Hindawi
Mobile Information Systems
Volume 2022, Article ID 8344814, 10 pages
https://doi.org/10.1155/2022/8344814

Research Article

AI-Based Heterogenous Large-Scale English Translation Strategy

Chuncheng Wang

Tongling University, Tongling 244061, China

Correspondence should be addressed to Chuncheng Wang; chad825@tlu.edu.cn

Received 3 January 2022; Revised 21 January 2022; Accepted 26 January 2022; Published 9 February 2022

Academic Editor: Jianhui Lv

English has become one of the most widely used languages in the world. If there is no good translation mechanism for such a widely used language, it will bring trouble to both study and life. At present, the world's major platforms are committed to the study of English translation strategies. There are translation platforms from different regions and different translation mechanisms. These translation data from different translation platforms have the characteristics of large-scale, multisource, heterogeneity, high dimensions, and poor quality. However, such inconsistent translation data will increase the translation difficulty and translation time. Therefore, it is necessary to improve the quality of translation data to achieve a better translation effect. How to provide a large-scale and efficient translation strategy needs to integrate the translation strategies of various platforms to perform heterogeneous translation data cleaning and fusion based on machine learning. At first, this paper represents the multisource, heterogeneous translation data model as tree-augmented naive Bayes networks (TANs) and naturally captures the relationship between the datasets through the learning of TANs structure and the probability distribution of input attributes and tuples, using data probability value to complete the classification of translation data cleaning. Then, a multisource, heterogeneous translation data fusion model based on recurrent neural network (RNN) is constructed, and RNN is used to control the node data of hidden layer to enhance the fault-tolerant ability in the fusion process and complete the construction of fusion model. Finally, experimental results show that TANs-based translation data cleaning method can effectively improve the cleaning rate with an average improvement of approximately 10% and cleaning time with an average reduce about 5%. In addition, RNN-based multisource translation data fusion method improves the shortcomings of the traditional fusion model and improves the practicability of the fusion model in terms of root mean square error (RMSE), mean absolute percentage error (MAPE), fusion time, and integrity.

1. Introduction

The importance of translation is seldom paid attention to in English learning. However, with the progress and development of society, the demand for English translation talents is increasing, making translation gradually receive attention in teaching [1, 2]. Translation is conducive to exercise students' divergent thinking, promote the comprehensive development of English level, and let students fully understand foreign culture. It is very important to cultivate students' translation ability, which is not only conducive to improve students' comprehensive English quality, but also enhance their ability and level of listening, speaking, reading, and writing. It can also exercise students' thinking ability and truly translate works that can be accepted universally. In the process of translation, students can fully understand foreign knowledge and foreign culture [3–5]. Based on this, various translation platforms emerge one after another. However, there is no large-scale English translation strategy [6]. The so-called large-scale refers to the text translation data from different regions, different platforms, and different translation mechanisms [7]. These data have large-scale, multisource, heterogeneous types and modes, high dimensions, and poor quality. Such data inconsistency will increase the difficulty of translation and the translation time. Therefore, it is necessary to improve the quality of data to achieve better translation effect. But how to provide a large-scale and efficient translation strategy? Data cleaning and data fusion are needed to integrate the translation strategies of each region.

Datasets inevitably exist as redundant data, missing data, uncertain data, and inconsistent data, and these data is called "dirty data" [8, 9]. The purpose of data cleaning is to detect the incorrect and inconsistent data in the English translation dataset and then delete or correct them so as to improve the quality of the translation data. Multisource, heterogeneous translation data cleaning refers to removing noise data and irrelevant data from datasets of different translation platforms, processing omitted data, and removing white noise in blank data fields and knowledge background. In multisource, heterogeneous translation environment, traditional data cleaning methods are generally divided into two ways: (1) Data is fused into the same data source through data integration or data fusion, and data cleaning is carried out during and after data fusion. (2) Through the development of unified data cleaning standards, each data source carries out inaccurate data cleaning at the same time. The latter method is more difficult to clean, and the cleaning efficiency is uncertain, but if the cleaning model is proper, the cleaning efficiency is greatly improved compared to the first method.

The successful application of machine learning (ML) [10] in pattern recognition, information retrieval, data mining, and other fields provides a new solution for data processing, but the traditional statistical ML method cannot be simply used to process multisource, heterogeneous data because traditional statistical ML methods assume that the data to be processed comes from the same feature space and has the same distribution. Classification and clustering are two main problems in ML [11], and the typical solution is to use Bayesian network. Bayesian network [12] is a method to obtain a priori probability from a specific domain by using graphic patterns, and Bayesian network is suitable for processing incomplete data. For processing instances with missing values, all possible attribute values can be accumulated or integrated. In addition, Bayesian network has strong robustness for the overfitting problem of the model. Based on the advantages of Bayesian network in data processing, this paper considers Bayesian network to process uncertain translation data. Considering the characteristics of large-scale, multisource, heterogeneous translation data, and the advantages of Bayesian network in processing inconsistent data, a multilayer reduction model based on Tree-Augmented Naive Bayes Networks (TANs) [13] is constructed, which includes data source reduction, data attribute layer and tuple layer reduction, and duplicated data cleaning.

Accordingly, the main contributions of this study can be summarized as follows. (i) TANs-based translation data cleaning method is proposed. (ii) The RNN-controlled translation data fusion method is proposed.

