J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 7135--7151

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Cross-domain Generalization Performance of a Neural Machine Translation System Incorporating Hierarchical Bayesian Models

Jing Li[1,✉]

[1] School of General Education, Hunan University of Information Technology, Changsha, Hunan, 410000, China

ABSTRACT

The study combines hierarchical Bayesian model and adversarial neural network according to the model architecture of neural machine translation, and introduces the domain generalization method based on cross-domain gating to solve the domain generalization problem, and constructs the neural machine translation system based on hierarchical Bayesian model. Translation performance experiments are conducted on this translation system to test the cross-domain generalization performance of the neural machine translation system based on hierarchical Bayesian model in this paper. The translation method of this paper significantly outperforms the baseline system of statistical machine translation in the direction of translation for all the intertranslated languages and medial languages of the European Parliament corpus. The statistical machine translation model and the standard neural machine translation model have maintained a stable performance during the growth of the interpolation coefficients, while the performance of this paper's hierarchical Bayesian-based neural machine translation system grows rapidly to the maximum when the interpolation coefficients grow to 0.3 or 0.4, and its overall average BLEU value always outperforms that of the statistical machine translation model and the standard neural machine translation model. The BLEU values of the hierarchical Bayesian-based neural machine translation system are 35.26% and 34.28% for bi-directional Chinese-English translation, and 26.42% and 25.96% for bi-directional Chinese-Western translation, which are better than those of the neural machine translation based on the attentional mechanism and variational scoring. And the hierarchical Bayesian-based neural machine translation system has strong stability on the translation of low-resource languages.

Keywords: neural machine translation, hierarchical Bayesian models, adversarial neural networks, cross-domain generalization methods

✉ Corresponding author.
E-mail address: ednaleelee@163.com (J. Li).
Received 12 January 2024; Revised 05 March 2024; Accepted 25 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-397

1. Introduction

In recent years, with the rapid development of artificial intelligence technology, machine translation systems play an important role in translation tasks between different languages [1-2]. However, traditional machine translation methods are usually based on statistical models, using phrase-based methods or template-based methods, which require manual feature extraction and parameter tuning, with limited effectiveness [3-5]. In contrast, neural machine translation, as a machine translation method using deep learning techniques, adopts an end-to-end training approach, does not need to manually design features, and is able to learn translation models directly from raw data [6-8]. However, due to the complexity and diversity among languages, neural machine translation systems often have difficulties and deficiencies when facing translations in some specific domains or languages. In order to solve this problem, generalization learning methods are introduced into neural machine translation systems to improve the performance and generalization of the system [9-12].

Generalization learning refers to the fact that the knowledge learned on a given task can be effectively applied one to other tasks without the need to relearn or adjust the parameters [13-14]. In machine translation, generalization learning methods learn rich linguistic knowledge and laws by training on large-scale multilingual corpora, thus improving the system's translation performance between different language pairs [15-17]. Through generalization learning, the neural machine translation system can better capture the syntactic, grammatical and semantic information between different languages, thus improving the translation quality and accuracy [18-20].

Literature [21] unfolds a review of Multilingual Neural Machine Translation (MNMT), pointing out that transfer learning helps to improve the quality of translation, that MNMT has more potential than statistical machine translation, and analyzes the existing literature and future directions of MNMT. Literature [22] proposes the use of a single neural machine translation model for translation between multiple languages, indicates the possibility of migration learning and zero-probability translation in neural translation, and aims to validate the existence of a universal inter-language representation of the model by showing a case study. Literature [23] aims to learn a model that can generalize well to unknown domains without training data and explores the problem of domain generalization for NMT, proposing a leave-one-domainout training method to simulate unknown data scenarios. Literature [24] describes the principles of neural machine translation systems and step-by-step introduces the key concepts used to describe these systems, including neural networks, learning algorithms, word embeddings, etc., in order to facilitate a better understanding of their inner workings and possibilities. Literature [25] proposes a new machine translation approach that can effectively improve the performance of the in-domain translation model and maintain or even improve the overall performance of the out-of-domain translation model, which is more complete than the existing approaches. Literature [26] emphasized that deep learning techniques have driven the development of Neural Machine Translation (NMT) models, pointed out the deficiencies of translation systems, and revealed that good domain adaptation can be achieved by fine-tuning the in-domain data. Literature [27] systematically analyzes the field of attention-based neural machine translation for multilingual communication, especially the revolutionary impact of the attention mechanism in improving cross-language comprehension, revealing the important role of the attention mechanism in multilingual communication. Literature [28] proposes a word-level domain-sensitive representation learning (WDSRL) approach that achieves substantial improvements on several representative baselines for multilingual pairs. Literature [29] explored the performance of neural machine translation models such as recurrent neural networks, long and short-term memories, and gated recurrent units on two different linguistic datasets, elaborating that the gated recurrent unit model had the highest test accuracy. Literature [30] describes activities that are considered “indirect tasks” in technical translation courses in order to promote the development of students' translation skills and to provide teachers with more resources to assess students' translation abilities.

In this paper, based on the neural machine translation model, we construct a hierarchical Bayesian model and an adversarial neural network, and integrate the two into the standard neural machine translation model to construct a neural machine translation system based on the hierarchical Bayesian model. To further improve the translation performance of the system, the domain generalization method based on cross-domain gating is used to solve the domain generalization problem in neural machine translation. The Hierarchical Bayesian-based Neural Machine Translation System is compared with the Statistical Machine Translation System and the Standard Neural Machine Translation System for corpus experiments to test the performance of each translation model. The hierarchical Bayesian-based neural machine translation system is compared with the neural machine translation methods based on the attention mechanism and variational scoring to further verify the superiority of the translation system in this paper. Finally, it is verified for generalizability.

2. Design of a Neural Machine Translation System Incorporating Hierarchical Bayesian

2.1. Model Architecture for Neural Machine Translation

Machine translation is a sequence-to-sequence (Seq2Seq) generative task [31], where given an input sentence $x_{1:S}$ in the source language, the translation model has to generate the corresponding sentence $y_{1:T}$ in the target language, which can be of different lengths S for the input sentence and T for the output sentence.

