Hindawi
Complexity
Volume 2024, Article ID 9787219, 1 page
https://doi.org/10.1155/2024/9787219

Retraction

Retracted: Adaptive Language Processing Based on Deep Learning in Cloud Computing Platform

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Xu and C. Yin, "Adaptive Language Processing Based on Deep Learning in Cloud Computing Platform," *Complexity*, vol. 2020, Article ID 5828130, 11 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 5828130, 11 pages
https://doi.org/10.1155/2020/5828130

Research Article

Adaptive Language Processing Based on Deep Learning in Cloud Computing Platform

Wenbin Xu[1] **and Chengbo Yin**[2]

[1]*Department of English Language and Literature, China University of Petroleum (East China), Qingdao, Shandong 266580, China*
[2]*School of Data Science, Qingdao Huanghai University, Qingdao 266427, Shandong, China*

Correspondence should be addressed to Wenbin Xu; xuwenbin@upc.edu.cn

Received 28 March 2020; Revised 7 May 2020; Accepted 27 May 2020; Published 19 June 2020

Guest Editor: Zhihan Lv

With the continuous advancement of technology, the amount of information and knowledge disseminated on the Internet every day has been developing several times. At the same time, a large amount of bilingual data has also been produced in the real world. These data are undoubtedly a great asset for statistical machine translation research. Based on the dual-sentence quality corpus screening, two corpus screening strategies are proposed first, based on the double-sentence pair length ratio method and the word-based alignment information method. The innovation of these two methods is that no additional linguistic resources such as bilingual dictionary and syntactic analyzer are needed as auxiliary. No manual intervention is required, and the poor quality sentence pairs can be automatically selected and can be applied to any language pair. Secondly, a domain adaptive method based on massive corpus is proposed. The method based on massive corpus utilizes massive corpus mechanism to carry out multidomain automatic model migration. In this domain, each domain learns the intradomain model independently, and different domains share the same general model. Through the method of massive corpus, these models can be combined and adjusted to make the model learning more accurate. Finally, the adaptive method of massive corpus filtering and statistical machine translation based on cloud platform is verified. Experiments show that both methods have good effects and can effectively improve the translation quality of statistical machines.

1. Introduction

Currently, corpus-based translation system relies on large-scale bilingual parallel corpus, uses the translation model to estimate the probability, and selects the final translation result based on the translation probability. The advantage of the corpus-based translation method over the rule-based translation method is that it does not require much human and material participation in the construction of the model. The researchers themselves do not need to master the level of linguistic experts in the mastery of the two languages. The threshold is not so high, which allows more interested scholars and researchers to invest in it. Depending on the specific translation strategy, corpus-based machine translation can be divided into statistical-based machine translation and instance-based machine translation. The statistical-based method is the mainstream method of current machine translation.

The early stages of statistical machine translation development use only some coarse-grained features, such as bidirectional phrase translation probabilities [1–3], bidirectional lexical translation probabilities [4], vocabulary length penalties [5], phrase lengths, punishment [6], language model [7], and sequence model [8–10]. Many systems use only these 10–20 features to complete the translation process and use the minimum error rate training (MERT) method [11–13] to perform feature weight adjustment. With the development of statistical translation models and the widespread use of massive data, researchers have found that the use of fine-grained

features [14] can further improve the accuracy of the translation system. However, the use of a large number of fine-grained features poses a great challenge to the adjustment of feature weights. The traditional MERT method can only adjust the weights of dozens of features but cannot do anything for a translation system with thousands of features. References [15–17] proposed a training algorithm based on max-violation perceptron and forced decoding [18], which can be used to translate the system by using all bilingual training data, large-scale discriminative training, and support for tens of millions of sparse features. Compared to the MERT and PRO methods, this approach can bring very significant performance improvements [19, 20] and further maximizes the use of perceptual machine training methods. The hierarchical phrase translation system has also achieved good results. The traditional statistical machine translation domain adaptive method usually migrates the model for a single domain. For example, the training data is news corpus, and the test data is network corpus. However, most practically in the application scenario, it is necessary to perform model migration on multiple domains at the same time. For example, for online translation services, the user's input is usually text from various fields, which requires the statistical machine translation model to process automatically according to the actual input. The field adaptive research of deep learning translation is still relatively few, and the existing work has not given a clear domain label. However, the actual translation of scientific and technological literature often faces multiple professional fields, and the use of existing knowledge to organize information, such as the keywords of the paper, the scientific and technological word system, and other knowledge to obtain more clear semantic tags, helps to divide the corpus more finely.

In view of this, this paper mainly studies the multidomain adaptive method of statistical machine translation based on massive corpus under the cloud computing platform. Firstly, two corpus screening strategies are proposed, based on the double-sentence pair length ratio method and the word alignment information based method. The innovation of these two methods is that no additional linguistic resources such as bilingual dictionary and syntactic analyzer are needed as auxiliary. No manual intervention is required, and the poor quality sentence pairs can be automatically selected, and can be applied to any language pair. Secondly, a domain adaptive method based on massive corpus is proposed. The method based on massive corpus utilizes massive corpus mechanism to carry out multidomain automatic model migration. In this domain, each domain learns the intradomain model independently, and different domains share the same general model. Through the method of massive corpus, these models can be combined and adjusted to make the model learning more accurate. Finally, the adaptive method of massive corpus filtering and statistical machine translation based on cloud platform is verified. Experiments show that both methods have good effects and can effectively improve the translation quality of statistical machines.

