International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 4728-4735

https://doi.org/10.70517/ijhsa464316

Multi-objective optimization algorithm for English translation information fusion and its big data implementation

Wenjing Jiao[1,*]
[1] English Department, Hegang Normal College, Hegang, Heilongjiang, 154107, China
Corresponding authors: (e-mail: jiaowenjing202412@163.com).

Abstract With the deep integration of big data technology and artificial intelligence, the field of English translation is experiencing a paradigm shift from single text processing to multimodal information fusion. In this paper, we design a multimodal English translation model based on Transformer, which fuses cross-modal information through the source language sequence encoding layer. Using the pre-trained cross-modal model CLIP to extract graphic features, combined with the dynamic contextual visual guidance vector mechanism, the adaptive fusion of text and image information is realized. In English-Chinese translation, the perplexity of the model in this paper is only 1.36 after 100 rounds of iterations, and the correct rate reaches 99.80%. The control model 1 still has a perplexity of more than 10, and the correct rate is only 94.49%. In Chinese-English translation, the model in this paper also achieves optimal results. The model's translation performance on the six sets of parallel corpus, including English-German, English-French, and Chinese-English, is significantly better than that of the baseline model, with the highest improvement of BLEU value of 16.89, which verifies the effect of multimodal information enhancement on the improvement of translation quality and context adaptation.

Index Terms English translation, Transformer, CLIP, multimodal information

I. Introduction

With the deepening of digital globalization and cross-border travel, the global demand for real-time English translation on a single day has been increasing, and the translation services supplied by enterprises have also risen [1], [2]. Although, in recent years, the research on translation has made good progress from theory to practice. However, with the advent of the era of big data, the theory and practice of translation has also changed from the traditional information restricted single state to the state of data and information sharing, which in turn has a certain impact on the theory and practice of translation [3]-[5]. And the current English translation, there is still the problem that the conversion between English and other languages is not smooth and unreasonable, which leads to some common-sense errors, or the translated sentences are hard, not quite in line with the professional standards of the industry, and can not correctly express the relevant meanings, showing that the translation level is relatively low [6]-[8]. While traditional statistical machine translation relies on a large number of bilingual corpora for translation, the accuracy of English translation in specialized fields is poor due to idioms, cultural differences, ambiguity, and the lack of small-language materials [9], [10]. These phenomena highlight the urgency of English translation innovation.

With the development of big data technology, the construction and application of corpus and English translation will move towards more intelligent and personalized. The open corpus resources in the Internet are rich, containing hundreds of languages and their linguistic and cultural backgrounds and idiomatic characteristics, and the user-generated content in the network is an important linguistic resource, through the big data technology to analyze and store these linguistic resources, keep pace with the times at the same time to obtain a large-scale corpus resources, for the optimization of the English translation to provide a direction of progress [11]-[13].

In this paper, we firstly share the parameters of the cross-modal encoder with the text encoder of the MNMT model through a multi-task learning approach. A cross-modal encoder sharing mechanism based on CLIP is proposed to realize the deep alignment of visual semantics and linguistic representations through text-image block segmentation and multi-head attention fusion. Dynamic contextual visual guidance vectors are introduced to enhance the model's real-time reasoning ability for graphic-text associations. The effectiveness of the improvement strategy in this paper is explored through ablation experiments. Verify the generalization performance of the proposed model based on massively parallel corpus with multilingual data enhancement strategy.

II. A Multimodal English Translation Model Integrating Visual Features and Semantic Information

Under the dual drive of globalization and digitization, language translation technology urgently needs to break through the limitations of pure text processing to cope with the problems of semantic ambiguity and lack of context in complex cross-cultural scenarios. Multimodal machine translation refers to the fusion of information from other modalities with text information on the basis of pure text corpus translation to improve the quality of machine translation. The current mainstream neural machine translation model relies on the local contextual attention mechanism, which is difficult to effectively capture the complementary information of the graphic and text, resulting in a bottleneck in the translation results in terms of cultural imagery conveyance and scene adaptability. Aiming at the above problems, this paper carries out the research work on the integration of English translation and information technology driven by big data.

II. A.Source Language Sequence Encoding Layer

The encoder is a stack of L identical layers, each of which is connected by two sub-layers, where the first sub-layer is a multi-head self-attention module, which is mainly responsible for performing the learning of the semantic correlation between the current text word and the rest of the words, while the second sub-layer is a location-based feed-forward neural network, which is mainly responsible for performing the transformation of the feature dimensions.