The goal of the Seq2Seq model is to estimate conditional probabilities:

$$p_\theta\left(y_{1:T} \mid x_{1:S}\right)=\prod_{t=1}^{T} \theta\left(y_t \mid y_{1:(t-1),x_{1:S}}\right) \tag{1}$$

where $y_t \in V$ is a word in word list V.

Given a set of training data $(x,y)\ldots$, maximum likelihood estimation can be used to train the parameters of a sequence-to-sequence model:

$$\max_\theta \sum_{n=1}^{N} \log_\theta\left(y_{1:T_n} \mid x_{1:Sn}\right) \tag{2}$$

After training, based on the input sentence x, the model generates the most likely translated sentence word by word:

$$\hat{y}=\arg\max_y \, p_\theta(y \mid x) \tag{3}$$

NMT, as a sequence-to-sequence task, is usually realized through an encoder-decoder architecture. The encoder encodes the input sentence as a fixed-length vector representation, and the decoder uses this vector representation as the input to incrementally generate the target sentence. The encoder-decoder architecture can flexibly and independently employ a variety of different neural networks as its own encoder or decoder.

2.2. Hierarchical Bayesian Model Construction

Hierarchical Bayesian models are based on exchangeability [32-33], and secondly the flexibility of the hierarchical Bayesian approach is that it is able to optimize complex models and does not require much model design or normality assumptions. Hierarchical Bayesian methods are also known as full Bayesian methods, in which it is first assumed that the parameters of the model obey a certain prior distribution, and then inference is made based on the posterior distribution.

Since the prior distribution of the hyperparameters in the second layer of the hierarchical Bayesian model is given by the empirical data or the subjective prior distribution of the parameters, it may be more difficult to make a direct inference on it, and it will also affect the estimation results of the hyperparameters λ. Therefore, in the process of the actual operation of the problem, it is also necessary to take into account the actual situation and analyze the pros and cons of the model according to the relevant areas of the problem under study and other factors before making the choice of the number of layers. Among them, the common layered model is generally two layers in most cases.

According to Bayes' theorem, there are:

$$f(\mu,\lambda \mid y) = \frac{f(y,\mu \mid \lambda) f(\lambda)}{f_1(y)} \tag{4}$$

where $f_1(y)$ is the a priori density of y and:

$$f_1(y) = \int f(y,\mu \mid \lambda) f(\lambda) d\lambda \tag{5}$$

The a posteriori density function is obtained from (4):

$$f(\mu \mid \lambda) = \int f(\mu,\lambda \mid y) d\lambda \tag{6}$$

$$f(\mu \mid y) = \int f(\mu \mid y,\lambda) f(\lambda \mid y) d\lambda \tag{7}$$

Here $f(\mu \mid y,\lambda)$ is commonly used for empirical Bayesian estimation inference, and Eq. (6) can be further simplified to the form of (7). It can be seen from Eq. (6) that $f(\mu \mid y)$ may involve the operation of high-dimensional complex integrals, which requires the MCMC method to solve such computational problems.

In a two-layer hierarchical Bayesian model, the expressions for the first and second layers can be represented by $f(y,\mu \mid \lambda_1)$ and $f(y,\mu \mid \lambda_2)$, where:

$$\lambda = \left(\lambda 1, \lambda_2^T\right)^T \tag{8}$$

From the Bayesian theory, the posterior distribution $f(\mu \mid y)$ can be obtained, and based on the posterior distribution $f(\mu \mid y)$, the estimates of other parameters are then inferred.

Compared with other models, the hierarchical Bayesian model is simpler, less demanding on data, and its estimation accuracy is more precise. The important thing in hierarchical Bayesian models is to specify the relevant prior distribution for parameter λ. $f(\lambda)$. The prior distribution for parameter λ may or may not be useful for estimation, as the prior information is the information gained from relevant experience and past knowledge before the overall sample size is obtained. However, in practice, especially in applications involving public policy, hierarchical Bayesian models seldom involve the use of prior information, which is intended to reflect the missing information of parameter λ, and a reasonable choice of prior information is one of the reasons for the accuracy of the posterior information.

In the vast majority of cases, much of the information about the prior distribution $\omega(x)$ is unknown, which also requires the researcher to make a more detailed description of the prior distribution before proceeding with the model estimation, and when parameterizing the parameters in the prior distribution, the distribution of the prior vector θ is first given a parameter λ to make it conditional on the prior distribution:

$$\theta \sim \omega_1(\theta \mid \lambda) \tag{9}$$

The parameter λ in the above equation is also unknown, and then let its prior distribution be:

$$\lambda \sim \omega(\lambda) \tag{10}$$

At this point, we consider λ as the hyperparameter of parameter θ under the hierarchical Bayesian model, i.e., a parameter θ is given to the prior distribution, and then another parameter λ

is given to parameter θ to make it a parameter that satisfies the condition, at this point, λ is the hyperparameter of parameter θ, and the prior distribution of λ is called the super-prior distribution, and the hierarchical Bayesian model is in fact a more detailed delineation of prior distributions, and a hierarchical hierarchy is made to classify the prior information layer by layer, so that its posterior density function can be obtained. The prior information in the model is graded layer by layer, and its posterior density function can be obtained as:

$$\omega(\theta \mid y) \propto f(y \mid \theta)_{\omega_1}(\theta \mid \lambda)_{\omega_2}(\lambda) \tag{11}$$

Since the researcher often focuses on the posterior distribution of parameter θ, and hyperparameter λ becomes a more variable in the calculation process, we often call hyperparameter λ a "useless parameter" in the calculation process using Bayesian model, and in order to obtain the marginal posterior density of parameter θ, it is necessary to perform an integral operation on the hyperparameter first, and at this time, the hierarchical prior density of θ can be expressed as:

$$\omega(\theta) = \int_{\omega_1} (\theta \mid \lambda)\omega_2(\lambda)d\lambda \tag{12}$$

The use of hierarchical Bayesian theory to construct the relevant model in this paper is to avoid the over-reliance of the estimation value on the prior distribution, so as to achieve the effect of enhancing the precision and accuracy of the model. In addition, compared with the single-layer prior distribution, the hierarchical prior distribution can accommodate certain unknown factors more, such as if in the acquisition of prior information encounters obstacles, the choice of hierarchical model is also our attention to the objective facts, the hierarchical model through the prior sample information to get the random effect of hyperparameters, and the direct estimation of the prior information of the method of the fixed effects of a combination of precise portrayal of the internal correlation of the data. The stratified Bayesian model, as a representative method of empirical Bayesian model, effectively combines Bayesian statistics and classical statistics, and its estimation results are also far more accurate than the estimation precision of traditional sampling survey under the condition of small samples.