2. Massive Corpus Screening Strategy under Cloud Computing Platform

2.1. Cloud Computing Platform Framework. The Hadoop Distributed File System (HDFS) can be deployed on a large number of inexpensive machines to store up terabytes and petabytes of data in a highly fault-tolerant and reliable manner. It combines well with the MapReduce model to provide high-throughput data access. The structure of DFS is shown in Figure 1.

As can be seen in Figure 1, an HDFS cluster that consists with a NameNode and multiple DataNodes was discussed. The metadata and the DataNode are actual data. The application accesses the NameNode to get the metadata of the file, and the actual I/O operation is directly interacting with the DataNode. The NameNode is the primary control server responsible for managing the file system namespace and coordinating application access to files, recording any changes to the namespace or changes to their properties. The DataNode is responsible for storage management on the physical node where the file is located. The feature of HDFS is that the data is "write once, read many times." The files of HDFS are generally divided into different data blocks according to a certain size, and each data block is dispersed into different DataNodes as much as possible. In addition to completing the namespace operation of the file system, the NameNode also determines the mapping of the data block to the DataNode.

2.2. Massive Corpus Screening Strategy. For statistical machine translation systems, the intuitive understanding is that increasing the size of the training data can help improve system performance. Massive data is easier to obtain in today's information environment than ever before. Scholars have built knowledge bases such as parallel sentence pairs and bilingual dictionaries by crawling bilingual web pages [21]. There are more and more sources of corpora, from multilingual websites, comparable bilingual corpora, human translated text, and more. The scale of building parallel corpora has been large, and it can be used for statistical machine translation system training. Too many errors must affect statistical machine translation systems that rely on data quality. In view of the fact that there is no qualitative change in the current statistical model, it is necessary to acquire the model features by training the corpus. Therefore, in order to train a high-performance statistical machine translation system, it is necessary to process and screen the training data. In this paper, two methods are used to filter the noise sentence pairs in the bilingual parallel corpus: the method based on the sentence pair length ratio and the method based on the alignment information.

2.2.1. Method Based on Sentence to Length Ratio. In general, the length of a pair of statements that are translated should be proportional to a certain ratio. However, most parallel corpora contain sentence pairs that do not match the length ratio. These sentence pairs are usually noise in the corpus. Noise phenomena caused by the length ratio include

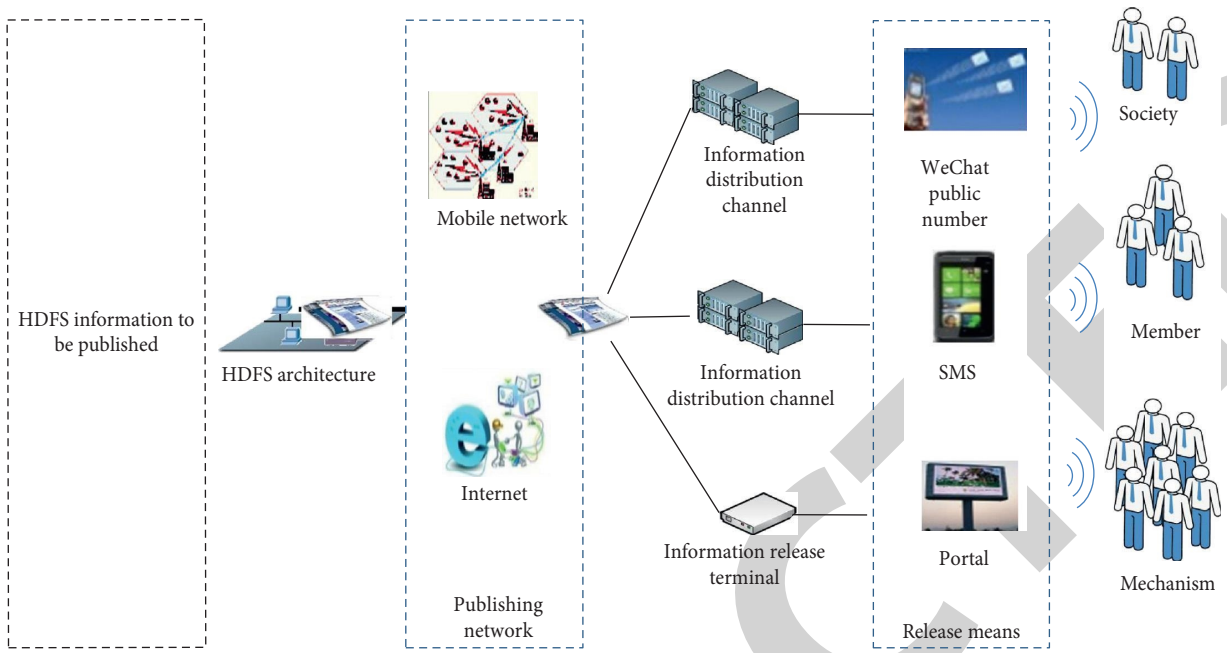

Figure 1: HDFS and DFS structure diagram.

monolingual errors, alignment errors, and inclusion of unknown tags (html tags, etc.). These phenomena have been observed, and many noise sentence pairs whose length ratios do not conform to the regularity are found in the experimental corpus. Some examples are listed in Table 1.

