels of interaction amongst different modal data can obtain comprehensive, accurate and convenient information. The multimodal data processing technology used in this study combines NLP and multimodal technology and utilises the complementarity and interaction relationship of different modal data to effectively improve the quality and efficiency of MT, providing great solutions for many practical application scenarios.

3. Multimodal Data Processing and Feature Extraction.

3.1. Multimodal Data Fusion.

Multimodal data fusion is the way to combine different types of text, image, voice and other data at the semantic and expression levels, strengthen the limitations of single modal data and improve the accuracy and efficiency of information processing and analysis [11,12] . Compared with traditional single-modal data, multimodal data has a wider audience and better enriches the dimensions and characteristics of the data. However, how to fuse these heterogeneous data remains a challenging issue.

The mainstream methods include feature and decision-level fusions. The former mainly integrates information based on features, whereas the latter constructs multidimensional structural models through decision trees [13,14]. Tables 1 and 2 show the specific implementation methods.

Table 1. Feature-level fusion methods

	Big difference in data sources	Accuracy	Eigenvalue or number of features changes	Aim
Simple splicing	_	_	_	Feature vectors of different modes are directly spliced
Weighted fusion	√	√	_	Multiple eigenvectors are weighted and averaged
Depth feature extraction	_	√	√	Integrate information to generate a feature tag set

Note: "_" indicates no, and "√" indicates yes.

Table 2. Decision-level fusion methods

	Precision	Preconditioning	Stability	Aim
Weighted average	↓	_	↓	Decision of multiple data is normalised weighted
Majority rule	↑	_	↑	Compare attributes by discriminator
Bayesian network	↑	√	↑	Find the joint probability distribution

Note: "_" indicates no; "√" indicates yes; "↑" indicates high and "↓" indicates low.

Feature and decision-level fusions can greatly extract considerable unknown factors in complex systems. The multimodal data fusion used in this article can effectively improve processing efficiency and accuracy when analysing multi-source information whilst also reducing misjudgements caused by noise or other interference and enhancing recognition ability. However, owing to its high computational complexity and limited application range, multimodal data fusion is still not widely applicable in practical scenarios. Calculate the score for each data source under a certain criterion. Normalize the scores of each data source to ensure their comparability under the same criteria. Calculate the weight of each data source using the average score. Use the weight of each data source to calculate the fusion result of the decision layer, and obtain the final decision result by weighted averaging the scores of each source.

3.2. **Multimodal Feature Extraction.** In the fusion of multi-source data, to extract complex feature information, several signals with similar scales, frequencies and time-frequency domain characteristics for classification should be used. The feature extraction methods of multimodal data used in this study include text feature extraction, image feature extraction and voice feature extraction. They can effectively describe various features in a single mode into multiple forms and combinations and can obtain further comprehensive and accurate information by analysing these characterisation results, as shown in Table 3.

Table 3. Multimodal feature extraction methods

Method	Type	Advantage	Shortcoming	Accuracy rate
Text	Bag model	Convenient and practical	Ignore text word order and sentence structure	85%-94%
	Topic model	Process large amounts of text	Training is complex and time-consuming	
Graphics	Convolutional neural networks	Image feature extraction and classification are effective	High data requirements	76%-90%
	Scale invariant feature transform	Rotational invariance	Sensitive to light changes and noise	
Voice	Short-time energy and zero crossing rate	High real-time	Less information	88%-94%
	Mel-scale frequency cepstral coefficients	Strong anti-noise performance	Affected by parameters	

The Bag model ignores word order and sentence structure, representing text as an unordered collection of vocabulary items. For example, for the text ”Cats like sunny days, dogs like rainy days”, vocabulary items cat, like, sunny days, dog, rainy will be extracted.

At the current level of technology, pure text MT has made significant progress, mainly focusing on translating one natural language into another. However, in the face of complex language environments and multilingual user groups, the difficulties brought by this text context-based conversion method are becoming increasingly apparent. Owing to its design based on specific objective functions and constraints, this method cannot effectively handle the differences amongst massive heterogeneous data sources, easily falling into local optima and being unable to distinguish noise, which greatly limits the reliability and application range of translation. In response to the above issues, the multimodal-based multilevel unified interactive translation model further considers various physical forms of input modes, such as text, images and speech. This model can dynamically adjust model parameters according to actual application situations, thereby enhancing pattern recognition ability. Moreover, the model has high interactivity and great semantic quality and is currently one of the ideal human–computer interaction methods [15, 16].

4. Design and Implementation of a Multilevel Unified Interaction Model.

4.1. **Design of Multilevel Interaction Architecture.** In MT systems, the architecture design and implementation of multilevel interaction is an important content. It maps the interoperability amongst various languages into semantic layer transformations, enabling each text to be understood and represented. The data are processed by machines to obtain the required information. Ultimately, the information is fed back to the user based on the processing results to perform the corresponding work [17]. Establishing a multilevel interaction model architecture includes modules, such as low-level feature extraction, middle-level semantic understanding and top-level decision generation. Such modules abstractly describe problems at different levels, forming a complex and clear hierarchical structure to achieve information transmission and interaction, as shown in Figure 1.