After obtaining the posterior distribution, then use the MCMC method for sampling to obtain the estimated value of the parameter, which is divided into three steps: firstly, extract the sample value of θ. $\theta^t, t = 0,1,\ldots n$. Secondly, randomly select the number from the uniform distribution $U(0,1)$. c. Finally, if the range of values of the random number is greater than 1, less than the value of the prior density of the adjacent layers, by the nature of Markov chain have: $\theta(t+1) = \theta^{(t)}$, otherwise, $\theta(t+1) = \theta^{(t+1)}$, repeat n times, you can get the sampling sample of the estimated parameter.

In general, the two-level hierarchical model is as follows:

First level:

$$y_{1j} = \beta_{0j} + \sum_{p=1}^{p} \alpha_p^x p1j + \sum_{q=1}^{Q} \beta_{\text{qj}} z_{q1j} + e_{1j} \tag{13}$$

Second layer:

$$\beta_{0j} = \gamma_{00} + \sum_{m=1}^{M} \gamma_{0m}\omega_{mj} + u_{0j} \tag{14}$$

$$\beta_{1j} = \gamma_{10} + \sum_{m=1}^{M} \gamma_{1m}\omega_{mj} + u_{1j} \tag{15}$$

……

$$\beta_{Qj} = \gamma_{Q0} + \sum_{m=1}^{M} \gamma_{Qm}\omega_{Qj} + u_{Qj} \tag{16}$$

In the above equation, the stochastic regression coefficient $(\beta_{0j}, \beta_{1j}, \ldots, \beta_{Qj}, q = 1, \ldots, Q)$ in the first layer is defined as a linear function of the M level second layer variables Ωmj, and layer 1 can then be transformed into a matrix based on this function:

$$Y = X\alpha + Z\beta + e \tag{17}$$

In Eq. (16), vector Y represents the response variable. X is the matrix of fixed effects α. Vector e represents the error in the first layer. The residual term e_{1j} of the model in the first layer represents the within-group variation, and the residual term $(u_{0j}, u_{1j}, \ldots, u_{Qj})$ in the second layer represents the between-group perturbations of the random regression coefficients in the first layer, respectively.

2.3. *Adversarial Neural Network Construction*

The adversarial neural network model in this paper is divided into two parts: generator and discriminator [34], and its structure is schematically shown in Fig. 1. Among them, the generator is an adversarial neural network translation model that can generate translated sentences. The discriminator, on the other hand, can be regarded as a discriminative model that can evaluate the translated sentences. However, at this time, it is still necessary to carry out gradient training for the generator to improve the translation accuracy.

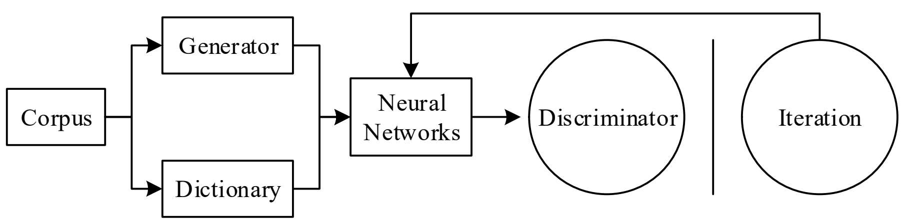


Fig. 1. Schematic diagram of countermeasure neural network

1) In the generator part, the underlying network uses an RNN model based on recurrent neural networks. The model consists of two parts, the encoder and the decoder, which performs encoding operations on the input sentence data sequence and generates the propagated states as shown in equations (18) and (19):

$$\vec{h} = \left(\overrightarrow{h_1}, \cdots, \overrightarrow{h_m}\right) \tag{18}$$

$$\overleftarrow{h} = \left(\overleftarrow{h_1}, \cdots, \overleftarrow{h_m}\right) \tag{19}$$

The decoder uses RNN network to decode the target data, and the decoded data is the current state and hidden state data set generated by the encoder.

2) The discriminator uses the convolutional neural network as the base network, which mainly discriminates the translation data generated by the generator. Since the length of the data generated by the generator is closely related to the input corpus to be translated, the discriminator will first serialize the sentences. Assuming that the default length of the sentence is set to T, the matrix to be built is:

$$X_{1:T} = x_1; x_2; \cdots; x_T \tag{20}$$

$$Y_{1:T} = y_1; y_2; \cdots; y_T \tag{21}$$

Once the building of the matrix is complete, the data is convolved to compute the vector values for each convolution kernel. Then pooling operation is performed on this vector value to get the feature vector as shown below:

$$c_x = \left[\varepsilon_1, c_2, \cdots, c_{T-I+1}\right] \tag{22}$$

The feature vectors in the target data matrix are finally extracted and the probability of the vectors is calculated as:

$$p = \varphi\left(V\left[\underline{E}_x ; c_y\right]\right) \tag{23}$$

3) Gradient training. After the generator generates the sequence, the generator needs to be trained with gradient with the objective function equation:

$$J(\theta) = \sum_{Y_1:T} G_\theta\left(Y_{1:T} \mid X\right) \cdot R_{D \cdot Q}^{G_\theta}\left(Y_{1:T-1}, X, y_T, Y^*\right) \tag{24}$$

And the feedback of the discriminator is calculated as:

$$\begin{aligned} & R_{D \cdot Q}^{G_\theta}\left(Y_{1:T-1}, X, y_T, Y^*\right) \\ & = \begin{cases} \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \lambda\left(D\left(X, Y_{1:T}^n\right) - b\left(X, Y_{1:T}^n\right) + (1-\lambda) Q\left(Y_{1:T}, Y^*\right)\right) \\ \lambda\left(D\left(X, Y_{1:T}\right) - b\left(X, Y_{1:T}\right) + (1-\lambda) Q\left(Y_{1:T-1}, Y^*\right)\right) \end{cases} \end{aligned} \tag{25}$$