In order to remove such pairs of erroneous sentences, we set a length rule that defines the length ratio as

$$\mathrm{LR}(f, e) = \frac{|f|}{|e|}, \tag{1}$$

where f is the source sentence, e is the target sentence, and $|f|$ and $|e|$ are the number of words after the source and target sentences.

The method based on the length ratio is usually based on linguistic knowledge, and the artificially set length is lower than the upper and lower limits. This paper assumes that the noise sentence in the corpus is less than the normal sentence pair; that is to say, there is a continuous range of ratios, which is the majority of the normal sentence pairs in this range. Therefore, the threshold is set according to the statistical distribution characteristics of the length ratio; that is, the sentence pairs whose length is less than the total number are filtered out. This has the advantage that thresholds can be set for different language pairs without the need for specific linguistic knowledge.

2.2.2. Method Based on Word Alignment. The word alignment problem is the task of finding the alignment of words in a given two-state pair. It is a key step in statistical machine translation. The word alignment model has been studied for a long time, and people use different methods for bilingual word alignment. Run the IBM model from both directions and merge the results of the two word alignments. In general, the intersection contains relatively reliable

Table 1: Noise sentence pair example.

Sentence example	Error
(of an enterprise, institution, etc.) Shift to another system	Sentence level alignment is incorrect
Falco jugger &bra; laggar falcon &ket;	Source unknown tag
c & w &bar; c-and-w &bar; country-and-western	Unknown label
1. (of a liquid) to settle, 2. clear; transparent	Sentence level alignment is incorrect
Take cognizance of	Incorrect content

alignment points; that is, the alignment point is highly accurate but does not contain all reliable alignment points; and the assembly contains most of the desired alignment points; that is, the recall rate is high but introduces additional errors. A good alignment point is adjacent to other good alignment points. Therefore, the algorithm starts from aligning the intersections. In the expansion step, adjacent alignment points located in the union but not in the intersection are added, and finally points that are not aligned in both directions are added. The pseudocode for this algorithm is given in Table 2.

The position of the two sentences in the corpus is adjacent. The occurrence of this situation is the automatic noise extraction of the parallel sentence to the technology, because it is impossible to judge the correct alignment sentence pair (the correct alignment sentence corresponds to <discountedant opinions, discordant opinions>, or <discondant opinions> which is the second sentence in the table). A similar situation has occurred many times in the corpus we use.

Figure 2 shows the alignment matrix of the two sets of sentence pairs. As shown in Figure 2, it shows the results of two unidirectional alignments in English and Chinese, and

Table 2: Pseudocode for grow-diag-and-final heuristic word alignment extension.

```
GROW DIAG-AND-FINAL(e2ffle):
neighboring = ((−1, 0), (0, −1), (1, 0), (0, 1),
                (−1, −1), (−1, 1), (1, −1), (1, 1))
alignment = intersect(e2f, f2e);
GROW-DIAGQ; FINAL(e2); FINAL(f2);
GROW DIAL-ANDQ:
  iterate until no new points added
    for English word e = 0, 1, 2. . .en
      for Foreign word f = 0, 1, 2. . .fh
        if (e aligned with f)
          for each neighboring point(e-new, f-new):
            if(e-new not aligned or f-new not aligned)and
              (e-new, f-new)in union(e2f, f2e))
              add alignment point(e-new, f-new)
FINAL(a):
  for English word e-new = 0, 1, 2. . .en
    for foreign word f-new = 0. . .fn
      if(e-new not aligned or f-new not aligned)and
        (e-new, f-new)in union(e2f, f2e))
add alignment point(e-new, f-new)
```

the bidirectional alignment matrix on the right is obtained by the grow-diag-and-final algorithm. It can be seen that the intersection of two unidirectional alignments is only a matter of discretion, discordant; that is to say, when the grow-diag-and-final extension is performed, there is only one alignment result that is originally considered reliable. After the expansion, the result of the obvious error alignment is obtained. This error not only affects the alignment quality of itself, but also affects the rule extraction result of the translation system. For example, in the phrase system, the above sentence pairs are extracted from the rules of discriminating, inconsistent. This kind of rule does not play any role in translation decoding; even if the decoder selects the rule, it will only reduce the translation quality. Therefore, similar problems should be avoided as much as to improve the quality of the translation or to reduce the size of the rule set.