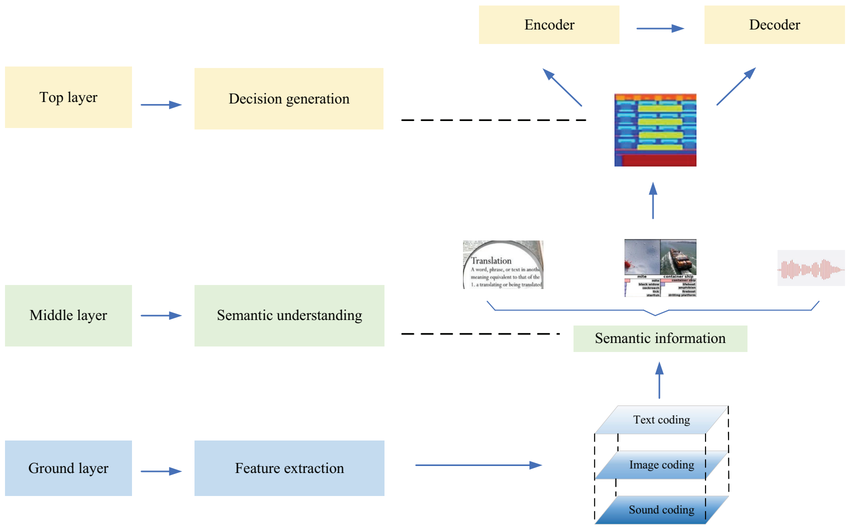


Figure 1. Multimodal MT interaction architecture

The low-level feature extraction of the multilevel unified interaction model designed in this study mainly deals with text input, image input and voice input. This approach effectively captures the long-distance dependency amongst text sequences through the self-attention mechanism, uses convolutional neural network [18, 19] and other models to extract and code image features or uses the acoustic model and time series model to process voice signals. This method also transmits the results to subsequent modules for the next translation operation. Middle-level semantic understanding generally extracts semantic information from the underlying features and uses related algorithms to analyse and interpret knowledge representation and reasoning, including reasoning, similarity calculation, classification, clustering and other advanced functions. Top-level decision generation is the process of transforming semantic representations into corresponding layer graph structures and logical rules to form the final target solution. Data synchronisation updates are achieved for each target object recognition by receiving heterogeneous data from different sources and previously calculated specific context vectors. Based on a pre-trained encoder [20, 21], the input sequence is modelled, achieving encoding and decoding functions and generating target-side translation results. Convolutional neural networks effectively extract image features through convolutional and pooling layers. Convolutional operations capture local features and pooling reduces the number of parameters. This mechanism is effective for image classification as it preserves spatial structural information. Convolutional neural networks have high data requirements because they rely on a large amount of annotated data to learn features. The larger the amount of data, the more the network can generalize, adapt to a wider range of image changes, and improve classification accuracy.

In the low-level feature extraction module, the text input adopts a self attention mechanism. This mechanism assigns different attention weights to words at different positions in the text sequence, enabling the model to capture long-distance dependencies. Through self attention, the model can focus on relevant vocabulary in the sequence, better understand contextual information, and help improve the accuracy and fluency of multimodal translation.

In the intermediate semantic understanding stage, use relevant algorithms for knowledge inference. Similarity calculation can measure the degree of semantic similarity between entities or concepts, thereby identifying correlations. Clustering algorithms help to combine similar knowledge and reveal hidden patterns. These algorithms can parse semantic relationships, detect connections between concepts, and provide deeper reasoning for understanding inputs.

In addition to MT, the multilevel interaction processing method can also be applied to other NLP problems, such as text summarisation and sentiment analysis, providing important ideas and technical support for the development of related applications. This study designed a multimodal MT platform based on this, as shown in Figure 2.

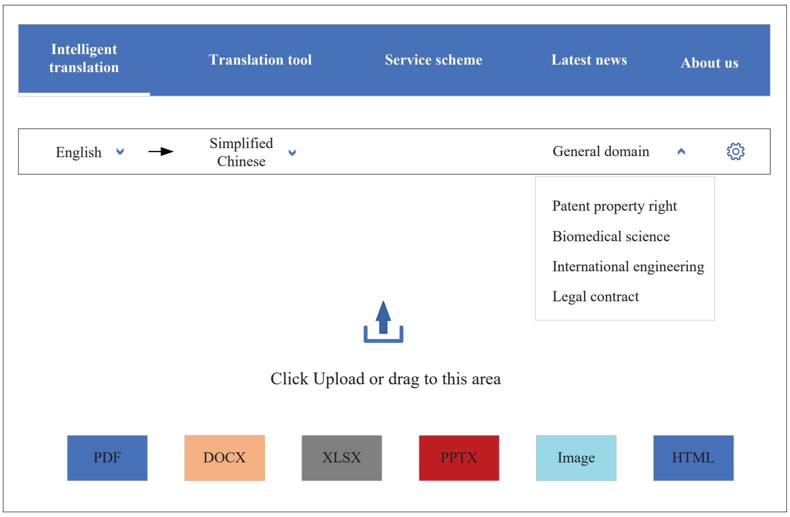


Figure 2. Multimodal MT platform

In this translation model, various data types, such as images, audio, video and text, can complement and support each other. The meanings and characteristics in natural language can be comprehensively recognised, understood and expressed through different information fusion methods and coordination modes. In addition, the system has information processing and learning capabilities at multiple levels, including word-level language modelling, sentence-level parsing and semantic parsing and even some complex semantic and social factors in cross-cultural aspects. The system has achieved effective processing and storage of various media forms and contents and can provide rich and complete information retrieval functions and personalised service functions. In summary, MT and NLP with multimodal and multilevel unified interaction are complex systems that integrate the advantages of multiple disciplines and technologies. They can help researchers and engineers solve problems in the field of NLP and provide great human–machine interaction methods and information communication channels for daily life and work scenarios [22].

In the architecture of a multi-level interaction model, various modules collaborate with each other to achieve efficient translation. The input module receives source text, images, and audio, and passes them to the feature extractor of the text and images. These extracted features are combined in the fusion module to form a multimodal representation. The encoding module uses self attention mechanism to capture long-distance dependencies of text sequences and generate context aware text representations. The decoding module utilizes a generative model and attention mechanism to generate the target language sequence through multimodal representation. The model optimizes parameters through gradient descent algorithm to achieve end-to-end training. This design enables the model to better understand and process multimodal inputs, improving translation performance.

4.2. **Multimodal Unified Representation Learning.** Multimodal unified representation learning eliminates the differences amongst different input modalities, thereby allowing multiple modalities to fuse closely and further promoting the development of multimodal tasks, such as MT, visual and language questions and answers. The end-to-end MT mentioned in this study is a typical representative form of unified representation learning for multimodal data, which can achieve cross-modal information fusion and expression in complex scenarios, as shown in Figure 3.