The outputs of the first sublayer are connected after being processed by residual processing and layer normalization, and then the source language context feature H_L is obtained by layer stacking computation, corresponding to the computational formula:

$$\bar{H}^l = LN(ATT^l(Q^{l-1}, K^{l-1}, V^{l-1}) + H^{l-1}) \tag{1}$$

$$H^l = LN(FFN^l(\bar{H}^l) + \bar{H}^l) \tag{2}$$

Among them, $ATT^l()$ represents the self-attention module of the l layer, $LN()$ represents the layer normalization operation, $FFN^l()$ represents the feedforward neural network of the l layer, Q^{l-1} represents the query vector obtained by the linear transformation of the context features of the source language of the $l-1$ layer, K^{l-1} represents the key vector obtained by the context features of the source language of the $l-1$ layer through linear transformation, V^{l-1} represents $l-1$ A value vector obtained by a linear transformation of the context features of the layer source language. Finally, the source language context feature H_L is obtained by stacking the encoding end.

Through the multi-task learning method, the parameters of the cross-modal encoder in the cross-modal entity reconstruction method are shared with the text encoder of the MNMT model, so as to achieve the purpose of fusing the cross-modal information into the MNMT model.

II. B.Text-image feature extraction

Computational vision systems predict the categories of a set of image information by means of a trained model (e.g. 1000 categories for ImageNet, 80 categories for COCO). This type of supervision is more limited because it requires additional labeling data, which limits generalization capabilities. It also requires additional data collection if there are new categories. Learning directly from the raw text associated with the images is a promising alternative that can utilize a wider range of supervised resources. Given some text descriptions and some images, predict which caption matches which image. This pre-training task is an efficient and scalable way to perform these tasks.

We use the pre-trained model CLIP for text and image feature extraction. There are two inputs, image and text, the dimension of the image is $[n,h,w,c]$, n scalars the number of images, h denotes the height of the image, w denotes the width of the image, and c denotes the number of channels in the image. The dimension of the text is $[n,l]$, n denotes the number of texts, and l is the length of the sequence.

A standard converter encodes the sequence as input and then divides the image into chunks when processing the image. Given an image of $H \times W \times C$ and a block size P, the image can be divided into N chunks with N being $H \times W / (P \times P)$. After obtaining the chunks, they are converted to D-dimensional feature vectors using a linear transformation, and then added to the position coding vectors. Like BERT, the image encoder also adds a flag bit $[class]$ before the sequence. The image encoder input sequence Z is shown in the following equation, where b denotes an image block.

$$Z_0 = [b_{cls}; b_p^1 E; b_p^2 E; \cdots; b_p^N E;] + E_{pos}$$
$$E \in \square^{P^2 \cdot c} \quad E_{pos} \in \square^{(n+1) \times D} \tag{3}$$

The coding model is basically the same as Transformer in that the input sequence is passed into the encoder and then the final features are output using $[cls]$-chi bits. The image encoder mainly consists of Multihead Self-Attention (MSA) and Multilayer Perceptron (MLP), and the two-layer fully connected network uses the RELU activation function. The image encoder has the following equation.

$$Z_l' = \mathrm{MAS}\left(LN\left(Z_{l-1}\right)\right) + Z_{l-1} l = 1 \cdots L \tag{4}$$

$$Z_l = \mathrm{MLP}\left(LN\left(Z_l'\right)\right) + Z_l' l = 1 \cdots L \tag{5}$$

$$F_{image} = LN(Z_l^0) \tag{6}$$

The text representation is generated from the text encoder in CLIP, whose text encoder is a Transformer structure. This is a 12-layer, 512-channel model with 8 attention heads.The activation processing in the highest layer of the Transformer is considered to be the feature representation of the text, which is denoted as $text_j$ for text T. After extracting the image representation $F^{img} = (img_1, img_2, \cdots, img_n)$ and the text representation T^{text}. The key step is similarity calculation.