To this end, we propose a sentence-pair filtering method based on the grow-diag-and-final extension method. We consider expanding the number of alignment results EC and the number of alignment results of the intersection alignment IC. When the extended alignment result exceeds the intersection alignment result by a certain amount, we think that the alignment result is unreliable. We set the filtering rules based on the word alignment extension and use the following to judge whether the word alignment results are reliable:

$$\mathrm{ER}(f, e) = \frac{\mathrm{IC} - \mathrm{EC}}{\mathrm{UC}}. \tag{2}$$

3. Statistical Machine Translation Adaptation Based on Massive Corpus

3.1. Statistical Machine Translation Adaptation. This section introduces a domain adaptive approach based on massive corpus. The main idea is that the training of our model is mainly divided into three steps: first, selecting the data in the domain is according to the predefined domain; second, training the domain model and the general model is to construct the statistical machine translation system; third, using massive corpus technology makes joint adjustments to multiple domain systems.

According to the above, the first step in this work is to select the in-domain bilingual control data from all the bilingual training data to train the translation model. Since the monolingual data in a specific field can be obtained in large quantities, we draw on the method of bilingual cross section data selection [22] to obtain bilingual data in the field:

$$\left[H_{I-\mathrm{src}}(s) - H_{G-\mathrm{src}}(s)\right] + \left[H_{I-\mathrm{igt}}(t) - H_{G-\mathrm{igt}}(t)\right]. \tag{3}$$

This bilingual cross-entropy-based criterion tends to choose a sentence pair that is more similar to the data distribution in the domain but different from the general data distribution. Therefore, this method considers that the sentence pair with larger cross-entropy difference should be selected.

In the second step, we use the training data in the selected domain to build a statistical machine translation system based on the hybrid model. Specifically, we adopted the idea of a hybrid model to build N machine translation systems for N predefined fields; each of which is a log-linear model. For each system, the optimal translation result f is given by

$$f = \arg\max_{r}\{P(f \mid e)\}. \tag{4}$$

For each machine translation system, two translation models and two language models are included. The translation model of a specific field is trained by the bilingual data selected by the data selection method introduced in the previous section, and the translation model of the general domain is trained using all bilingual data. For the language model, we reuse the language-specific and general-language models of the specific domain trained for data selection in the previous section. Compared to a translation system that does not do domain migration, this system with a hybrid model can better balance the general translation knowledge and domain-specific translation knowledge and can benefit from two aspects.

In the third step, it is necessary to adjust the feature weights in different machine translation systems. The traditional method of arranging is generally directed to a single system. The method described in this section regards translation systems in different fields as related translation tasks, and joints are coordinated under the framework of massive corpus. There are two reasons for using massive corpora:

(1) The translation system of a specific domain shares the same general domain translation model and language model, and the massive corpus mechanism can make better use of the common translation knowledge of translation tasks in different fields.

(2) By forcing the general domain translation model and the language model to behave the same in different

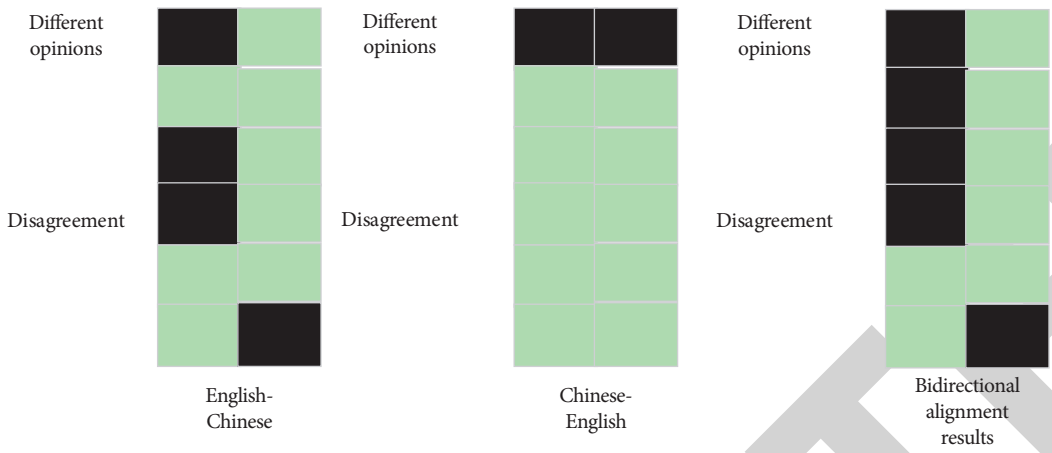

Figure 2: Schematic diagram of the alignment matrix of the wrong sentence pair.

fields, massive corpus provides a regularization mechanism to prevent model overfitting. Formally, the objective function of using massive corpus to adjust parameters is represented by the following formula:

$$\min_{w}\left\{\sum_{i=1}^{N}\mathrm{Loss}\left(E_i, e\left(F_i, w_i\right)\right)\right\}. \tag{5}$$

In order to be able to efficiently coordinate the parameters, we have improved an asynchronous stochastic gradient descent algorithm to optimize and borrowed the idea of pairwise ranking to use the perceptron algorithm to update the feature weights.

$$L\left(w_i\right) = \left(-\left(w_i, v^1 - v^2\right)\right). \tag{6}$$

We first use the machine translation system to generate the N best translation result candidates (N-best), which are reordered and combined into pairs by scoring with smooth sentence level BLEU. Specifically, similar to the asynchronous gradient descent algorithm, we divide the N best translation result candidates into three parts: the best 10% (high), the middle 80% (middle), and the worst 10% (low). These three parts of the translation result candidates are used for two-two sorting, in which we choose "high one," "medium one low," and "high one low" to combine in pairs, but will not select two of the same part Candidate combinations that are paired. The basic idea of constructing a sample in this way is that the algorithm can better have the discriminability of distinguishing between high quality and low quality translation results.