Then the final gradient formula is:

$$\begin{aligned} \nabla J(\theta) = & \frac{1}{T} \sum_{t=1}^{T} \sum_{y_t} R_{D \cdot Q}^{G_\theta}\left(Y_{1:T-1}, X, y_T, Y^*\right) \\ & \nabla_\theta\left(G_\theta\left(y_t \mid Y_{1:T-1}, X\right)\right) \\ & = \frac{1}{T} \sum_{t=1}^{T} E_{y_t \in G}\left[R_{D \cdot Q}^{G_\theta}\left(Y_{1:T-1}, X, y_T, Y^*\right) \cdot \nabla_\theta \log\left(G_\theta\left(y_t \mid Y_{1:T-1}, X\right)\right)\right] \end{aligned} \tag{26}$$

4) Evaluation mechanism. After the model calculates the evaluation score of BLEU using the BLEU evaluation criteria, it is compared with the complexity of the standard statement score. The lower the complexity the better the performance of the model, the target evaluation is calculated as:

$$PPL(S) = 2^{-\frac{1}{N}\sum \lg\left(P(\omega_i)\right)} \tag{27}$$

2.4. System Framework

The system framework of this paper is shown in Fig. 2. First, the corpus prepared in this paper is trained using the hierarchical Bayesian algorithm, while the corpus dictionary is built, and the word frequencies in the dictionary are counted. Then, the corpus is encoded using the generator in the adversarial generative neural network to train this data. The iterative training using the discriminator while verifying the BLEU score value can be obtained to improve the translation performance by the hierarchical Bayesian algorithm and the adversarial neural generative network.

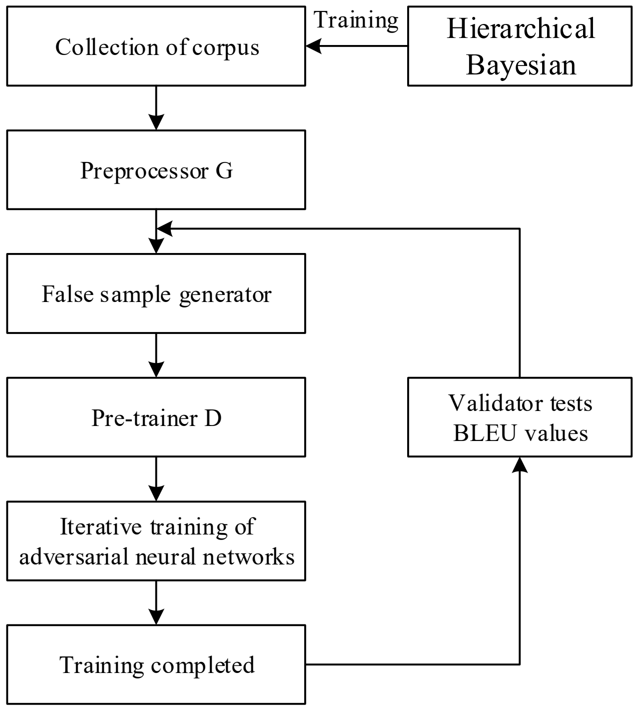


Fig. 2. System framework

2.5. Cross-domain Generalization Methods

2.5.1. Domain Generalization. Most of the existing approaches to domain generalization can be classified into the following categories: feature-based approaches, splitter-based approaches, data augmentation approaches, and meta-learning-based approaches [35]. The self of feature-based methods is to generate a domain invariant representation. Classifier-based methods aim to enhance generalization by fusing multiple sub-classifiers learned from the source domain. Data augmentation refers to the fact that some methods directly generate images based on these current source domains at the time of training that are stylistically very different from the current source domains, but with similar semantic differences (categorical features), with the goal of engaging a more diverse set of data in training so that the model can learn a richer representation of the features, and can perform better on the new domain data. Another approach is to utilize adversarial learning to directly generate new distributed data domains. In recent years, many approaches based on meta-learning strategies have been proposed to solve the domain generalization problem, aiming to allow models to obtain more general knowledge representations using a series of related tasks.

2.5.2. Domain Generalization Methods Based on Cross-domain Gating. In this subsection, a cross-domain gating based feature learning method, CDG for short, is proposed to solve the domain generalization problem during neural network training. When training for domain generalization, the training data used comes from a collection of multiple source domains with different data distributions, denoted as $D=\{D_1,\ldots,D_n\}$. Where, n denotes the number of source domains. For each training domain, we train an auxiliary domain specific classification network (ADS model) separately. Each ADS model is trained for object classification using only the corresponding single-domain data to obtain distribution (domain)-specific recognition capabilities.The training process of CDG contains two branches, the main branch for training the target network and the cross-domain gating branch for cross-domain activation of the target features. For the information from the data domain, we first extract the features of its low-level network, and then follow the following steps to train the CDG:

1) Feed the feature into the auxiliary domain-specific network of the cross-domain gating branch;

2) Calculate the gating mask for this input feature based on the labeled information return gradient;

3) Filtering information from the original feature by the gating mask. When performing testing or inference, the CDG uses only the target network of the primary branch.

3. Translation Cross-domain Generalization Performance Analysis

3.1. Analysis of the Corpus Experiment

In this section, the architecture of the neural machine translation system incorporating the hierarchical Bayesian model is compared with the baseline system, which proves the effectiveness of the neural machine translation method based on hierarchical Bayesian and effectively integrates the neural machine translation method based on hierarchical Bayesian with the standard model. Finally, this section provides a comprehensive analysis of the experimental results. In this corpus experiment, the author used the European Parliament corpus for the experiment.

3.1.1. Comparison of Different Fitting Functions. In the experiments based on the European Parliament corpus, this paper uses a statistical machine translation system as a baseline system for comparison experiments. This section first compares the hierarchical Bayesian-based neural machine translation method proposed in this paper on the mutual translation of Spanish (es), German (de), and French (fr). Secondly, in order to exclude the influence of other factors, this paper tries to use other European national languages (Danish da, English en, Finnish fi, Italian it, Norwegian nw) as the medial language in the direction of translation from Portuguese (pt) to Swedish (sv). In the choice of fitting functions for the number of phrase-pair co-occurrences, this paper compares the minimum value function, the maximum value function, the arithmetic mean function and the geometric mean function. The results of comparing different fitting functions with the baseline system are shown in Table 1, and the results of comparing different fitting functions with the baseline system under different medial languages are shown in Table 2.