The rest of this paper is organized as follows. Section 2 reviews related work. In Section 3, TANs-based heterogeneous translation data cleaning method is presented. In Section 4, RNN-controlled heterogeneous translation data fusion is proposed. Experimental results are presented in Section 5. Section 6 concludes this paper.

2. Related Work

In the process of data acquisition, based on the comprehensiveness of data collection and the integrity of relevant data, data collection usually involves multiple data sources, including a variety of databases, file systems and service interfaces, resulting in complex data types and large data scale. Therefore, it is necessary for data cleaning after data acquisition. In [14], a novel data cleaning technique was introduced to remove dirty data effectively, and based on optimization, a new hybrid firefly update enabled rider optimization algorithm was proposed. In [15], a hybrid data cleaning system that integrated bias detection and repair was proposed to process multiple bias types. In [16], a federated data cleaning protocol was presented to realize data cleaning without damaging data privacy in edge intelligence. In [17], a privacy-aware data cleaning as a service model was proposed, which promoted the interaction with the parties requesting data query from the client, as well as the service provider using the data pricing scheme, which calculated the price according to the data sensitivity. In [18], a set of conditional functional dependencies was introduced into the density-based data cleaning algorithm, and the algorithm used the set to repair inconsistent data. Internet of things is a multisource information integration technology, which collects multisource data every day, but most of the data may be irrelevant and redundant. In [19], a deep Q-network-based feature selection method was proposed for multisource data cleaning. In clinical research, the abundance of data resources also brings the challenge of data cleaning; some data cleaning methods had been proposed [20–22].

Multisource, heterogeneous data refers to data with multiple sources and different component characteristics. The fusion of multisource, heterogeneous data can assist researchers to obtain effective information in data. Therefore, multisource, heterogeneous data fusion has become a hot topic in current research. In [23], a multiscale deep coupled neural networks was proposed to fully fuse the fault information. In [24], the author presented a multisource, heterogeneous data fusion method for narrowband Internet of things based on perceptual semantics. In [25], a novel and dynamic opportunistic clustering and data fusion scheme based on self-organizing hesitation fuzzy entropy was presented to overcome energy consumption and the bottleneck of network lifetime. In [26], the authors had made a comprehensive research on the data fusion method based on ML. Data fusion also plays a pivotal role in providing environmental information. In [27], the authors listed data fusion algorithms provided by the general data fusion framework and illustrated the method of the general data fusion framework by an example of three-dimensional reconstruction from two-dimensional images. In [28], the authors analyzed the detection process of massage chair intelligent detection robot and made theoretical research from decision-level fusion and data-level fusion. Data generated by sensors in the Internet of Things always is large-scale, multisource, and heterogeneous, so data fusion is necessary for providing intelligent services [29–32].

Translation data cleaning and fusion are pivotal steps for the following data processing, so this paper pays attention to translation data cleaning and fusion in heterogeneous large-scale English translation.

3. TANs-Based Translation Data Cleaning

The main method to classify multisource, heterogeneous data is to establish and examine the multisource, heterogeneous data network model formed by data relationship. TANs model is a method of knowledge representation, learning, and reasoning on the structure and relationship between data based on probability framework, which can better describe the uncertainty of translation data. Due to the dependence between various attributes in the translation data source, this paper represents the multisource, heterogeneous translation data model as TANs and naturally captures the relationship between the datasets through the learning of TANs structure and the probability distribution of input attributes and tuples.

The basic idea of translation data cleaning based on TANs is as follows: according to different eigenvectors of translation data, the translation data attributes to be cleaned are divided into different classes to form multiple Bayesian network structures. A parent node is made up of several child nodes. The translation data generated by different data objects in Bayesian networks in different Bayesian network structures indicate a data object according to the importance of data attributes and tuples in the network. Therefore, the Bayesian network formed by data objects in one data domain becomes a comparable quantity with the Bayesian network formed by data objects in another data domain. Divide all translation data generated by data objects containing the same task into a dataset. Finally, the associated data objects of different data sources are combined into a dataset to be cleaned using TANs to represent data attributes.

Suppose a dataset $D = \{T_1, T_2, \ldots, T_n, C\}$, C is a class variable and its value ranges are $\{c_1, c_2, \ldots, c_m\}$, m is the total number of classes, $\{t_1, t_2, \ldots, t_n\}$ is the attribute value of $\{T_1, T_2, \ldots, T_n\}$ that represents classification feature, and n is the number of attributes of the classification. TANs classifier is assumed by attribute nodes $\{T_1, T_2, \ldots, T_n\}$. The structure of TANs network composed of attribute nodes $\{t_1, t_2, \ldots, t_n\}$ is a tree, and each attribute variable has no more than one attribute parent node besides the parent class. Each attribute variable has no more than one attribute parent node except the parent class, and a tree is formed between the attribute nodes as the maximum weight span tree.

For probability distribution $P(T_1, T_2, \ldots, T_n, C)$, and we have

$$\arg\max\left\{ P(C)\prod_{i=1}^{n} P(t_i)|\delta_{t_i}, C, WS_T \right\}. \tag{1}$$

The classifier for predicting variable C is TANs classifier, where WS_T represents the maximum weight span tree of $T_1, T_2, \ldots, T_n$ under the constraint of class variable C, and δ_{t_i} is the value of the attribute parent $\prod (t_i)$ of t_i in the maximum weight span tree.