3.2. Neural Network Deep Fusion Model. The algorithm based on domain knowledge uses the explicit discrete features of domain knowledge, and the deep labeling algorithm uses the hidden continuous features of deep learning. The sentence domain probability vectors obtained by the two methods are different.

Combining the domain labeling algorithm based on domain knowledge and the domain labeler based on deep learning, a multilayer perceptron based on the top layer is designed as a deep fusion model of the neural network. The architecture is shown in Figure 3. The preprocessing of the sentence to be labeled is mainly word segmentation and garbled filtering. The preprocessed results are input to the knowledge-based domain tagger and deep learning-based domain tagger to obtain the domain knowledge-based probability vector and probability vectors for deep learning. The top-level neural network deep fusion model is a two-layer perceptron, and the hidden layer is two receiving four-dimensional vectors. Neuron, the activation function, is set to the ReLU (Rectified Linear Unit) function.

$$\begin{aligned} y_1 &= \mathrm{ReQ}\left(W_1 x_1 + b_1\right), \\ y_2 &= \mathrm{ReQ}\left(W_2 x_2 + b_2\right). \end{aligned} \tag{7}$$

The deep mixed neural model obtained through this fusion will well combine explicit and invisible knowledge, merging the advantages of discrete features and continuous features, and make the probability vector and decision category of each sentence more accurate. Thereby, the adaptation problem in the field of machine translation is better improved, and, for the data to be translated in a specific field, a higher quality translation output will be obtained.

4. Experiments and Results

4.1. Massive Corpus Screening Experiment Verification. The experimental part of this paper runs on a separate server. The specific software and hardware configuration is shown in Table 3. Because Hadoop installation is a stand-alone mode, the comparison and analysis of experimental results focus on the impact of the proposed method on translation quality.

Under the cloud translation platform, bilingual parallel corpora are a wide range of sources, such as translated manuscripts completed by translators, officially published bilingual materials, and automatic extraction of multilingual web pages. The quality of corpus is uneven. Therefore, in order to test whether the method is effective, the training