Since the models in this paper are constructed based on pre-trained graphical models. In this paper, we use a parameter-free method as the averaging pool. Large-scale pre-training is performed by using CLIP on image-text pairs. The image representation F^{image} and the text representation T^{text} are layer normalized and linearly projected into the multimodal embedding space. The parameter-free type first uses average pooling

The parameterless type first uses average pooling to aggregate all the image features, and finally solves the similarity with the image features to obtain the image with the largest degree of resemblance to the text, defining the similarity function $sim(Image, Text)$ cosine similarity.

$$sim\left(\mathrm{Im}\, age, Text\right) = \frac{Text \cdot \mathrm{Im}\, age}{\left\|Text\right\| \left\|\mathrm{Im}\, age\right\|} \tag{7}$$

II. C.Image visual semantic and textual information fusion strategy

Neural Machine Translation (NMT) uses a sequence-to-sequence translation model based on encoding and decoding. Taking the input as a sequence of source language words $p = \frac{f-1}{2}$ and the output as a sequence of target language words $Y = (y_1, y_2, \ldots, y_n)$, the goal of the NMT model is to learn the model that maximizes the probability that Y is a given X, i.e., $P(X \mid Y)$. The way to merge other modal information is by manipulating the sequence of elements in a Transformer-based multi-head mechanism to form a new sequence. Each layer of the encoder layer consists of two sub-layers: multi-head attention and dot feedforward network. For a given source text sentence x, each word source is the sum of word embeddings and positional encodings.

$$PE(pos, 2i) = \sin(pos / 10000^{2i/d_{\mathrm{mod}\,el}}) \tag{8}$$

$$PE(pos, 2i+1) = \cos(pos / 10000^{2i/d_{\mathrm{mod}\,el}}) \tag{9}$$

Can be encoded using the same multilayer stacking approach, where each layer consists of two fully connected sublayers. In addition, batch residual connectivity and normalization are introduced for each sublayer. It should be noted that when iterating the customized layers, the list cannot be used directly, otherwise it may lead to the model weights and data inputs not being on the same device during the training process; meanwhile, when performing the model training, the updating of the parameters of each layer can easily lead to the instability of the computed values due to the drastic changes of the output layer. At this point, it is necessary to introduce batch normalization and use a small batch of mean and standard deviation to output the intermediate layers of the visual network to stabilize the stability of the output values of each layer of the neural network.

Each decoder consists of three sub-layers: masked multi-head attention, multi-head attention, and dot feedforward network. In this case, the query matrix Q, the key matrix K and the value matrix V in the multi-head self-attention function receive the output of the masked multi-head attention sublayer and the output of the

encoder as inputs, respectively. After the target is embedded, the output is embedded and positional coding is added to obtain the input of the decoder layer.

$$Attention(Q,K,V) = softmax(k)\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

MMT can more accurately translate “source text” into “target language” through images. In conclusion, this is important for the development of MMT and demonstrates the importance of visual features in further improving the accuracy of the predictive translation task. The scaled dot product attention process is shown in Fig. 1, and the approach in this paper is based on Transformer. The most important feature of the improved model is the introduction of contextual visual information guidance vectors, which are used to dynamically learn to generate multimodal context vectors for translation, a mechanism that allows the model to take full advantage of multimodal data when processing translation tasks. Specifically, the model is architecturally enhanced with a key addition to extract valid contextual visual information guidance vectors from the input visual content by capturing and fusing the correlations between text and images through a specially designed component.

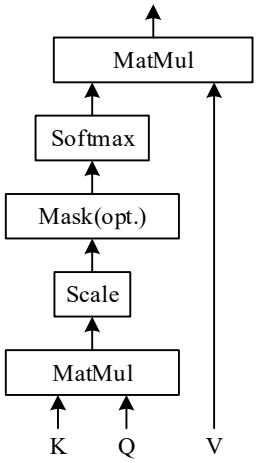


Figure 1: Zooming dot product attention

The generation process of this bootstrap vector is dynamic and adaptive, and it can be adjusted in real time according to specific input scenarios (e.g., graphic content), resulting in a highly relevant multimodal contextual representation. This representation not only reflects the grammatical structures and lexical meanings in plain text, but also integrates the unique contextual information conveyed by visual elements, such as character expressions, object locations, action indications, and other non-textual descriptions of important content.

In practice, the model guides translation decisions based on this multimodal context vector when generating translation results, ensuring that the translated text not only accurately corresponds to the literal meaning of the original text, but also aptly reflects the specific context in which the original text is embedded, thus greatly enhancing the quality of translation and the efficiency of cross-lingual and cross-cultural communication.