The TANs attribute tree can be characterized by the function δ: $\{1, 2, \ldots, n\} \longrightarrow \{0, 1, \ldots, n\}$, where the node $\pi(i) = 0$ is the parent node. There is no sequence $\{i_1, i_2, \ldots, i_k\}$ to make $\delta(i_j) = i_j + 1$, where $i \leq j < k$, and $\delta(i_k) = i_1$; that is, no undirected loop can be generated. When $\delta(i) > 0$, $\prod t_i = \left\{t_{\delta(i)}\right\}$. When $\delta(i) = 0$, $\prod t_i = \varnothing$. Therefore, the function δ defines the structure of the TANs classifier.

The weights of TANs attributes are constructed by calculating the mutual information of attributes among variables. Mutual information refers to the degree of association between two random variables, that is, the degree of uncertainty weakening of another random variable given a random variable. According to the chain rule of entropy, $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$. Therefore, the difference $H(X) - H(X|Y) = H(Y) - H(Y|X)$ is denoted as $MI(X, Y)$, and mutual information $MI(X, Y)$ is defined as follows.

$$MI(X, Y) = \sum_{xy} P(x, y)\log\frac{P(x, y)}{P(x)P(y)}, \tag{2}$$

where $P(x, y)$ is the joint distribution of variable (X, Y) and $P(x)$ and $P(y)$ are the marginal distribution, respectively.

The mutual information between attributes is the correlation of attributes, and the correlation values calculated by different class attributes are also different. Considering the addition of TANs class variable attributes, the mutual information equation of a certain classification attribute needs to be redefined, so the mutual information of TANs can be calculated as follows.

$$\begin{aligned} MI_{ij} &= \left(c_i, c_j \mid C\right) \\ &= \sum_{c_i, c_j, c} P\left(c_i, c_j, c\right) D\left(c_i \| c_j\right) \log\frac{P\left(c_i, c_j \mid C\right)}{P(c_i \mid C)P\left(c_j \mid C\right)}, \end{aligned} \tag{3}$$

where c_i and c_j are attribute variables and C are class variables, and $D(c_i \| c_j)$ is the relative entropy of c_i and c_j.

The construction for TANs is summarized as follows.

Step 1. Obtain the mutual-information values $MI(T_i, T_j)$ of all attribute pairs through equation (3).

Step 2. $MI(T_i, T_j)$ is sorted in descending order, and node pairs are output successively.

Step 3. According to the principle that TANs does not generate a loop, edges are selected in descending order of edge weight until $n - 1$ edges are selected, and a completely undirected graph with mutual-information value as weight is constructed.

Step 4. Select any node in the completely undirected graph as the root of TANs and set the direction of all edges outward from the root node. The process of converting an undirected tree into a directed tree is completed by setting the direction between attribute nodes.

Step 5. Add a class node (i.e., class attribute node) to each node in TANs and the directed edges of the class node pointing to all attribute nodes.

Therefore, TANs-based translation data cleaning steps are as follows.

(1) The translation data attribute dataset TD is sampled.

(2) The importance measurement algorithm [33] is used to reduce the data level of the sampled translation data.

(3) Construct TANs.

(4) The mutual-information value is used to score the TANs.

(5) The Top-k problem is to get the maximum number of k from an array or list. Judge the Top-k mutual-information value $MI(T_i, T_j)$ and the empirical parameter value. If the Top-k mutual-information value $MI(T_i, T_j)$ is less than the empirical parameter value, remove the top-level node and query the score R of the result of removing the top-level node.

(6) The TANs nodes are sorted and output in descending order according to the score R.

4. RNN-Controlled Translation Data Fusion

4.1. Position of Multisource, Heterogeneous Translation Data Nodes. Due to the structural diversity of multisource, heterogeneous translation data, the data structure will produce diverse fusion results. Therefore, before multisource, heterogeneous fusion, nodes in multisource, heterogeneous translation data need to be located. The improved firefly algorithm [34] is used to calculate the distance of the coordinates of heterogeneous anchor nodes. DV-HOP algorithm is used to calculate the hop count of nodes and anchor nodes as follows.

$$t_{k+1/k} = NP_{k|k}N^{T} + S, \tag{4}$$

where $t_{k+1/k}$ represents the prior estimate value of state at time $K+1$, $P_{k/k}$ represents the posterior estimate matrix at time K, S is the state estimate value, T is the time, and N is the anchor node.

Equation (3) is used to preliminarily lock the region where the heterogeneous translation data is located, collect all sample points in this region, and predict the location of unknown mobile nodes. Suppose that the moving speed of the unknown node meets the interval $[0, v_{\max}]$ and presents an interval uniform distribution, and its position is defined as follows.

$$\text{pos} = \begin{cases} \dfrac{1}{\delta v^{2}}, & d \leq v_{\max}, \\ \\ 0, & d > v_{\max}, \end{cases} \tag{5}$$

where d represents the distance of moving node from time $K-1$ to time K, and v represents the average speed of moving node. Set the communication radius between hops and remove nonconforming nodes, if the nodes are within a communication range and less than the communication radius with anchor nodes. On the contrary, the nodes that do not meet the conditions are filtered, the calculation results of all nodes are integrated, the data fusion mapping relationship of nodes is described, and the fusion model is constructed.

4.2. Analysis of Translation Data Fusion Mapping. Any data fusion process can be viewed as a process of external to internal mapping. When constructing the multisource, heterogeneous translation data model, the node relation is obtained to describe the translation data fusion mapping relation. The quintuple in the fusion model is defined as follows.

$$FM = \{S, M, MS, TS, Rm\}, \tag{6}$$

where S represents the state data in the prefusion space, M represents the measurement space, MS represents the fusion space, TS represents the target space for fusion judgment, and Rm represents the mapping set relationship between different spaces.