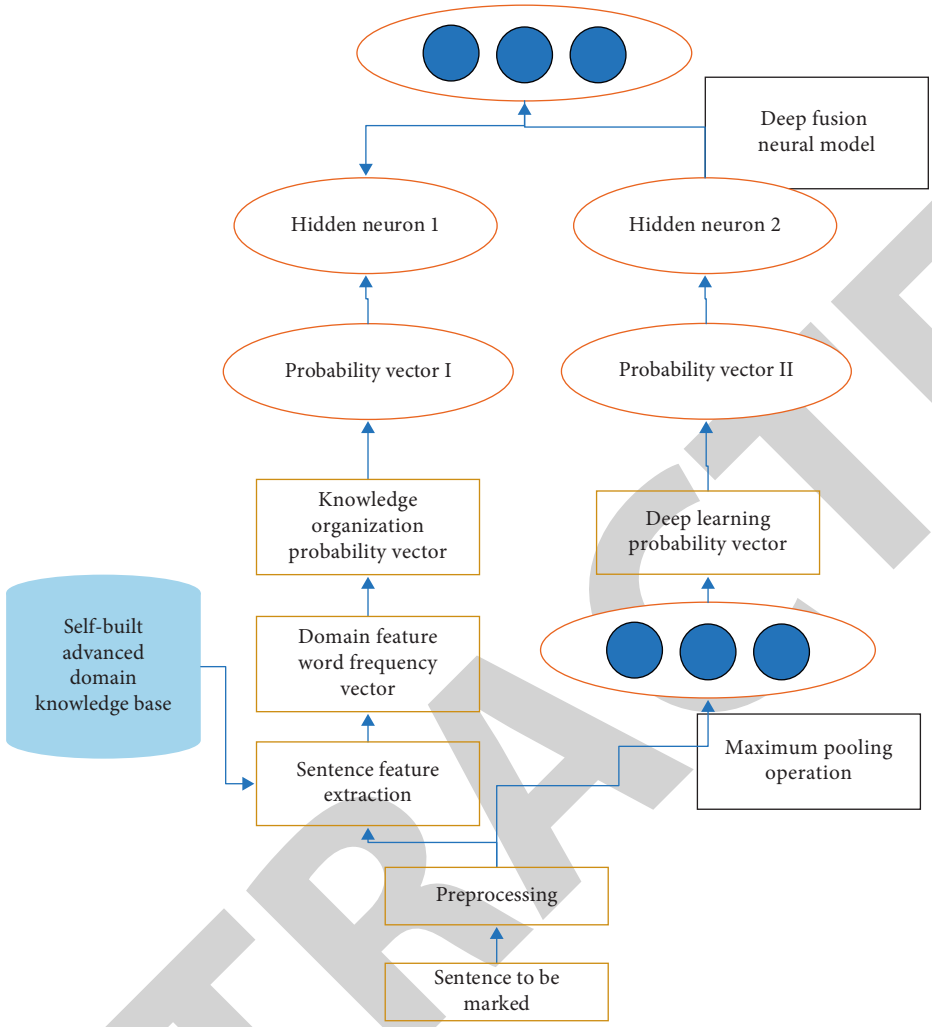

Figure 3: Neural network deep fusion model framework.

Table 3: Experimental environment.

CPU	RAM (GB)	Operating system
Intel (R)-Xeon (R) 11 2.93 Hz	96	Ubuntu 12.04.1

corpus selected is mostly from the network. It contains 1,937,289 bilingual parallel sentences as a training set, which mixes multiple fields of content. The English language model uses the Xinhua part of LDC2007T07, the test set, and the development set. The test set contains a British news review in cwmt2011, and nist02 to nist08. The data sets are specifically listed in Table 4.

The tools used in the experiment include open source word alignment tool GIZA ten +, open source statistical machine translation system Moses, open source language model training tool irstlm, and Chinese word segmentation tool ictclas. The translation model used is the phrase model, and the parameters use the standard settings of Moses.

First, we count the distribution of the length ratio of the English-Chinese sentence in the training set. The result is shown in Figure 4. The ordinate indicates the number of sentence pairs, and the abscissa indicates the ratio of the

Table 4: Source of experimental data.

Corpus type	Corpus name	Corpus size
Training set	web.ch-en	1,937,289 sentence pair
Test set	cwmt2011	1006 sentence pair
	nist02	878 sentence pair
	nist03	919 sentence pair
	nist04	1788 sentence pair
	nist05	1082 sentence pair
	nist06	1664 sentence pair
	nist08	1357 sentence pair
	LDC2007T07	9,685,593 sentence pair

length of the sentence between English and Chinese. We can find the ratio of the length of the sentence to a certain distribution law. When the ratio is 1.0, the sentence number is the most, and there are 297,341 pairs of sentences. The figure shows that the highest point of the ratio (1.0) is also relatively large in number of sentences. This verifies our hypothesis that the length ratios of the two languages conform to the law in a continuous range.

As shown in the above figure, the contrast value of the sentences appearing in the corpus is 0.1–66.0 due to the

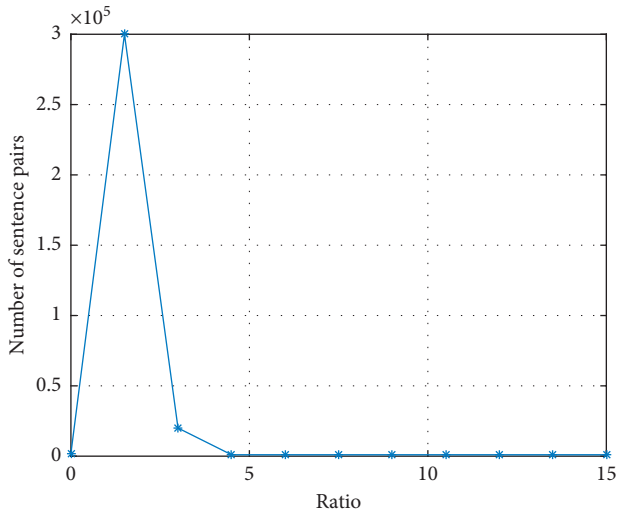

Figure 4: Statistics of length-to-length ratios of English-Chinese sentences.

influence of noise and domain differences. We use the training corpus used for statistics. The comparison value of %.06% falls within the range of 0.4–3.8. Even in the range of 0.5–4.5, there is still 90.11% of the sentence pairs. To this end, we screened the corpus training comparison system for the different ratio ranges of more than 90% of the total corpus and compared the BLEU scores of the systems on the test set. As shown in Table 5, the statistical distribution of the percentage of total corpus pairs in different ratio ranges is listed. The first line represents the corpus used, and the remaining lines represent the number of sentence pairs contained in the different ratio ranges and the percentage of the corpus. Table 5 shows the number of sentence pairs retained and the proportion of the total corpus when the ER filters different pairs of sentences. The higher the score of ER, the better the effect of word alignment, and the more reliable we think the result is. Based on the method of word alignment information screening corpus, we consider two cases: use the filtered sentence to align the retraining words and then get the translation model; directly use the filtered sentence pairs and alignment information to train the translation model. In the first case, the filtered noise information may affect the calculation of the word alignment probability during the iterative process of word alignment. Realigning after filtering out may improve the word alignment quality and improve the translation effect. In the second case, we use ER to retain the word alignment that is considered reliable, so reword alignment or different alignment results may occur due to the change of word alignment probability, and there may be unreliable alignment results in these results. The experimental results are shown in Table 6.