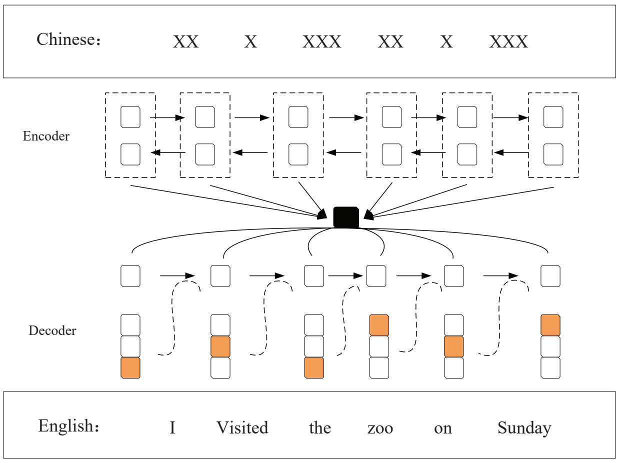


FIGURE 3. End-to-end multimodal MT

The end-to-end multimodal MT used in this article integrates information, such as text, image and sound, and achieves high-quality translation tasks through a unified deep neural network model [23, 24]. In this process, the semantic differences amongst different languages and speech processing technology can be used to establish the corresponding target word mapping relationship, eliminate many manual steps required in the traditional process and strengthen the context understanding ability, to generate further natural and smooth translation results and enhance the translation accuracy. In practical applications, end-to-end multimodal translation models are suitable for multiple scenarios, from image description and video subtitle production to emotion recognition. Moreover, they are widely used in various intelligent terminals and human–computer interaction systems.

4.3. **Multilevel Interactive Optimisation and Training.** When processing different types of data in multimodal machine learning, interacting with data across multiple levels, such as visual, language, audio and knowledge, and analysing the data at each level are often necessary. Owing to the differences in information amongst different levels and the lack of correlation between them, multi-layer modelling often finds it difficult to achieve global and local optima simultaneously. The method of multilevel interactive training has become an important approach to solving this type of problem.

4.3.1. *Theme Thesaurus Structure Model for Multilevel Interactive MT.* Assuming a training sample for MT exists, its definition formula is as follows:

$$A = \{Q_{ab}(i)W_{ab}(i)R_{ab}(i)\}\,. \tag{1}$$

In the formula, $Q_{ab}(i)$ is the information input vector, $W_{ab}(i)$ is the set of semantic correction vectors during the translation process and $R_{ab}(i)$ is a frequent itemset with semantic autocorrelation. Their respective calculation formulas are as follows:

$$Q_{ab}(i) = \frac{t_{ab}(i) - \alpha t_{ab}(i)}{t_{ab}(i-1)}. \tag{2}$$

$$W_{ab}(i) = \frac{|t_{ab}(i) - \Delta t(i)|}{t_{ab}(i)}. \tag{3}$$

$$R_{ab}(i) = exp\left[-u\left[x_a(i) - x_b(i)\right]^2\right] \tag{4}$$

In the formula, $t_{ab}(i-1)$ is the output of MT's topic word information, and α is the probability of correct translation.

A MT thesaurus structure model is constructed based on the above analysis, as shown in Figure 4.

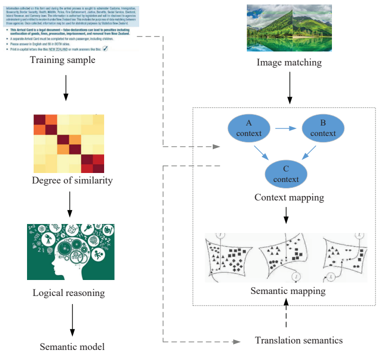


Figure 4. MT thesaurus structure model

The thesaurus structure model of multilevel interactive MT has been widely used in the field of computer language translation and has achieved fruitful results. This model has many innovations in semantic description, information expression and inference. This model also achieves the extraction of semantic associations between words and phrases in the corpus by introducing training samples and calculating their similarity. The tuple method is used to establish the mapping relationship between ontology and knowledge base to enhance the overall expression ability of the system, enabling machines to provide further accurate text input to users from a massive corpus, thereby laying a solid foundation for human–machine interaction based on multimodal knowledge.

The gradient descent algorithm is used in multimodal interactive translation models to minimize the loss function and optimize parameters. At each training step, update the parameters to reduce the loss by calculating the gradient of the loss function with respect to the parameters. This prompts the model to gradually adjust weights and improve performance. The learning rate determines the parameter update step size. Gradient descent ensures that the model better adapts to multimodal data and achieves accurate translation results.

4.3.2. *MT Automatic Evaluation.* The evaluation of MT quality is a complex process that needs to be considered from multiple aspects, and manual evaluation is the most important and effective way. However, manual participation in evaluation requires a significant investment of time, energy and resources, including considerable translations and diverse language types. Innovative algorithms and efficient automated alternative solutions are needed to address these limitations. Automatic evaluation technology can provide new solutions to this problem. The commonly used automation evaluation indicators include [25] bilingual evaluation understudy (BLEU), translation error rate (TER) and recall-oriented understudy for gisting evaluation (ROUGE).

BLEU is the most common automation evaluation indicator [26]. It uses the n-ary grammar model method to compare the similarity between MT and reference texts and then obtains a relatively accurate result. The BLEU index evaluates translation quality by comparing the lexical overlap between machine translation results and reference translations. The parameter N (1 to 4) represents the length of the matched phrase, and the higher the score, the better the translation. BLEU considers both precise matching and n-gram matching, reflecting the accuracy and fluency of translation. The calculation formula is as follows:

$$\mathrm{BLEU} = BP \times \exp(\sum_{n=1}^{N} W_n \log P_n), \tag{5}$$

where P_n is the improved n-gram accuracy value, N is the length of $N - gram$ and BP is the penalty factor. If the length of the translation is less than the shortest reference translation, then BP is less than 1, and the specific expression is as follows:

$$BP = \begin{cases} 1 & lc > lr \\ \exp(1 - \frac{lr}{lc}) & lc < lr \end{cases}, \tag{6}$$

where lc is the length of MT, and lr is the length of the shortest reference translation sentence.