Suppose the mapping set has the following triplet relationship.

$$Rm = \{\alpha, \beta, f\}, \tag{7}$$

where α represents the measurement space mapping in the space to be fused, β represents the process of transforming original multisource, heterogeneous translation data into integrated spatial data after mapping processing, and f represents the mapping spatial relationship. The space before the fusion of j multisource, heterogeneous translation data can be expressed as follows.

$$S = \begin{pmatrix} S_{11} & S_{1j} \\ S_{j1} & S_{ij} \end{pmatrix}, \tag{8}$$

where rows represent the target contained in the space before translation data fusion, and columns have multisource, heterogeneous attributes. i represents the maximum number of features of the fusion target, which is zero if the target does not contain the related feature. Assume that the fusion space M at time t is expressed as follows.

$$M_t = \begin{pmatrix} m_{11} & m_{1q} \\ m_{q1} & m_{pq} \end{pmatrix}, \tag{9}$$

where m_{pq} represents the qth heterogeneous translation data obtained by information source p in the fusion model at time t. The maximum value of translation data provided by each information source is u, and the number of heterogeneous translation data sources is v, so the matrix of the fusion space is expressed as follows.

$$MS = \begin{pmatrix} ms_{11} & ms_{1v} \\ ms_{v1} & ms_{uv} \end{pmatrix}. \tag{10}$$

Combining equations (9) and (10), and the mapping relation can be calculated as follows.

$$MS = \sigma(M_t). \tag{11}$$

In the underlying dataset fusion, the translation data corresponding to nodes has been simply preprocessed. At this point, the mapping relation of σ is $1:1$, and the final space TS is composed of the final result of the fusion model, and the space can be expressed as follows.

$$TS = (d_1, d_1, \ldots, d_n)^T, \tag{12}$$

where d_i is the final fusion degree of fusion target i, and the fusion mapping relation can be expressed as follows.

$$TS = f\left(\sum_{t=0}^{q\Delta t} MS_t\right), \tag{13}$$

where Δt represents the time interval between fused translation data and q represents fusion times of translation data. Finally, the translation data fusion mapping relationship is described by using equation (13). By using the characteristics of the recurrent neural network (RNN), the fusion process of the fusion model is controlled and the multisource heterogeneous translation data fusion model is constructed.

4.3. RNN-Controlled Fusion Process. Before the fusion model is completed, RNN is used to control the multisource heterogeneous translation data fusion process, and the neural network structure is used to control. For the mapping set ts formed after mapping, it is assumed that the output of the input set of the neural network is O_{ts}, where the ith input is O_{ts_i}; that is,

$$W = \sum_{t=0}^{n} O_{ts\_i} O_{ts}. \tag{14}$$

The sigmoid function is a common "S" shape that is often used as the threshold function of neural networks, mapping variables to between 0 and 1, so equation (14) can be regarded as sigmoid function, and then it can be transformed into

$$\begin{aligned} O_{ts\_i} &= \text{sigmoid}(W) \\ &= \frac{1}{1 + e^{-O_{ts}}}. \end{aligned} \tag{15}$$

In order to reduce the bias in the neural network, the bias function [35] is normalized, and the bias function calculates the biases in the whole fusion process.

$$B_{ts} = \frac{1}{3} \sum_{ts} (w_i - O_{ts\_i})^2, \tag{16}$$

where w_i represents the training weight of the RNN, which is used to limit the error function to a minimum. In order to ensure that the RNN can control all hidden layer nodes and enhance fault-tolerant during the whole fusion process, hidden layer node information is selected, and the calculation is as follows.

$$h = \frac{I + O}{2} + \varepsilon, \tag{17}$$

where h represents the number of nodes in the hidden layer, I represents the number of input nodes, O represents the number of output nodes, and ε represents the constant ranges from 0 to 1. Constantly delete and increase the number of nodes, determine the nodes in the control fusion process, realize the neural network control fusion process, and complete the construction of multisource, heterogeneous translation data fusion model based on RNN.

5. Experiment and Results Analysis

5.1. Setup. This paper selected five different types of English corpora, which were Corpus of Contemporary American English (COCA), 20 Newsgroups, Auslan, Reuters-21578, and UNIX_user_data. Aiming at imprecise corpus data and similar repeated corpus data, the effectiveness of TANs-based translation data cleaning method proposed in this paper was verified by using two metrics: cleaning rate and cleaning time. FU-ROA [14], IHCS [15], and FedClean [16] are selected for comparison.

Root mean square error (RMSE), mean absolute percentage error (MAPE), fusion time, and integrity were used for comparison of heterogeneous translation data fusion. In comparison, MDCN [23], NB-IoT [24], and HFECS [25] were used.