It can be seen from the experimental results that the BLEU scores of each test set are improved in both cases. As far as the overall effect is concerned, it is better not to retrain the word alignment. However, in both cases, the improvement effect on the nist03 and nist05 test sets is not very obvious, the effect of reword alignment on nist03 is slightly better than the latter,

Table 5: Scores of 8 length ratios on the development set after filtering.

t	Ratio range	Number of sentences	Percentage (%)
0.000	03–2.0	1,911,491	100
0.004	0.4–2.0	1,876,177	98.15
0.005	0.4–1.8	1,867,066	97.68
0.010	0.4–1.7	1,837,419	96.12
0.012	0.5–1.7	1,814,507	94.93
0.014	0.5-i.b	1,791,698	93.73
0.020	0.5-I.5	1,764,233	92.30
0.022	*	1,722,475	90.11

Table 6: BLEU scores based on word alignment filtering (reword alignment).

ER	cwmt 2011	nist 02	nist 03	nist 04	nist 05	nist 06	nist 08
*	22.87	30.11	28.31	29.44	27.93	24.02	18.87
>−0.5	23.00	30.03	27.87	29.53	27.59	24.30	19.06
>−0.4	22.94	30.43	28.05	29.37	27.65	24.35	19.02
>−0.3	23.21	30.64	28.41	29.52	27.47	24.09	19.20
>−0.2	23.11	30.36	28.16	29.45	27.64	24.08	19.40
>−0.1	23.56	29.59	28.21	29.93	27.14	24.42	19.91

and the opposite is on nist45. Use ER to determine whether the word alignment is reliable. When ER is lower than the given threshold, we think that the word alignment result of the sentence pair is not reliable overall. We will filter out the sentence pair, that is, the sentence pair. All alignment information is deleted. In fact, in the word alignment result of the sentence pair, there will be some correct word alignment information; that is, the correct word alignment information is also deleted while deleting the error alignment information. Although the wrong information is not useful for translation tasks, the correct information to be deleted may be helpful for translation tasks. Therefore, there is also the possibility of reducing the BLEU score.

As shown in Figure 5, we find an instance from the filtered sentence pair to illustrate. The thick solid line in the figure is the intersection of two aligned directions, the thin solid line is the correct result of the extended alignment, and the dashed line is the wrong result of the extended alignment. The sentence in Figure 4 is correct for itself, but its alignment information is incorrect; only partial alignment is correct, and its ER value is −0.22; it can be seen that there is some correct word alignment information, and this part of information can be extracted. A rule facilitates translation. Because the ER value is lower than the given threshold, all alignment information of the sentence pair is filtered out, but the correct information is also filtered out, so there is a problem of degraded translation quality.

4.2. Adaptive Experimental Verification of Statistical Machine Translation Based on Massive Corpus. We compared the impact of different searched documents and hidden layer lengths on the accuracy of the translation system. The results are shown in Figure 6.

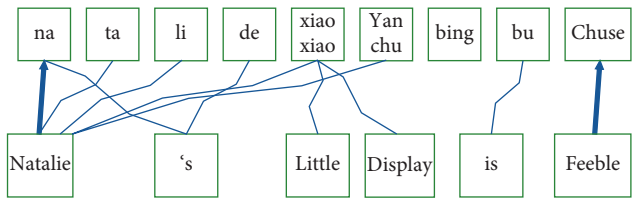

Figure 5: Example of incorrect alignment of correct sentence pairs.

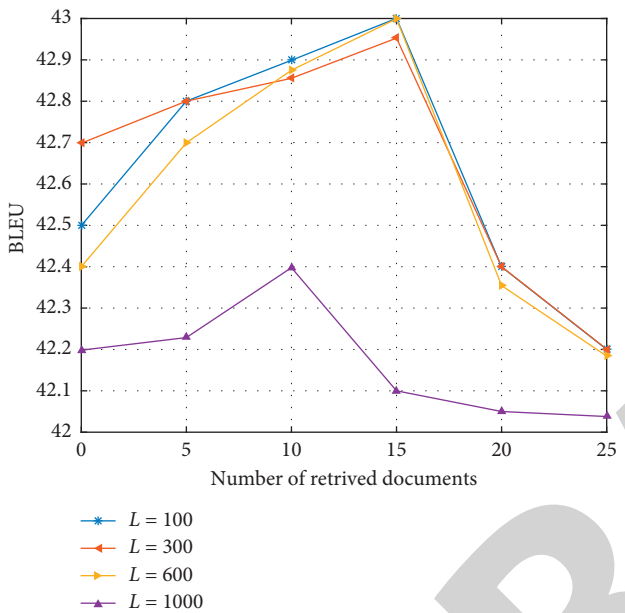

Figure 6: The effect of different number of retrieved documents and hidden layer length on machine translation accuracy.