Optimal results can be achieved using the hierarchical Bayesian-based neural machine translation method in all the interlanguage and medial language translation directions. The BLEU values of this paper's translation system exceed 29% and 22% in both the interlanguage and medial language translation directions, and exceed the baseline system in every interlanguage and medial language translation direction. This proves that the neural machine translation method based on hierarchical Bayes proposed in this paper has good results as well as robustness.

Table 1. Comparison of different merging methods against baseline (BLEU%)

Language pair	Baseline system	Ours	Minimum	Maximum	Arithmetic mean	Geometric mean
de-es	32.06	38.66	33.79	31.84	30.41	34.56
de-fr	28.58	32.39	28.75	29.77	30.02	30.85
es-de	34.06	37.34	24.87	31.65	25.27	28.32
es-fr	29.32	33.25	25.83	28.01	26.57	30.78
fr-de	25.35	29.34	23.66	28.43	25.55	21.75
fr-es	22.02	28.74	21.69	25.94	21.45	25.98

Table 2. Comparison of different merging methods against baseline in pivot languages (BLEU%)

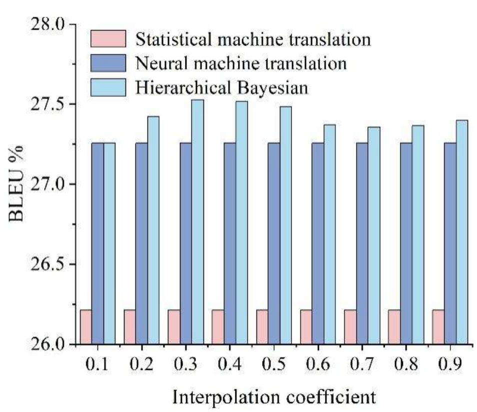
Language pair	Baseline system	Ours	Minimum	Maximum	Arithmetic mean	Geometric mean
pt-da-sv	22.12	22.84	20.25	19.69	22.35	22.29
pt-de-sv	22.87	24.02	21.93	21.38	22.11	21.82
pt-el-sv	22.48	22.83	21.16	20.77	19.99	20.86
pt-es-sv	21.36	25.09	21.11	22.42	19.39	23.16

pt-fi-sv	22.79	22.92	22.56	20.63	20.54	20.17
pt-fr-sv	21.96	24.28	20.98	21.89	19.85	22.42
pt-it-sv	20.79	25.02	21.01	22.69	22.55	20.92
pt-nw-sv	22.32	23.16	20.97	20.09	22.13	22.13

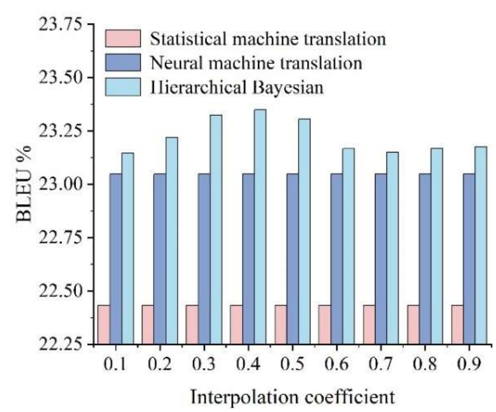
3.1.2. Comparison of the Fusion of This Paper's Translation Model and the Standard Model. The study is carried out on the example of medial language translation, which is adapted to the following scenarios: 1) There is a large-scale bilingual parallel corpus between source language-medial language and medial language-target language. 2) There is no bilingual corpus or only a small-scale bilingual corpus between source language-target language. Therefore, in this paper, 10,000, 50,000, 100,000, 200,000, and 500,000 bilingual parallel corpora are randomly selected from the source-target language bilingual corpus to simulate the scarcity of the standard corpus, which is used in this paper to train a standard neural translation model on a small scale, and then linearly interpolates the standard model with the hierarchical Bayesian machine translation model in this paper.

In this paper, experiments are conducted in six language directions: German-Spanish, German-French, Spanish-German, Spanish-French, French-German, and French-Spanish.

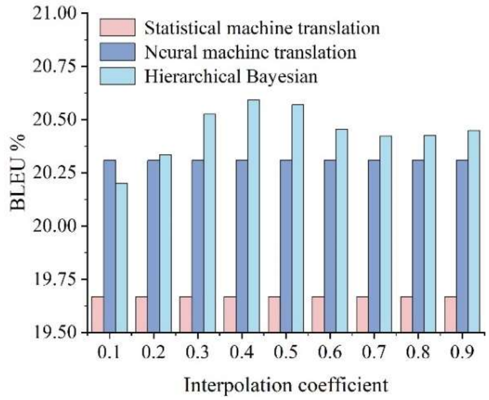
The experimental results are shown in Fig. 3, and (a) to (f) show the results on the six translation directions of German-Spanish, German-French, Spanish-German, Spanish-French, French-German, and French-Spanish, respectively. The horizontal axis of the graph represents the size of the standard corpus used, while the vertical axis represents the BLEU values. As the size of the standard corpus increases, the BLEU values of the standard statistical machine translation system and the neural machine translation system remain stable. And the performance of the neural machine translation system based on the hierarchical Bayesian model is also improving rapidly when the interpolation system is improved to 0.3 or 0.4. The BLEU values of the statistical machine translation system on the six language translations of German-Spanish, German-French, Spanish-German, Spanish-French, French-German, and French-Spanish are 26.22%, 22.43%, 19.67%, 32.91%, 20.02%, and 37.50%, respectively, and the neuro-machine translation system's six language direction BLEU values are 27.26%, 23.05%, 20.31%, 33.41%, 20.54%, 38.75%, and the average BLEU values of the neural machine translation system based on the hierarchical Bayesian model are 27.41%, 23.22%, 20.44%, 33.60%, 20.71%, and 38.91%. It can be seen that the neural machine translation system based on hierarchical Bayesian model in this paper achieves the best performance on the translation of six medial languages.