In the log field, the rating effect of BLEU is as follows:

$$\log \mathrm{BLEU} = \min(1 - \frac{lr}{lc},\ 0) + \sum_{n=1}^{N} W_n \log P_n. \tag{7}$$

In Baseline, $N = 4$; W_n is the uniform weight; $W_n = \frac{1}{n}$.

Although BLEU is simple and easy to implement, it cannot meet complex NLP requirements and cannot effectively describe the relationships amongst distant discontinuous words. However, the ROUGE method can integrate considerable information in a short period and automatically perform semantic analysis on these texts.

ROUGE-L reflects semantic coherence and sentence structure similarity by evaluating the longest common subsequence between MT output text and reference target. Based on the similarity between the measurement reference translation and MT output of the longest common subsequence, this study defines reference translation as P and MT output as Q. The similarities are as follows:

$$S_1 = \frac{LCS(P,\ Q)}{x}. \tag{8}$$

$$S_2 = \frac{LCS(P,\ Q)}{y}. \tag{9}$$

ROUGE-L can be represented as follows:

$$F = \frac{(1+\mu^2)S_1 S_2}{S_1 + \mu^2 S_2}. \tag{10}$$

In the formula, $LCS(P,Q)$ is the longest common subsequence between the reference translation and MT output, and μ is the relative weight. When $\frac{\partial S}{\partial S_1} = \frac{\partial S}{\partial S_2}$, its expression is as follows:

$$\mu = \frac{S_2}{S_1}. \tag{11}$$

The implementation of a multimodal interaction model includes the following key steps. Firstly, determine the model architecture, including the processing methods for text and image inputs, attention mechanisms, etc. Then, adjust the parameters and optimize the model parameters through feedback from experiments and validation sets to ensure optimal performance. The training process includes providing labeled data to the model and using optimization algorithms such as gradient descent for training. During the training process, monitor the loss function and performance metrics to ensure that the model learns effective representations. Regularly evaluate the performance on the validation set to avoid overfitting. Finally, evaluate the generalization performance of the model using test data.

5. **Evaluation and Evaluation of MT and NLP.**

5.1. **Experimental Design and Dataset Construction.**

To evaluate the effectiveness of pure text MT and multimodal MT, a dataset that can reflect actual application scenarios to ensure its accuracy and reliability should be established. For pure text MT, this article mainly utilises existing public datasets, such as the Workshop on Machine Translation (WMT) or International Workshop on Spoken Language Translation (IWSLT), with the main goal of completing language tasks in Germany, English and France. For multimodal MT, in addition to text, other forms of information also need to be considered, such as image description translation MSCOCO, Flickr30K, Multi30k, audio content translation MuST-SHE and IWSLT21-ST-MUST-C. On this basis, combinatorial optimisation can be carried out for different types of machines (e.g., combining speech recognition with semantic understanding) to achieve high performance.

5.2. **Evaluation Indicators.** This study mainly measures pure text and multimodal MT output based on human evaluation indicators, such as language fluency, translation accuracy and cross-language consistency. The study then combines automatic evaluation indicators, such as BLEU, ROUGE and TER, to comprehensively evaluate machine understanding and ultimately obtain objective and comprehensive MT and NLP effects. Language fluency can be evaluated using language model metrics such as Perceptibility, and translation accuracy can be quantified through editing distance or Word Error Rate. Cross linguistic consistency can be measured by combining semantic similarity and cultural adaptability.

5.3. **Data Evaluation and Result Presentation.** According to task requirements and translation types, text resources that need to be translated and evaluated are selected, including source language text, MT output and reference translation. Amongst the automatic evaluation indicators, WMT14 is selected for the pure text MT test set, and Multi30k is selected for the multimodal MT dataset.

5.3.1. *BLEU.* Translation quality is quantitatively evaluated by comparing the similarity between machine-generated and reference translation results. In BLEU, the values of N are 1, 2, 3 and 4, so they are named BLEU-1, BLEU-2, BLEU-3 and BLEU-4, respectively. The value range of each indicator is [0,1]. The higher the score, the better the translation quality. Figure 5 shows the results when faced with the same source language text to be translated.

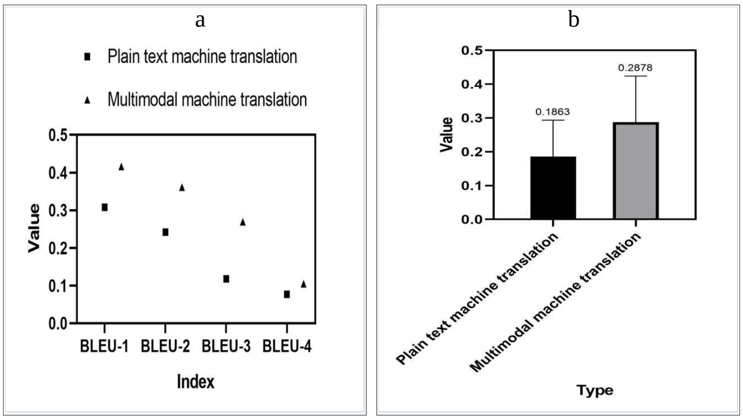


Figure 5. BLEU score comparison. Panel a. BLEU score distribution of two types of MT. Panel b. BLEU mean of two types of MT

In Figure 5a, the horizontal axis represents the four BLEU indicators, whereas the vertical axis represents the score of each indicator. Amongst them, blocks represent pure text MT, and triangles represent multimodal MT. Figure 5b shows the average BLEU indicators for the two translation methods. The horizontal axis represents two different MT modes, whereas the vertical axis represents the score. The performance of pure text MT in both images was not ideal. Amongst them, the scores of BLEU-1, BLEU-2 and BLEU-3 were 0.308, 0.242 and 0.118, respectively, which differed greatly from those of multimodal translation, reaching a maximum of approximately 0.151. In BLEU-4, the scores of the two were relatively similar, with values of 0.077 and 0.105, respectively. The difference value was only approximately 0.028. Based on the scores of the four BLEU indicators, the average BLEU of pure text MT was approximately 0.1863, whereas the average BLEU of multimodal MT was approximately 0.2878. Therefore, the results of multimodal MT had higher similarity and better quality compared with those of the reference translation. That is to say, plain text MT was more suitable for users to use on English websites, whereas multimodal MT could better support translation and communication amongst multiple languages and was also more suitable for fields with high requirements for speech recognition and NLP.