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \\ \text{MAPE} &= \sqrt{\frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}. \end{aligned} \tag{18}$$

5.2. Comparison Analysis

5.2.1. Translation Data Cleaning Experiment. Figure 1 shows that the cleaning rate of TANs-based translation data cleaning method proposed in this paper is generally in COCA, 20 Newsgroups, UNIX_user_data, and Reuters-21578 showed significantly higher cleaning rates in four English corpora than in the other three baselines. However, the improvement in Auslan English corpus is not obvious because there is a conditional independence relationship between the sample attributes. Except for Auslan English corpus, the method in this paper maintains a relatively stable cleaning rate, basically maintaining or even exceeding 90%, and can maintain such a high level of classification accuracy, indicating that the method in this paper is feasible. In the heterogeneous large-scale English translation strategy, due to the heterogeneity of translation data on different platforms, the good translation data cleaning rate solves the problem that the correlation of translation data in multisource, heterogeneous environment may lead to overfitting. The core tuple and boundary tuple are retained according to

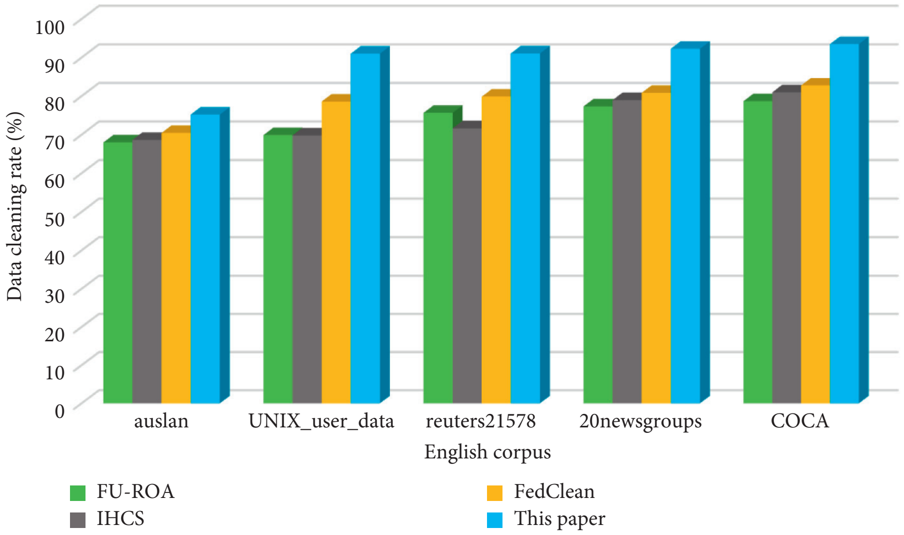

Figure 1: Translation data cleaning rate under different English corpora.

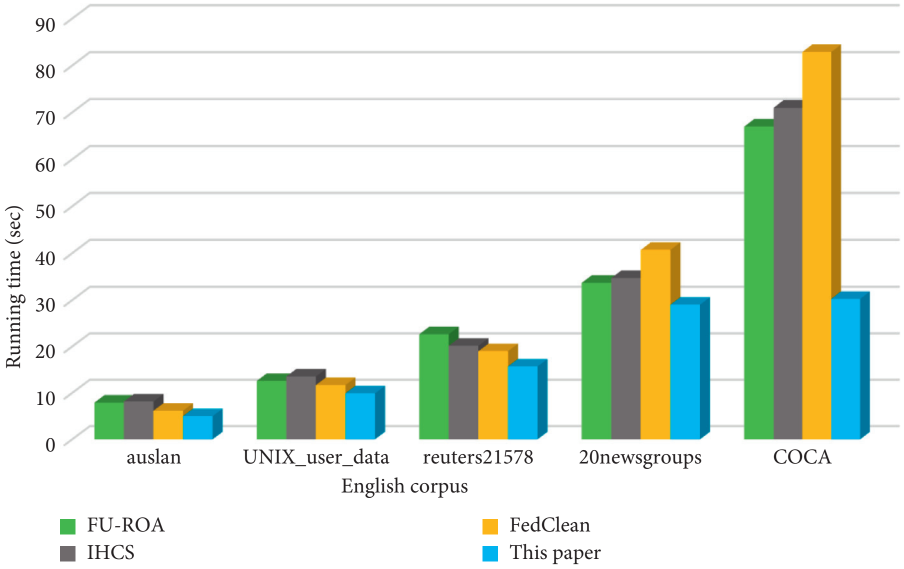

Figure 2: Translation data cleaning time under different English corpora.

the weight to reduce the outlier tuple, which greatly reduces the workload of data cleaning and the difficulty of translation.

As can be seen from Figure 2, for the small-scale English corpora Auslan, UNIX_user_data, Reuters-21578, and 20 Newsgroups, the performance effects of the four algorithms are not different. However, for the English corpus COCA with high dimension and large scale, the time spent by this algorithm is relatively less and the performance improvement is relatively obvious. The main reason is that when the amount of data is large, the improved attribute reduction algorithm can be mapped to a less search space, and the simplest attribute set can be solved by searching only a few simplified elements, which reduces the time consumption of attribute reduction and improves the efficiency of the algorithm, especially for large-scale English corpus. At the same time, it also solves the problem of redundancy and dependence of data features extracted from different interrelated English corpus. In heterogeneous large-scale English translation strategies, less translation data cleaning time corresponds to less translation time, and the translation effect is better.