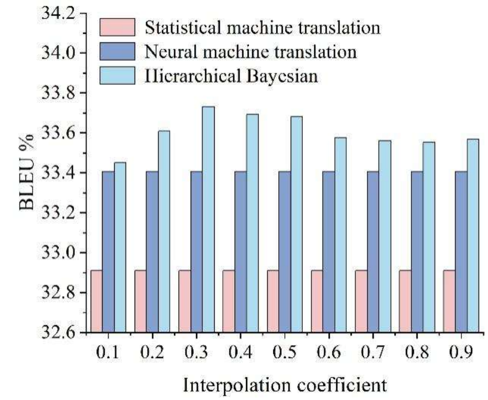
(a) German-English-Spanish translation



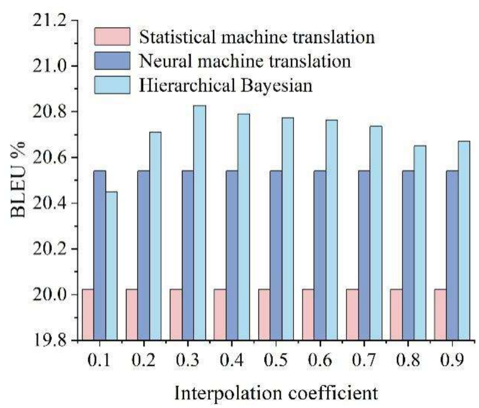
(b) German-English-French translation



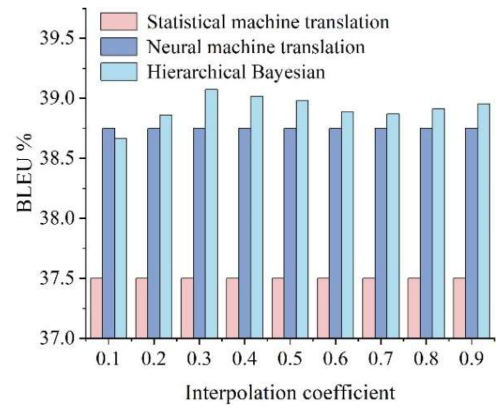
(c) Spanish-English-German translation



(d) Spanish-English-French translation



(e) French-English-German translation



(f) French-English-Spanish translation

Fig. 3. Comparison of pivot translation integrated with standard translation

3.2. Two-way Translation Analysis

Due to label smoothing, bundle embedding, double-byte encoding, and pre-trained word vectors, it is difficult to make a fair comparison between different methods. In order to make a fair comparison with other methods, Glove, a 100-dimensional pre-trained word vector, is used to train the model.In addition, in order to enhance the fairness of the comparison and to validate the reliability of the proposed model, the model is trained five times each on the Chinese-English and Chinese-Spanish bidirectional translations, for example, on the WMT14 cn-en, WMT14 en-cn, IWSLT14 cn-es, IWSLT14 es-cn datasets were trained on the model five times each and their average results are reported. However, for the WMT14 cn-en and WMT16 cn-en translation tasks, the model was trained only three times and the average results were reported considering the time cost and the consumption of computational resources. In addition, for the NER task, the model uses a Gaussian encoder, whereas for the machine translation task, the model uses a deterministic encoder to alleviate the difficulties during training.

1) Model results on the IWSLT14 Chinese-English translation task. On the IWSLT14 Chinese-English bi-directional translation task, other state-of-the-art models were experimentally compared with this paper's hierarchical Bayesian model-based neural machine translation approach, including models based on the attention mechanism (RNNSearch, CWE, NPMT, BiRNN+GCN) and variationally based models (VNMT, RNMT, VRNMT, AEVNMT, IntroVNMT, WAE).The Corpus BLEU scores on the WMT14 Chinese-English bidirectional translation task are shown in Table 3.

As shown in Table 3, the first part lists the performance of the methods based on the attention mechanism on the translation models, among which RNNSearch is one of the most popular and basic methods in machine translation, which achieves a BLEU value of 29.54% on Chinese-English translation.The CWE model learns inter-word similarity by Gumbel-softmax in order to reduce redundancy of embedding vectors, and its BLEU value is increased to 29.74%.NPMT model is a phrase-based neural machine translation method, in order to make full use of the structure of the phrases, sleep-wake Networks (SWAN) is used to explicitly model the phrases, and the BLEU value on Chinese-English and English-Chinese translation reaches 31.69% and 27.49% respectively, which achieves the then state-of-the-art results.The BiRNN+GCN model applies Byte Pair Encoding (BPE) to convert text to the subword level, while the penalty mechanism is applied to penalize candidate sentences that are too short, and these techniques help the BiRNN+GCN model to some extent to achieve higher performance, with BLEU values of 29.27% and 29.42% on both Chinese and English translations.

Table 3 then lists the performance of the variational-based approach on the translation model, where the difference between VNMT and VRNMT is that the former uses static latent variables to provide global semantic information for the model, while the latter uses dynamic latent variables to accomplish sentence translation. Next, the variational autoencoder-based neural machine translation model (AEVNMT) generates both source and target sentences from a shared latent representation with BLEU values of 29.45% and 27.16% on Chinese-English and English-Chinese translations, respectively. Inspired by the successful application of the variational selfencoder, an introspective VNMT model (IntroVNMT) was successively proposed to distinguish between the generated sentence's and the real sentence by means of the latent variables, and its BLEU values on Chinese-English and English-Chinese translations were 30.21% and 28.04%, respectively.The WAE-RN model is very close to the AEVNMT model in idea, and is also based on the variational self-encoder, and its BLEU values are also much closer to the AEVNMT model, which are 29.53% and 25.46%, respectively. From Table 3, it can be seen that without using BPE or other techniques, the neural machine translation method based on hierarchical Bayesian model (Hierarchical Bayesian) obtains a better performance, with performance BLEU values of 35.26% and 34.28% on bi-directional translation between Chinese and English, respectively. Compared with several other ATTENTION- and VARIATION-based neural machine translation models, the BLEU values outperform some of the state-of-the-art models, which proves that the Hierarchical Bayesian-based neural machine translation model is able to capture more useful information, and can further improve the performance of the neural machine translation system.