The architecture of the model is shown in Fig. 2. The decoder in this paper is an extension of the Transformer decoder, which takes the generated sequences as inputs, and the hidden state of the target layer generates the multi-head attention through the stacking of the LDs in the same layer, which enables the model to represent the subspace information at different locations as input $D(l-1) \in R$, where $D(0)$ is the source sentence in the $L-1$ th layer of the dimension, d_w is the dimension of the model, and $D(0)$ denotes the concatenation of all source words in the corpus:

$$H^{(l)}e = MultiHead(D^{(l-1)}, D^{(l-1)}, D^{(l-1)}) \tag{11}$$

$$H^{(l)}d = MultiHead(T^{(l-1)}, T^{(l-1)}, T^{(l-1)}), 1 \le l \le Ld \tag{12}$$

A trained model is introduced for experimentation and then predictions are made on the specified data:

$$MultiHead(K,Q,V) = Concat(head_1, \cdots, head_h)W_o \tag{13}$$

$$wherehead_i = Attention(QW_{iq}, KW_{ik}, VW_{iv}) \quad (14)$$

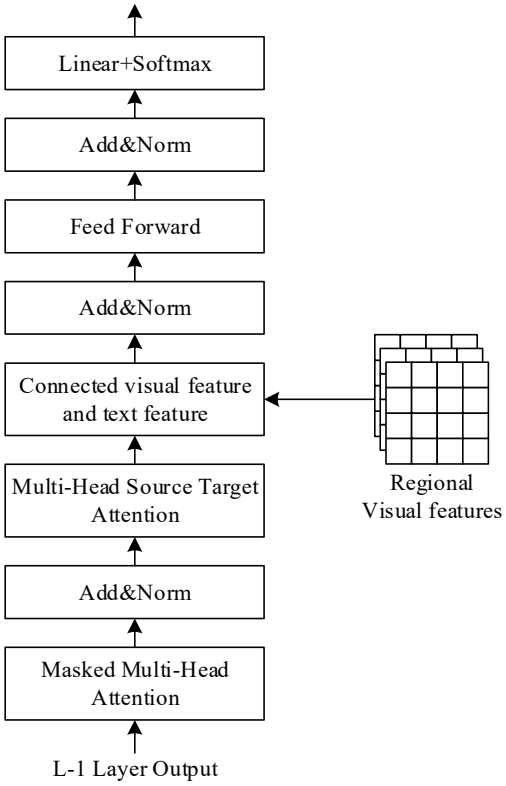


Figure 2: The architecture of the model

III. Analysis of the effectiveness of big data-driven English translation models

III. A. Experimental setup

The experiments in this paper are selected from the datasets of three parallel corpora of English 1 German, English 1 French, and English 1 Chinese in Muti30K, and we also achieve six sets of experiments by switching the source and target languages of the three sets of languages and translating them into each other to validate the performance of the model. During the experiments, the training stopping time is appropriately adjusted according to the results of each iteration with at least 100 rounds of iterations to achieve the best results.

In order to check the performance of the model, the proposed method is compared with two state-of-the-art baseline methods under the same parameter settings and they are briefly introduced.

MMT: Text-only pure Transformer model, which enables the encoder part to realize the interaction between text and image through the attention mechanism to enhance the richness of semantic information, thus improving the translation effect.

GSMT: A Transformer-based real-time machine translation model with a Poisson a priori approach to effectively balance translation quality and latency.

Agent: the agent strategy uses a fully trained translation model to create an agent for generating optimal segments in real-time machine translation.