5.2.2. Translation Data Fusion Experiment. As can be seen from Figure 3, RMSE of the proposed method in this paper is generally low in all English corpora. RMSE of NB-IoT algorithm performed well in Auslan, UNIX_user_data, Reuters-21578, 20 Newsgroups English corpus, better than MDCN and HFECS, but weakened in COCA corpus. The RMSE of MDCN algorithm on COCA corpus even exceeds 10. RMSE actually describes a degree of dispersion, while the

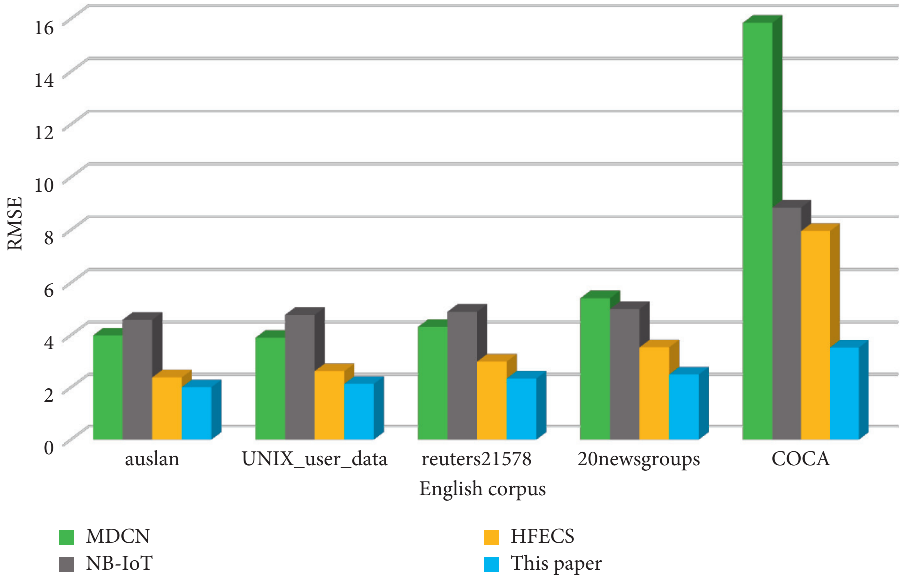

Figure 3: RMSE of translation data fusion under different English corpora.

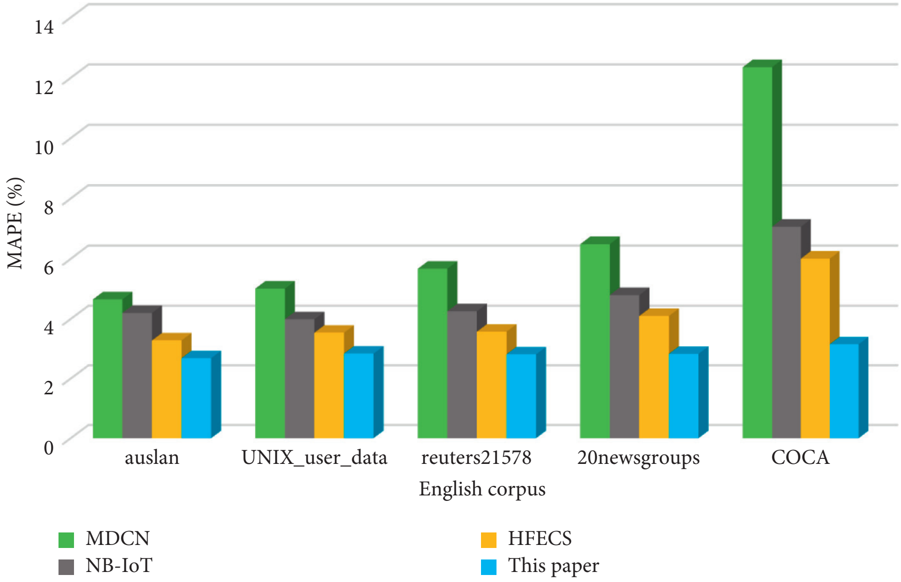

Figure 4: MAPE of translation data fusion under different English corpora.

translation data fusion algorithm proposed in this paper has a low RMSE, which further indicates that the algorithm presented in this paper has a good stability, while the other three baselines have great differences in different English corpora. It is not conducive to translation data fusion of heterogeneous translation platforms. From Figure 4, it can be seen that the MAPE of the translation data fusion algorithm in this paper always remains below 5% in each English corpus, showing a good fitting effect.

As indicated in Figure 5, the translation data fusion time of the four algorithms on Auslan, UNIX_user_data, and Reuters-21578 English corpora is similar. However, in 20 Newsgroups and COCA English corpus, the fusion time of the translation data fusion algorithm in this paper is much lower than the other three baselines. In large-scale, heterogeneous English translation, less heterogeneous data fusion time provides strong support for translation strategies. Figure 6 shows the integrity of the translation data fusion algorithm in this paper is always the best and remains above 80%. A large-scale and efficient translation strategy requires a complete integration of translation strategies in all platforms and the fusion of translation data. However, the algorithm in this paper maintains the integrity of heterogeneous translation data.