Table 3. Corpus BLEU in WMT14 Chinese-English two-way translation task (%)

Method	Model	WMT14 cn-en	WMT14 en-cn
Method based on attention mechanism	RNNSearch	29.54	29.37
	CWE	29.74	28.45
	NPMT	31.69	27.49
	BiRNN+GCN	29.27	29.42
Method based on variation	VNMT	27.69	26.49
	RNMT	29.18	28.42
	VRNMT	29.15	27.54
	AEVNMT	29.45	27.16
	IntroVNMT	30.21	28.04
	WAE-RN	29.53	25.46
Method based on hierarchical bayesian	Hierarchical Bayesian	35.26	34.28

2) Model results on the IWSLT Chinese-Western translation task. Next, experimental validation is carried out on the IWSLT14 Chinese-Western bidirectional translation task using the same method, and the comparison method is the same as the previous one, but with the addition of the unk replace

module to the RNNSearch and VNMT models.The Corpus BLEU scores on the IWSLT14 Chinese-Western translation task are shown in Table 4.

From the results shown in Table 4, it can be seen that the neural machine translation model based on hierarchical Bayesian model obtains the best translation performance on both datasets (26.42% and 25.96%, respectively) in comparison with the variational neural machine translation model and the attentional neural machine translation model, indicating that the neural machine translation model based on the hierarchical Bayesian model is able to both capture the complex relationships between objects and retains the semantic structure between objects, which is effective in improving the translation quality of neural machine translation.

Table 4. Corpus BLEU in WMT14 Chinese-Spanish two-way translation task (%)

Method	Model	WMT14 cn-en	WMT14 en-cn
Method based on attention mechanism	RNNSearch	17.15	16.84
	RNNSearch+unk replace	19.05	18.66
	CWE	17.69	15.28
	NPMT	16.89	16.08
	BiRNN+GCN	24.85	24.16
Method based on variation	VNMT	18.42	16.71
	VNMT+unk replace	20.52	19.38
	RNMT	26.16	24.09
	VRNMT	25.42	22.93
	AEVNMT	26.25	24.73
	IntroVNMT	24.63	23.41
	WAE-RN	24.87	22.44
Method based on hierarchical bayesian	Hierarchical Bayesian	26.42	25.96

3.3. *Verification of Generalizability*

The purpose of this paper proposing a hierarchical Bayesian-based neural machine translation method is to obtain a meta-parameter with better generalization, in order to verify the generalization of the meta-parameter. In this paper, we take Mongolian-Chinese language translation as an example, and simulate the generalizability of the meta-parameter under different low-resource scenarios by changing the size of the parallel corpus dataset of Chinese and English. The dataset division is shown in Table 5.

Table 5. Mongolian-Chinese datasets of different scales

Mongolian-Chinese parallel corpus			
Sentence quantity	Training set	Verification set	Test set
30000	20000	4000	4000
80000	60000	8000	8000
120000	100000	12000	12000
180000	150000	15000	15000
240000	200000	24000	24000

The BLEU values under the meta-learning method, the first-order meta-learning method and the method proposed in this paper using datasets of different sizes are shown in Figure 4. From the figure, it can be seen that when the size of the Mongolian-Chinese data set is 30,000 sentence pairs, the BLEU values obtained under the hierarchical Bayesian-based neural machine translation method are higher than those of the meta-learning method and the first-order meta-learning method, and it can be proved that in the case of a smaller data size, the learned meta-parameter generalization of the hierarchical Bayesian-based neural machine translation method is better than that of the other two meta-learning methods. With the expansion of the dataset, although the hierarchical Bayesian-based neural machine translation method is lower than the meta-learning method BLEU value of 0.65 when the Mongolian-Chinese dataset is 120,000 sentence pairs, it is higher than the BLEU value under the meta-learning method in most of the cases, which demonstrates that the hierarchical Bayesian-based neural machine translation method is more stable on the low-resource language translation task.

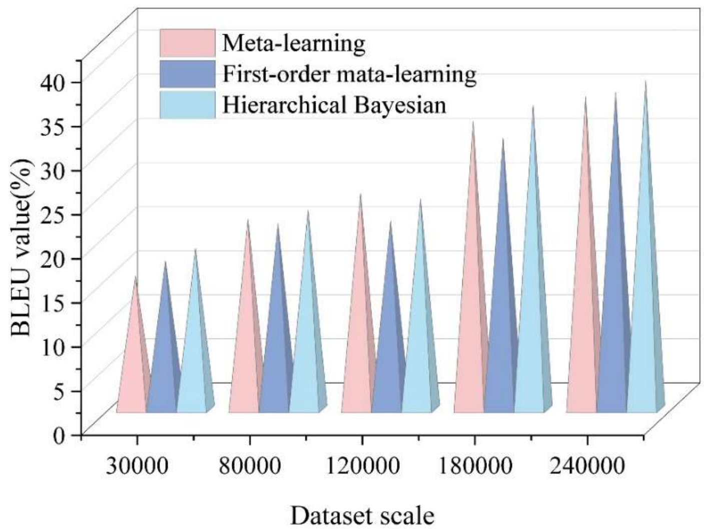


Fig. 4. BLEU values of three methods under different scale datasets

4. Conclusion

According to the model framework of neural machine translation, the article constructs hierarchical Bayesian model and adversarial neural network, integrates the two in standard neural machine translation, builds the neural machine translation system based on hierarchical Bayes, and adds the domain generalization method based on cross-domain gating into it.

1) Based on the European Parliament corpus for experiments, the statistical machine translation system is compared as the baseline system, and the mutual translation of several European languages and the medial language mutual translation are tested respectively, and it is found that in the direction of all the mutual translation languages and the medial language, this paper's hierarchical Bayesian-based neural machine translation method is obviously better than the baseline system.

2) Comparing the statistical machine translation model, the standard neural machine translation model with the hierarchical Bayesian-based neural machine translation system in this paper, the former two always maintain a stable translation performance. The hierarchical Bayesian-based neural machine translation system, on the other hand, rapidly grows to the maximum value as the interpolation coefficient grows to 0.3 or 0.4, and the overall average performance of the hierarchical Bayesian-based neural machine translation system always exceeds that of the statistical machine translation model and the standard neural machine translation model in the six axial language translation directions.