5.3.2. *ROUGE.* The size of the ROUGE could be used to measure the degree of matching between text summaries and automatically generated description text results and reference translations. The ROUGE indicators were mainly divided into three types: ROUGE-1, ROUGE-2 and ROUGE-L, and their scores when simultaneously executing an MT task were calculated. Figure 6 shows the results.

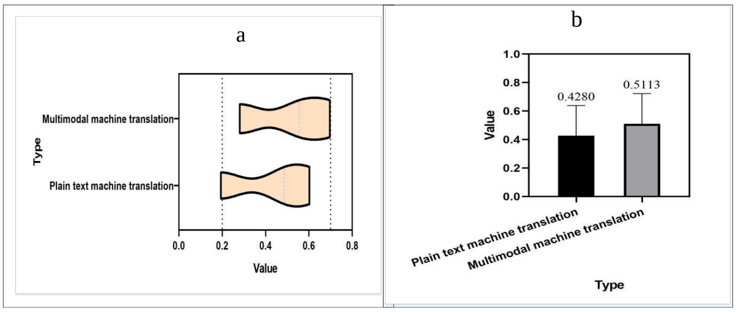


Figure 6. Comparison of ROUGE scores. Panel a. ROUGE score range for two types of MT. Panel b. The ROUGE mean of two types of MT

In Figure 6a, the horizontal axis represents the ROUGE score, with a ROUGE value range of [0,1]. The data tested in this article ranged from 0.1 to 0.7. The vertical axis represents two types of MT: pure text and multimodal. ROUGE-1 represents the number of matches between a single word in the translation result and the reference translation, including their overlap. ROUGE-2 also reflects the contextual relationship between the translation result and the reference translation. ROUGE-L measures the similarity between translation results and reference translations by comparing the length of the longest common subsequence between them. The score of multimodal MT ranged from 0.2 to 0.7, and that of pure text MT ranged from 0.15 to 0.65. Therefore, translation based on multilingual information showed great advantages and can provide further functions.

Figure 6b shows the comparison of the average ROUGE scores for two translation modes. The horizontal axis represents the type of MT, whereas the vertical axis represents the size of the average. After calculation, the average ROUGE score for pure text MT was approximately 0.428, and the average ROUGE score for multimodal MT was approximately 0.5113. The multimodal form was approximately 0.0833 higher than the pure text form. By evaluating BLEU and ROUGE indicators, multimodal MT had good decoding quality and improved translation accuracy to a certain extent. In practical life, the promotion and use of this translation system should be strengthened.

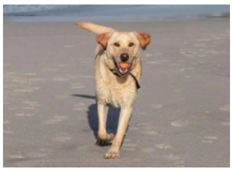
5.3.3. *TER.* As an automated evaluation indicator, TER is mainly used to compare the degree of difference between MT results and reference translations. The lower the score, the smaller the difference between translation results and reference translations. TER can accurately translate the source language into the target language and match it with the real reference text. Three source language texts were randomly selected for testing, and Tables 4 and 5 show the TER scores of the two MT results.

Based on the above results, the mean TER scores of the two MT methods were plotted, as shown in Figure 7.

Table 4. TER scores of pure text MT

Reference translation	Machine translation output	TER score
It is raining outside.	XXXXXXX	0.5
I like to go for a walk in the park.	XXXXXXXXX	0.17
They are having dinner together.	XXXXXXX	0.25

Table 5. TER scores for multimodal MT

Image description	Reference translation	Machine translation output	TER score
	A yellow dog playing on the beach.	XXXXXXXXXX	0.33
	Three children are playing a game under the tree.	XXXXXXXXXXX	0.17
	This is a blue car.	XXXXXXXXX	0.14

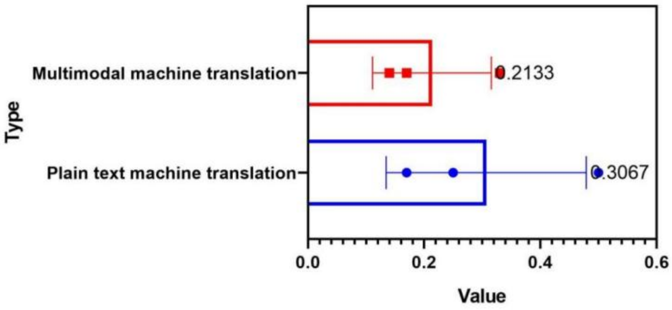


Figure 7. Mean TER scores of two MT methods

In Figure 7, the horizontal axis represents the score of TER, whereas the vertical axis represents the translation type. Amongst them, blue represents pure text MT, and red represents multimodal MT. The distribution of TER values for several text translation results was represented by dots and squares. In both MT results, two texts have TER values below the average. Moreover, the average value of multimodal translation was approximately 0.2133, whereas the average value of pure text translation was approximately 0.3067. A low TER score meant that the MT system could accurately translate the source language into the target language and match it with the actual reference text. Therefore, the translation that integrated image, voice and other information was more efficient than simple text translation.