III. B. Ablation experiments

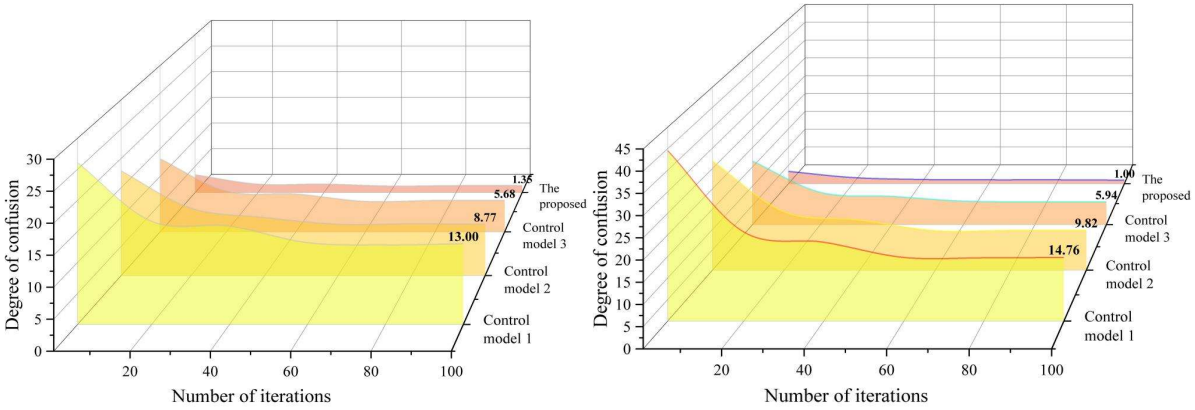
III. B. 1) Confusion level

The pure Transformer model MMT is set as the control model 1, the pre-trained model CLIP is used for text-image feature extraction on the basis of the control model 1 as the control model 2, and the contextual visual information guidance vector is introduced on the basis of the control model 1 as the control model 3.

Taking English one Chinese as an example, the results of the perplexity comparison of different models in English-Chinese translation and Chinese-English translation are shown in Fig. 3 (a~b). In English-Chinese translation, the model of this paper has the lowest perplexity, which is only 1.36 after 100 rounds of iteration, followed by the control model 3, and the control model 1 has the worst perplexity, which is still more than 10 in the end. In Chinese-English

translation, the model of this paper also achieves the optimal effect, and the confusion degree reaches 1.00 after 100 rounds of iteration.

The confusion degree of the three control models has decreased substantially compared with that of this paper's model, proving the effectiveness of this paper's improvement scheme.



(a) English-Chinese translation (b) Chinese-English translation

Figure 3: Comparison results of confusion

III. B. 2) Correctness

Taking English one Chinese as an example, the results of the comparison of the correct rates of different models in English-Chinese translation and Chinese-English translation are shown in Figure 4. In English-Chinese translation, the correct rates of the four models are 94.49%, 97.79%, 98.33% and 99.80% respectively, and the model of this paper is much higher than the three control models. In Chinese-English translation, the correct rate of this paper's model is 5.51%, 2.2% and 1.33% higher than that of the control model, respectively. The experiment verifies that text-image matching and image information integration are beneficial to translation information mining and can help improve the effect of machine translation.

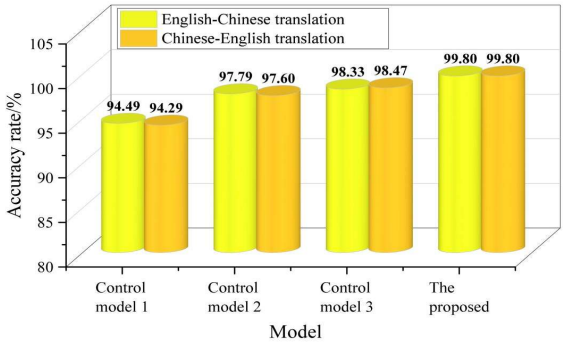


Figure 4: Comparison results of the accuracy rates of different models

III. C. Control experiments

The experimental results obtained by using parallel corpora of English-German, English-French, and English-Chinese are shown in Table 1. The ∞ in the table represent the traditional whole-sentence translation strategy, the plain-text translation model, and both models as a baseline were experimented with using whole-sentence translation. In this model experiment, the values of k were set to 2, 4 and 6 respectively, and the scores of BLEU and AL were compared according to different k values. " ↑ " means that the higher the value, the better, and " ↓ " means that the lower the value, the better. Compared to the text-only model, the model in this paper improves by up to 16.89 BLEU values. Compared with the two mainstream models of GSMT and Agent, the proposed model improves the BLEU values by up to 11.58 and 9.69 under the traditional whole sentence translation strategy, respectively.

Table 1: Experimental results under three parallel corpora

Model	k	English-German		English-French		English-Chinese	
MMT	∞	25.25	-	42.97	-	29.99	-
GSMT	∞	30.11	10.62	48.28	10.93	32.17	8.93
Agent	∞	32.78	10.46	50.17	10.72	33.18	8.61
The proposed	2	33.69	2.78	51.22	2.83	34.66	2.04
	4	34.77	4.11	53.29	4.26	35.07	3.85
	6	35.18	6.22	54.44	6.45	36.22	5.73
	∞	39.22	10.38	59.86	10.51	37.78	9.12

Not only that, this paper also conducted three more sets of experiments by swapping the source and target utterances of these three parallel sentence pairs to explore the translation performance of the model. The experimental results are shown in Table 2. Our model improves up to 12.82 BLEU values compared to the real-time machine translation model for plain text. Compared with the other two state-of-the-art methods, the model in this paper is still quite competitive for a similar range of AL.