3) In the comparison with the translation methods based on attention and variational scores, the hierarchical Bayesian-based neural machine translation system also achieves the best translation performance, with BLEU values of 35.26% and 34.28% on Chinese-English bi-directional translation, and 26.42% and 25.96% on Chinese-Western bi-directional translation, respectively. The Hierarchical Bayesian-based Neural Machine Translation has high stability on the translation of low-resource languages.

Funding

Application Research on Blended Training for University ESP Course Teachers in the Era of AI + Education (No.230906121261230).

References

[1] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, 56(2), 593-619.

[2] Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering, 23(1), 3-30.

[3] Maruf, S., Saleh, F., & Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. ACM Computing Surveys (CSUR), 54(2), 1-36.

[4] Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. Engineering, 18, 143-153.

[5] Li, P., Wang, M., & Wang, J. (2021). Named entity translation method based on machine translation lexicon. Neural Computing and Applications, 33, 3977-3985.

[6] Koehn, P. (2020). Neural machine translation. Cambridge University Press.

[7] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4396-4415.

[8] Liu, X., Zeng, J., Wang, X., Wang, Z., & Su, J. (2024). Exploring iterative dual domain adaptation for neural machine translation. Knowledge-Based Systems, 283, 111182.

[9] Tang, J., Luo, Y., Ouzzani, M., Li, G., & Chen, H. (2022, June). Sevi: Speech-to-visualization through neural machine translation. In Proceedings of the 2022 International Conference on Management of Data (pp. 2353-2356).

[10] Muzaffar, S., & Behera, P. (2018). A qualitative evaluation of Google's translate: A comparative analysis of English-Urdu phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT) systems. Language in India, 18(10), 154-164.

[11] Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. Psychological review, 129(5), 999.

[12] Du, D., Chen, J., Li, Y., Ma, K., Wu, G., Zheng, Y., & Wang, L. (2022). Cross-domain gated learning for domain generalization. International Journal of Computer Vision, 130(11), 2842-2857.

[13] Ponti, F., Frezza, F., Simeoni, P., & Parisi, R. (2024). A Generalized Learning Approach to Deep Neural Networks. Journal of Telecommunications and Information Technology.

[14] Cho, W. I., Kim, J., Yang, J., & Kim, N. S. (2021, March). Towards cross-lingual generalization of translation gender bias. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 449-457).

[15] Zhou, A. (2023). Optimization of Unsupervised Neural Machine Translation Based on Syntactic Knowledge Improvement. International Journal of Advanced Computer Science & Applications, 14(11).

[16] Yin, Y., Fu, L., Li, Y., & Zhang, Y. (2024). On compositional generalization of transformer-based neural machine translation. Information Fusion, 111, 102491.

[17] Russin, J., Jo, J., O'Reilly, R., & Bengio, Y. (2020, July). Compositional generalization by factorizing alignment and translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 313-327).

[18] Lake, B., & Baroni, M. (2018, July). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In International conference on machine learning (pp. 2873-2882). PMLR.

[19] Singh, M., Kumar, R., & Chana, I. (2021). Machine translation systems for Indian languages: review of modelling techniques, challenges, open issues and future research directions. Archives of Computational Methods in Engineering, 28(4), 2165-2193.

[20] Behnke, M., & Heafield, K. (2021, November). Pruning neural machine translation for speed using group lasso. In Proceedings of the sixth conference on machine translation (pp. 1074-1086).

[21] Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5), 1-38.

[22] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339-351.

[23] Vu, T., Khadivi, S., Phung, D., & Haffari, G. (2022). Domain Generalisation for Neural Machine Translation. Learning to Adapt Neural Models with Limited Human Supervision in Natural Language Processing, 95.

[24] Pérez-Ortiz, J. A., Forcada, M. L., & Sánchez-Martínez, F. (2022). How neural machine translation works. Machine translation for everyone: Empowering users in the age of artificial intelligence, 18, 141.

[25] Liu, J., Zhang, X., Tian, X., Wang, J., & Sangaiah, A. K. (2020). A novel domain adaption approach for neural machine translation. International Journal of Computational Science and Engineering, 22(4), 445-453.

[26] Saunders, D. (2022). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75, 351-424.

[27] Pandey, M., Pandey, R., & Nazarov, A. (2023, November). Dynamic Multihead Attention for Enhancing Neural Machine Translation Performance. In International Conference on Trends in Computational and Cognitive Engineering (pp. 87-99). Singapore: Springer Nature Singapore.

[28] Man, Z., Huang, Z., Zhang, Y., Li, Y., Chen, Y., Chen, Y., & Xu, J. (2023). WDSRL: Multi-domain neural machine translation with word-level domain-sensitive representation learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 577-590.

[29] Naeem, M., Siddique, A. B., Ali, R. H., Arshad, U., ul Abideen, Z., Khan, T. A., ... & Ali, N. (2023, December). Performance Evaluation of Popular Deep Neural Networks for Neural Machine Translation. In 2023 International Conference on Frontiers of Information Technology (FIT) (pp. 220-225). IEEE.

[30] Tavares, C., Tallone, L., Oliveira, L., & Ribeiro, S. (2023). The Challenges of Teaching and Assessing Technical Translation in an Era of Neural Machine Translation. Education Sciences, 13(6), 541.

[31] Shaolin Zhu,Leiyu Pan,Dong Jian & Deyi Xiong.(2025).Overcoming language barriers via machine translation with sparse Mixture-of-Experts fusion of large language models.Information Processing and Management(3),104078-104078.

[32] Changuk Mun,Jong Wha Bai & Junho Song. (2025). Hierarchical Bayesian models with subdomain clustering for parameter estimation of discrete Bayesian network. Structural Safety102570-102570.

[33] Maura Mezzetti & Ilia Negri. (2024). Hierarchical Bayesian model to estimate and compare research productivity of Italian academic statisticians. Scientometrics(12),1-32.

[34] Xiangna Chen,Feiyi Liu,Weibing Deng,Shiyang Chen,Jianmin Shen,Gábor Papp... & Chunbin Yang. (2024). Applications of Domain Adversarial Neural Network in phase transition of 3D Potts model. Physica A: Statistical Mechanics and its Applications129533-.

[35] Jing Liu,Wei Zhu,Di Li,Xing Hu & Liang Song. (2024). Domain generalization with semi-supervised learning for people-centric activity recognition. Science China Information Sciences(1),112103-112103.