5.3.4. *Metric for Evaluation of Translation with Explicit Ordering (METEOR).* METEOR is a MT automated evaluation method that comprehensively considers factors such as strategy, accuracy [27,28]and recall [29,30]. Compared with TER, METEOR can more comprehensively evaluate the effectiveness of translation systems. Assuming there are 49 texts to be translated, of which 32 are pure language texts, and 17 are image and audio texts, the METEOR scores of pure text and multimodal MT are calculated, as shown in Figure 8.

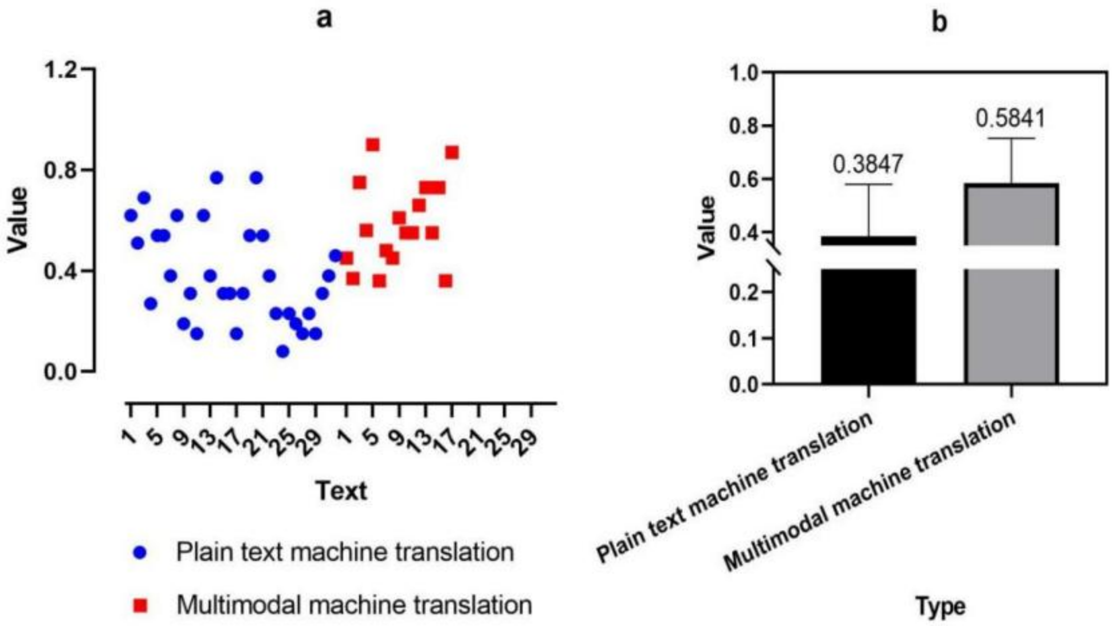


FIGURE 8. Meteor scores for two types of MT. Figure 8a METEOR score range for two types of MT. Figure 8b METEOR mean of two types of MT

The horizontal axis in Figure 8a represents the number of text quantities, whereas the vertical axis represents the METEOR score. Blue dots represent pure text MT, and red squares represent multimodal MT. Amongst them, the maximum value of pure text MT was 0.77, the minimum value was 0.08 and the maximum difference was 0.69. The maximum value of multimodal MT was 0.9, the minimum value was 0.36, and the maximum difference was 0.54. Therefore, the distribution of METEOR scores in multimodal MT was concentrated, and the translation results were accurate and coherent.

In Figure 8b, the horizontal axis represents the translation type, and the vertical axis represents the average METEOR score. Evidently, the average for pure language text translation was approximately 0.3847, whereas the average for multimodal translation was approximately 0.5841. Therefore, multimodal MT could collect further contextual information and language features and help improve MT technology and translation quality to promote continuous improvement of MT products and provide guidance for the development of products.

In addition to automated evaluation indicators, this study also designed a manual evaluation experiment to evaluate three multimodal MT and two pure text MT. A total of 100 Chinese sentences were randomly selected from a test corpus of 500 sentences. Each source sentence was paired with each translation to obtain a total of 50 pairs of Chinese source sentences and English translations. These translation pairs were randomly sorted on the webpage to disperse the translation of each source sentence. All manual evaluators used the same webpage and view sentence pairs in the same order. Translation scores range from 0 to 5, with 0 being very poor, 5 being best and 2.5 or above being qualified. W1 and W2 are named as two traditional pure text MT, and S1, S2 and S3 are named three multimodal MT. These 50 samples are evaluated in terms of language fluency and cross-language consistency.

In the evaluation, pure text machine translation (W1, W2) scored lower on BLEU, ROUGE, TER, and had average language fluency. Multimodal machine translation (S1, S2, S3) performs better on these metrics, especially in terms of cross language consistency and language fluency. S1, S2, and S3 are more in line with human perception, providing a vivid and intuitive translation experience. Overall, multimodal machine translation excels in handling word selection, grammar, and cultural differences, providing more flexible and scalable translation solutions for multi domain applications.

5.3.5. *Language Fluency.* Language fluency mainly focuses on conversion accuracy and expression clarity amongst different languages, and Figure 9 shows the comparison results.