Table 2: Experimental results after the swap of the three parallel corpora

Model	k	German-English		French-English		Chinese-English	
MMT	∞	27.71	-	40.65	-	28.13	-
GSMT	∞	32.66	10.98	46.25	11.47	30.37	10.22
Agent	∞	34.98	10.63	48.26	11.22	31.69	10.08
The proposed	2	35.82	2.84	50.42	3.87	32.95	2.01
	4	36.77	4.02	51.56	5.03	33.58	3.27
	6	37.96	6.47	52.84	8.22	34.46	5.92
	∞	38.83	10.41	53.47	10.94	36.02	9.18

IV. Conclusion

This study constructs a multimodal English translation model that integrates visual features and semantic information through multimodal fusion and big data analysis technology innovation.

The ablation experiments show that in English-Chinese translation, the model of this paper has the lowest perplexity, which is only 1.36 after 100 rounds of iteration, followed by the control model 3, and the control model 1 has the worst perplexity, which is still more than 10 in the final perplexity. In Chinese-English translation, the model of this paper also achieves the optimal effect, and after 100 rounds of iteration, the perplexity reaches 1.00. In English-Chinese translation, the correct rates of the four models are 94.49%, 97.79%, 98.33% and 99.80% respectively, and the model of this paper is much higher than that of the three control models. In Chinese-English translation, the correct rate of this paper's model is 5.51%, 2.2%, and 1.33% higher than that of the control model, respectively.

The control experiments show that the model in this paper improves up to 16.89 BLEU values compared with the text-only model under parallel corpus. Compared with the two mainstream models, GSMT and Agent, this paper's model improves at most 11.58 and 9.69 BLEU values under the traditional whole sentence translation strategy, respectively. After parallel corpus swapping, our model improves up to 12.82 BLEU values compared to the real-time machine translation model for plain text. Compared with the other two state-of-the-art methods, the model in this paper is still quite competitive under a similar AL range.

References

[1] Ekuerhare, A., & Udoka, O. P. (2024). The Effect of Globalization on Language Services and Translation Strategies. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 507-516.

[2] Bakhodirovna, T. P., & Qizi, I. S. Z. (2024). INTERNATIONAL TOURISM TERMS IN SIMULTANEOUS TRANSLATION. Science and innovation, 3(Special Issue 19), 705-708.

[3] Wang, H. (2019). The development of translation technology in the era of big data. Restructuring Translation Education: Implications from China for the Rest of the World, 13-26.

[4] Guo, Z., & Chen, C. (2021, August). Practice and research of computer-aided medical translation based on big data. In Journal of Physics: Conference Series (Vol. 2004, No. 1, p. 012014). IOP Publishing.

[5] Roig-Sanz, D., & Fólica, L. (2021). Big translation history: Data science applied to translated literature in the Spanish-speaking world, 1898–1945. Translation Spaces, 10(2), 231-259.

[6] Zaykova, I., & Shilnikova, I. (2019). Economic translation: Theoretical and practical issues. In SHS Web of Conferences (Vol. 69, p. 00139). EDP Sciences.
[7] Ameen, J. A. M., & Sherwani, K. A. (2023). Semantic Challenges of Business Translation with Reference to English and Arabic: A Product-Oriented Approach. journal of Language Studies, 6(3, 2), 1-13.
[8] Seljan, S. (2018). Quality Assurance (QA) of Terminology in a Translation Quality Management System (QMS) in the business environment. European Parliament: Translation Services in the Digital World, 92-145.
[9] Benkova, L., Munkova, D., Benko, Ľ., & Munk, M. (2021). Evaluation of English–Slovak neural and statistical machine translation. Applied Sciences, 11(7), 2948.
[10] Cuong, H., & Sima'an, K. (2017). A survey of domain adaptation for statistical machine translation. Machine Translation, 31, 187-224.
[11] Li, C. (2024, April). Construction of Translation Corpus and Training of Translation Models Supported by Big Data. In 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT) (pp. 981-986). IEEE.
[12] Ma, W. (2025). Optimizing English translation processing of conceptual metaphors in big data technology texts. Advances in Continuous and Discrete Models, 2025(1), 1-17.
[13] Hu, J. (2023). Analysis of the feasibility and advantages of using big data technology for English translation. Soft Computing, 27(16), 11755-11766.