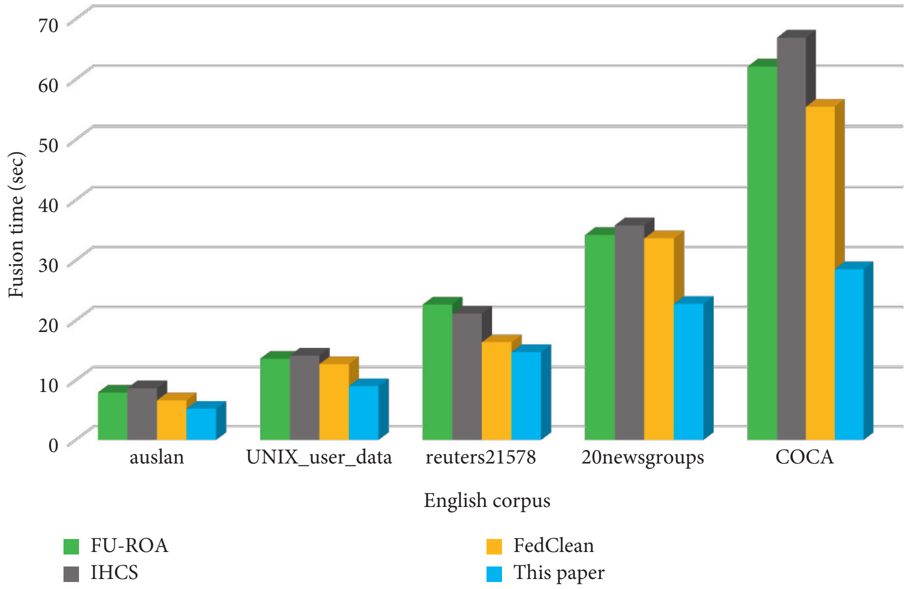

Figure 5: Fusion time of translation data fusion under different English corpora.

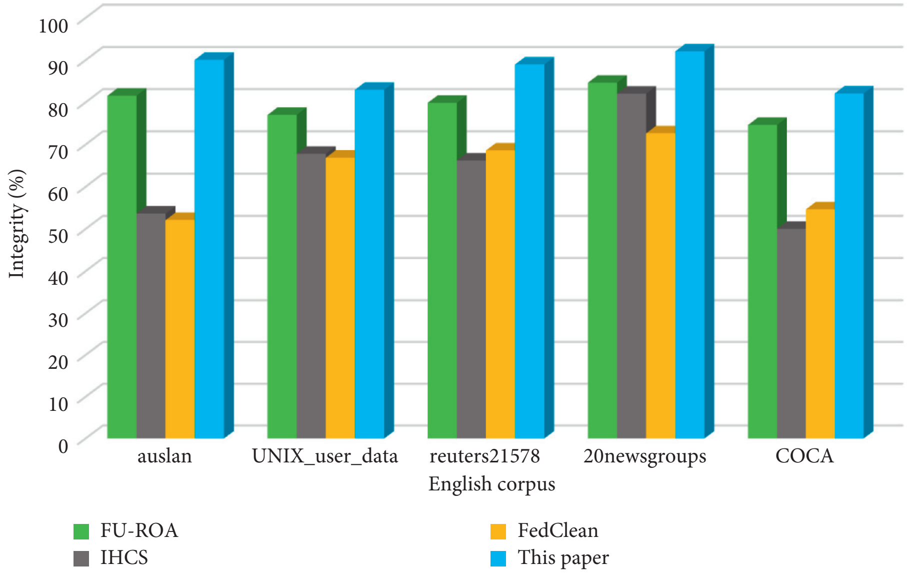

Figure 6: Integrity of translation data fusion under different English corpora.

6. Conclusions

Aiming at the problem of large amount of translation data in multisource, heterogeneous translation environment, TANs-based translation data cleaning is proposed to clean translation data. Translation data attributes and tuples are weighted by data hierarchical reduction, which solves the problem that translation data correlation may lead to overfitting in multisource, heterogeneous environment. Core and boundary tuples are retained according to the weight, and outlier tuples are reduced, which greatly reduces the workload of translation data cleaning. In addition, aiming at the problem that the number of unique elements obtained by the traditional multisource, heterogeneous translation data fusion model is small, resulting in the lack of strong integrity of the final fusion data, a multisource, heterogeneous translation data fusion model based on RNN is constructed. By locating the node data in the multisource, heterogeneous data and analyzing the mapping relationship, the multisource heterogeneous translation data fusion model is realized. Experimental results show that TANs-based strategy can effectively improve the cleaning rate and cleaning time of translation data cleaning compared with baselines. Moreover, the translation data fusion experiment results demonstrate that the RNN-based translation data fusion method outperforms baseline with respect to RMSE, MAPE, fusion time, and integrity.

In the future work, we will deeply study various data types, data storage methods, data association status, and errors in the process of data cleaning. At the same time, although the translation data fusion model in this paper improves some metrics to a certain extent, there are still some deficiencies. It still needs to be improved in future research to obtain a better fusion model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] S. Bi, "Intelligent system for English translation using automated knowledge base," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 5057–5066, 2020.

[2] S. Wang, "Simulation of English translation text filtering based on machine learning and embedded system," *Microprocessors and Microsystems*, vol. 83, 2021.

[3] Y. Liu and H. Bai, "Teaching research on college English translation in the era of big data," *International Journal of Electrical Engineering Education*, 2021.

[4] Z. Li, "Simulation of English Education Translation Platform Based on Web Remote Embedded Platform and 5G Network," *Microprocessors and Microsystems*, vol. 81, 2021.

[5] L. Shi and X. Wang, "Strategies of cross-cultural eco-education in college English translation teaching," *Ekoloji*, vol. 28, no. 107, pp. 3045–3050, 2019.

[6] M. Kolhar and A. Alameen, "Artificial intelligence based language translation platform," *Intelligent Automation and Soft Computing*, vol. 28, no. 1, 2021.

[7] L. Ma, M. Huang, S. Yang, R. Wang, and X. Wang, "An adaptive localized decision variable analysis approach to large-scale multiobjective and many-objective optimization," *IEEE Transactions on Cybernetics*, 2021.