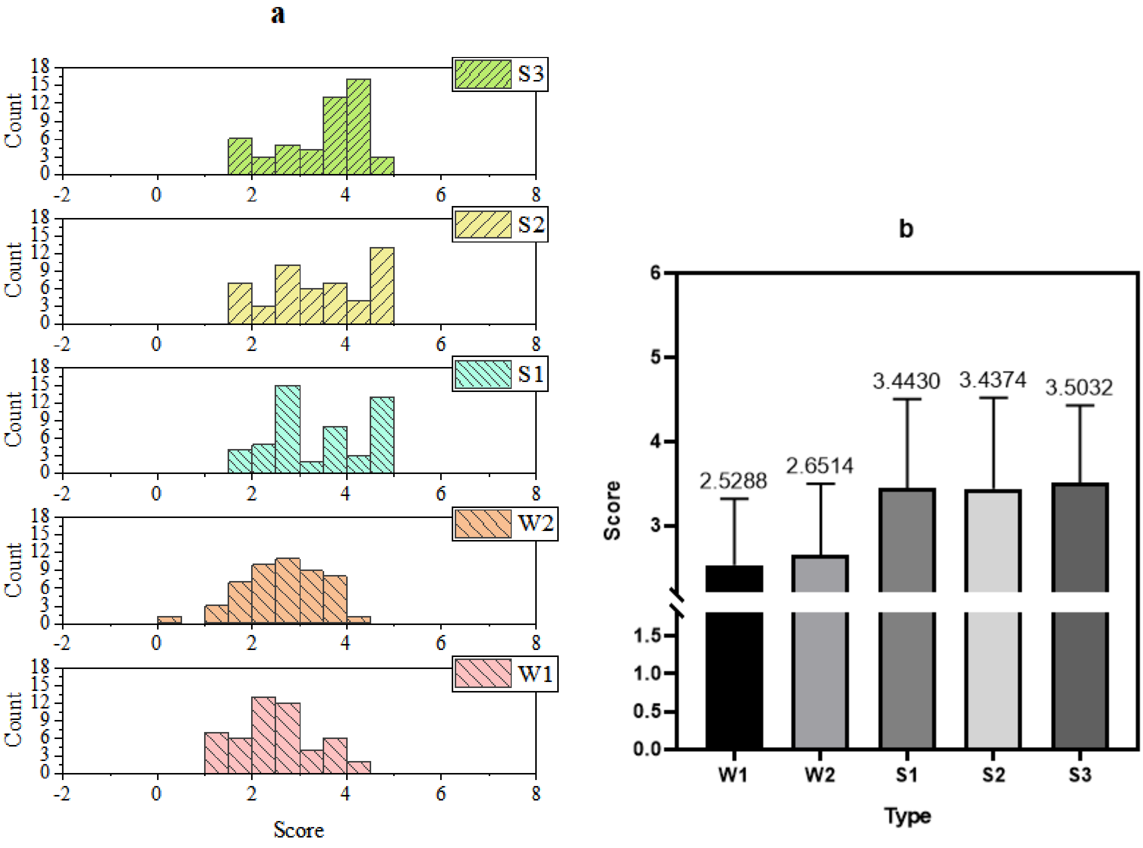


FIGURE 9. Comparison of language fluency between two types of MT. Panel a. The distribution of language fluency scores for two types of MT. Panel b. Mean language fluency scores for two types of MT

The horizontal axis in Figure 9a represents the score, whereas the vertical axis represents the number of samples within each score interval. Evidently, the distribution of the bar charts in W1 and W2 was more to the left of the coordinate axis than S1, S2 and S3, with even scores within the 0–1 range in W2. Significantly more samples exist within the

4–5 interval in S1, S2 and S3. This case indicated that MT types such as S1, S2 and S3 were in line with human sensory cognition, providing users with a vivid and intuitive translation experience.

The horizontal axis of Figure 9b represents two types of MT models with a total of five, whereas the vertical axis represents the average smoothness evaluation of 50 translation samples completed by each model. The average values of W1 and W2 on the left side were approximately 2.5288 and 2.6514, respectively, barely reaching the qualified level. The average values of S1, S2 and S3 on the right side were approximately 3.443, 3.4374 and 3.5032, respectively. This result proved that this type of multimodal MT could integrate multimedia data, such as images, videos and sounds, into the translation process, effectively increasing the interaction and communication between the source and target languages, thereby enhancing the coherence of translation results and improving translation fluency.

5.3.6. *Cross Language Consistency.* The information of text, images, audio and video should be consistent with the language translation results to avoid errors caused by integrating modal data. Figure 10 shows the specific evaluation results.

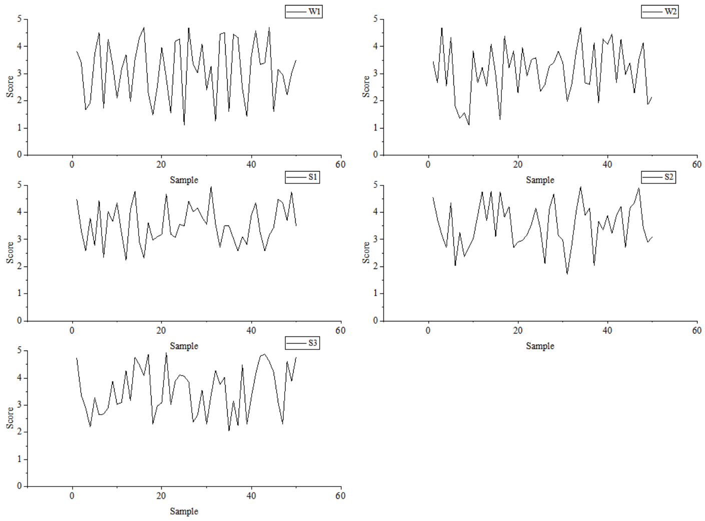


Figure 10. Cross-language consistency scoring results for two types of MT

The horizontal axis of each line chart represents the number of samples. The vertical axis represents the score and the tick marks are all from 0 to 5. In this case, the scores of S1, S2 and S3 were more concentrated. The lowest scores of S1 and S3 were both above 2, and the lowest scores of S2 were also above 1.5. They could have high flexibility and scalability in different application scenarios. The lowest scores for W1 and W2 were both approximately 1 point, which may lead to the selection of words in the translation results not in line with the customary habits of the target language, thereby affecting the overall consistency of the translation. In summary, multimodal MT had a comprehensive understanding and diverse information sources and performed better in solving problems related to single vocabulary, word order, grammar and cultural differences, achieving a natural and authentic correspondence between the target and source language. In addition, in real-time voice translation scenarios, such as audio conferences and video calls, multimodal MT could simultaneously process audio and text inputs, providing further timely and accurate translation results.