[8] J. A. DeSimone and P. D. Harms, "Dirty data: the effects of screening respondents who provide low-quality data in survey research," *Journal of Business and Psychology*, vol. 33, no. 5, pp. 559–577, 2018.

[9] J. Rammelaere and F. Geerts, "Cleaning data with forbidden itemsets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1489–1501, 2020.

[10] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[11] R. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Chambell, "Introduction to machine learning, neural networks, and deep learning," *Translational Vision Science & Technology*, vol. 9, no. 2, 2020.

[12] L. Azzimonti, G. Corani, and M. Zaffalon, "Hierarchical estimation of parameters in Bayesian networks," *Computational Statistics & Data Analysis*, vol. 137, pp. 67–91, 2019.

[13] Y. Long, L. Wang, and M. Sun, "Structure extension of tree-augmented naive Bayes," *Entropy*, vol. 21, no. 8, 2019.

[14] K. Rahul and R. K. Banyal, "Detection and correction of abnormal data with optimized dirty data: a new data cleaning model," *International Journal of Information Technology and Decision Making*, vol. 20, no. 02, pp. 809–841, 2021.

[15] C. Ge, Y. Gao, X. Miao, L. Chen, C. S. Jensen, and Z. Zhu, "IHCS: an integrated hybrid cleaning system," *Proceedings of the Vldb Endowment*, vol. 12, no. 12, pp. 1874–1877, 2019.

[16] L. Ma, Q. Pei, L. Zhou, H. Zho, L. Wang, and Y. Ji, "Federated data cleaning: collaborative and privacy-preserving data cleaning for edge intelligence," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6757–6770, 2021.

[17] Y. Huang, M. Milani, and F. Chiang, "Privacy-aware data cleaning-as-a-service," *Information Systems*, vol. 94, 2020.

[18] S. Al-Janabi and R. Janicki, "Data repair of density-based data cleaning approach using conditional functional dependencies," *Data Technologies and Applications*, 2021.

[19] Q. Wang, Y. Guo, L. Yu, X. Chen, and P. Li, "Deep Q-network-based feature selection for multisourced data cleaning," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 16153–16164, 2021.

[20] X. Shi, C. Prins, G. Van Pottelbergh, P. Mamouris, B. Veas, and B. D. Moor, "An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, 2021.

[21] M. B. Gesicho, M. C. Were, and A. Babic, "Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: a case study of Kenya," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 2020.

[22] S. Liu, G. Li, S. Jiang et al., "Investigating data cleaning methods to improve performance of brain-computer interfaces based on stereo-electroencephalography," *Frontiers in Neuroscience*, vol. 15, 2021.

[23] J. Tian, D. Han, L. Xiao, and P. Shi, "Multi-scale deep coupling convolutional neural network with heterogeneous sensor data for intelligent fault diagnosis," *Journal of Intelligent and Fuzzy Systems*, vol. 41, no. 1, pp. 2225–2238, 2021.

[24] Y. Liu, "Multi-source heterogeneous data fusion based on perceptual semantics in narrow-band Internet of Things," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 413–420, 2019.

[25] J. Anees, H. Zhang, S. Baig, B. G. Lougou, and T. G. Robert Bona, "Hesitant fuzzy entropy-based opportunistic clustering and data fusion algorithm for heterogeneous wireless sensor networks," *Sensors*, vol. 20, no. 3, 2020.

[26] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020.

[27] R. Dominguez, M. Post, and A. Fabisch, "Common data fusion framework: an open-source common data fusion framework for space robotics," *International Journal of Advanced Robotic Systems*, vol. 17, no. 2, 2020.

[28] B. He, X. Cao, and Y. Hua, "Data fusion-based sustainable digital twin system of intelligent detection robotics," *Journal of Cleaner Production*, vol. 280, 2021.

[29] I. A. Al-Baltah, A. A. Abd Ghani, G. M. Al-Gomaei, F. H. Abdulrazzak, and A. A. Al Kharusi, "A scalable semantic data fusion framework for heterogeneous sensors data," *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[30] M. Simjanoska, S. Kochev, J. Tanevski, A. Madevska Bogdanova, G. Papa, and T. Eftimove, "Multi-level information fusion for learning a blood pressure predictive model using sensor data," *Information Fusion*, vol. 58, pp. 24–39, 2020.

[31] I. Ullah and H. Y. Youn, "Intelligent data fusion for smart IoT environment: a survey," *Wireless Personal Communications*, vol. 114, no. 1, pp. 409–430, 2020.

[32] K. K. Kumar, E. Ramaraj, and P. Geetha, “Multi-sensor data fusion for an efficient object tracking in internet of things (IoT),” *Applied Nanscience*, 2021.

[33] J. Lv, X. Wang, K. Ren, M. Huang, and K. Li, “ACO-inspired Information-Centric Networking routing mechanism,” *Computer Networks*, vol. 126, pp. 200–217, 2017.

[34] A. M. Altabeeb, A. M. Mohsen, L. Abualigah, and A. S. Ghalllab, “Solving capacitated vehicle routing problem using cooperative firefly algorithm,” *Applied Soft Computing*, vol. 108, pp. 1–10, 2021.

[35] R. T. Godwin and D. E. Giles, “Analytic bias correction for maximum likelihood estimators when the bias function is non-constant,” *Communications in Statistics-Simulation and Computation*, vol. 48, no. 1, pp. 15–26, 2019.