6. **Conclusion.** Traditional text translation converts pure text from the source language into pure text from the target language. Although machine learning and NLP have made significant progress, some translation problems have been successfully solved, and the translation effect is not ideal. By contrast, MT based on multimodal and multilevel unified interaction can fully consider multiple data sources, making the translation results highly intuitive, expressive and in line with the needs of practical scenarios. This study introduced several commonly used data fusion methods and their characteristics and then classified and extracted features for the target sentences under multi-mode, multi-layer unified interaction mode. The study established a multilevel interaction architecture and its corresponding relational model and finally completed the design and implementation of a multi-mode MT system with the help of NLP technology, which had good support ability for different language corpora.

REFERENCES

[1] I. Rivera-Trigueros, “Machine translation systems and quality assessment: a systematic review,” *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593-619, 2022.

[2] S.-Z. Wu, D. -D Zhang, Z.-R Zhang, N. Yang, M. Li, M. Zhou, “Dependency-to-dependency neural machine translation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, PP. 2132-2141, 2018.

[3] D. Khurana, A. Koli, K. Khatter, S. Singh, “Natural language processing: State of the art, current trends and challenges.” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713-3744, 2023.

[4] H. Li, “Deep learning for natural language processing: advantages and challenges,” *National Science Review*, vol. 5, no. 1, pp. 24-26, 2018.

[5] J.-s. Su, J.-L. Zeng, D.-Y. Xiong, Y. Liu, M.-X. Wang, J. Xie, “A hierarchy-to-sequence attentional neural machine translation model,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 623-632, 2018.

[6] A. Galassi, M. Lippi, and P. Torroni, “Attention in natural language processing.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291-4308, 2020.

[7] J. Guo, H. He, T. He, T. He, M. Li, H.-B. Lin, “Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 845-851, 2020.

[8] P. Lohar, G.-D. Xie, D. Gallagher, A. Way, “Building Neural Machine Translation Systems for Multilingual Participatory Spaces,” *Analytics*, vol. 2, no. 2, pp. 393-409, 2023.

[9] T. B, C. Ahuja, L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2018.

[10] H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, “Clinical natural language processing in languages other than English: opportunities and challenges,” *Journal of Biomedical Semantics*, vol. 9, no. 1, pp. 1-13, 2018.

[11] C, Zhang, Z.-C. Yang, X.-D. He, L. Deng, “Multimodal intelligence: Representation learning, information fusion, and applications,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478-493, 2020.

[12] J. Gao, P. Li, Z.-K. Chen, J.-N. Zhang, “A survey on deep learning for multimodal data fusion,” *Neural Computation*, vol. 32, no. 5, pp. 829-864, 2020.

[13] Charbuty, Bahzad, A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20-28, 2021.

[14] Z.-D. Zhang, C. Jung, “GBDT-MO: Gradient-boosted decision trees for multiple outputs,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3156-3167, 2020.

[15] S. Parida, O. Bojar, S.-R. Dash, “Hindi visual genome: A dataset for multi-modal English to Hindi machine translation,” *Computacion y Sistemas*, vol. 23, no. 4, pp. 1499-1505, 2019.

[16] A. Fan, S. Bhosale, H. Schwenk, Z.-Y. Ma, A. El-Kishky, “Beyond English-centric multilingual machine translation,” *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4839-4886, 2021.

[17] E.-D. Vries, M. Schoonvelde, G. Schumacher, “No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications,” *Political Analysis*, vol. 26, no. 4 pp. 417-430, 2018.

[18] D. Anamika, G.- K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85-112, 2020.

[19] G.-W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future," *Journal of Cognitive Neuroscience*, vol. 33, no. 10, pp. 2017-2031, 2021.

[20] Y.-R. Ji, Z.-H. Zhou, H. Liu, R.-V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no.15, pp. 2112-2120, 2021.

[21] Gogineni, K. Ajay, S. Swayamjyoti, S. Devadatta, K.-K Sahu, K. Raj, "Multi-Class classification of vulnerabilities in Smart Contracts using AWD-LSTM, with pre-trained encoder inspired from natural language processing," *IOP SciNotes*, vol. 1, no. 3, pp. 035002, 2020.

[22] L.-N. Vieira, O. Minako, O. Carol, "Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases," *Information, Communication & Society*, vol. 24, no. 11, pp. 1515-1532, 2021.

[23] Mehrer, Johannes, C.-J. Spoerer, N. Kriegeskorte, T.-C. Kietzmann, "Individual differences among deep neural network models," *Nature Communications*, vol. 11, no. 1, pp. 5725, 2020.

[24] G. Burak, "A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images," *Annals of Operations Research*, vol. 328, no. 1, pp. 617-641, 2023.

[25] Chauhan, Shweta, P. Daniel, A. Mishra, A. Kumar, "Adableu: A modified BLEU score for morphologically rich languages," *IETE Journal of Research*, vol. 69, no. 8, pp, 5112-5123, 2023.

[26] N. Han, "Google, Youdao Neural machine translation System Chinese English Translation Evaluation," *Journal of Shanxi Institute of Energy*, vol. 31, no. 5, pp. 123-124, 2018.

[27] Pennycook, Gordon, M. Jonathon, Y.-Z. Zhang, J.-G. Lu, D.-G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychological Science*, vol. 31, no. 7, pp. 770-780, 2020.

[28] J. John, D. Hassabis, "Protein structure predictions to atomic accuracy with AlphaFold," *Nature Methods*, vol. 19, no. 1, pp. 11-12, 2020.

[29] Bernstein, H. Michael, G.-L. Baird, A.-P. Lourenco, "Digital breast tomosynthesis and digital mammography recall and false-positive rates by time of day and reader experience," *Radiology*, vol. 303, no.1, pp. 63-68, 2022.

[30] Z.-Q. M, Z. Lu, R.-J. Brooke, M.-M. Hudson, Y. Yuan, "A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve," *Diagnostic and Prognostic Research*, vol. 5, no. 1, pp. 1-15, 2021.