As shown in Figure 6, we found that, for most results, the optimal translation accuracy was obtained when the number of retrieved documents was $N = 10$. This result confirms that extending the source language input by means of information retrieval is very helpful for determining topic information. It plays an important role in the selection of translation rules. However, in the experiment, when N is large, for example, $N = \_50$, the translation performance is drastically lowered. This is because as the number of retrieved documents increases further, the introduction of topic-independent documents into the neural network will be introduced, and irrelevant documents will bring about unrelated real words, thus affecting the performance of neural network learning.

In Figure 7, it can be seen that when L is small, the translation system is relatively accurate. In fact, in the case of $L < 600$, the difference in translation performance is small. However, when $L = 1000$, the translation accuracy is worse than other cases. The main reason is that the amount of parameters in the neural network is so large that it cannot be well studied. We know that when $L = 1000$, there are 100000×1000 parameters between the linear and nonlinear layers of the network. The current training data size is not enough to support this network parameter level training, so the model is likely to fall into the local Optimal and unacceptable topic representation information.

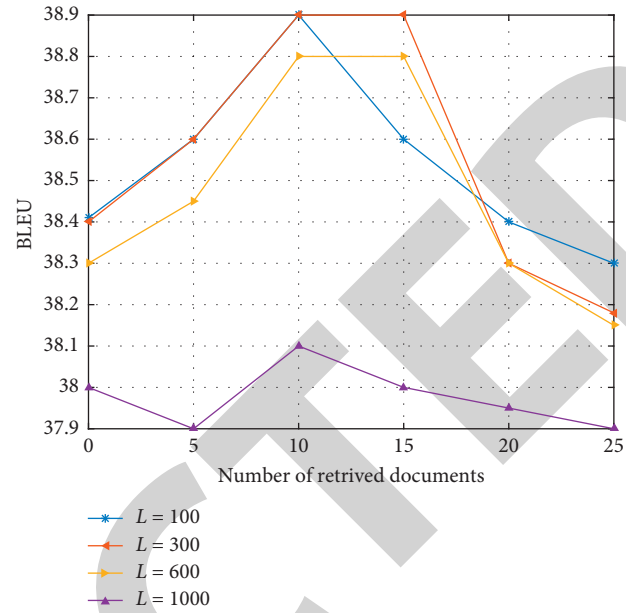

Figure 7: The impact of topic-related features on machine translation accuracy.

As shown in Figure 8, we find that the topic similarity feature on the source language side is slightly better for the system than the target language-side similarity feature, and the enhancements that they bring can be accumulated, which means that the neural network based on bilingual data training can help the statistical machine translation system Translation result candidates perform better disambiguation. Further, based on the similarity feature, the topic sensitivity feature of the translation rule can bring more performance improvement, because the translation rules of specific topics are usually more sensitive, when the similarity is similar. The system tends to choose translation rules for specific topics rather than general translation rules. Finally, we see that our methods perform best when using all topic-related features, with an average of 0.39 BLEU points higher than the LDA-based method.

As shown in Figure 9, we use the method of information retrieval to extend the original input, thus avoiding restrictions on bilingual chapters. We use neural network technology for topic modeling. The algorithm is more practical and has good scalability. Under the deep learning framework, our method directly optimizes the bilingual topic similarity, so that the learned topic representation can be easily integrated into statistical machine translation.

4.3. Domain Labeling Performance. The statistics of the data set are used by the domain tagger; 1% is randomly selected as the test data for the domain tagging performance experiment. The training data for training deep learning is selected from the remaining 99% of the size of the data set. At the same time, the domain labeler based on the domain knowledge and the trained deep learning domain labeler are used to label the test data to obtain the category and probability vector, and then the results are simply linearly

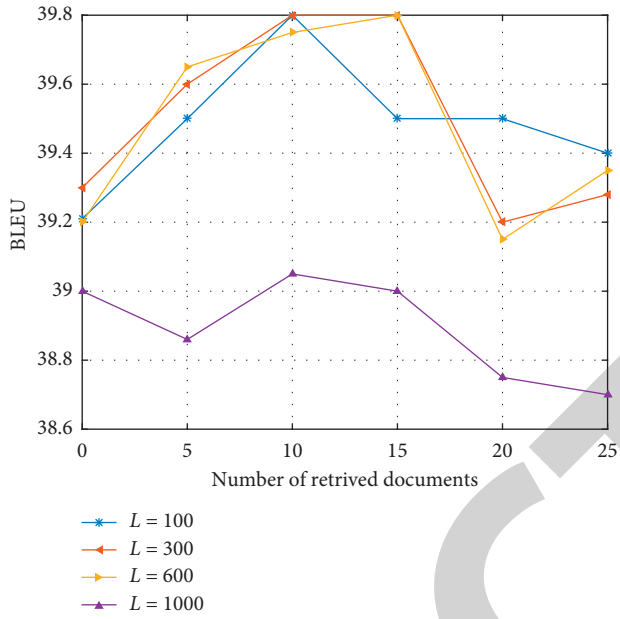

Figure 8: The impact of different features on the translation system.

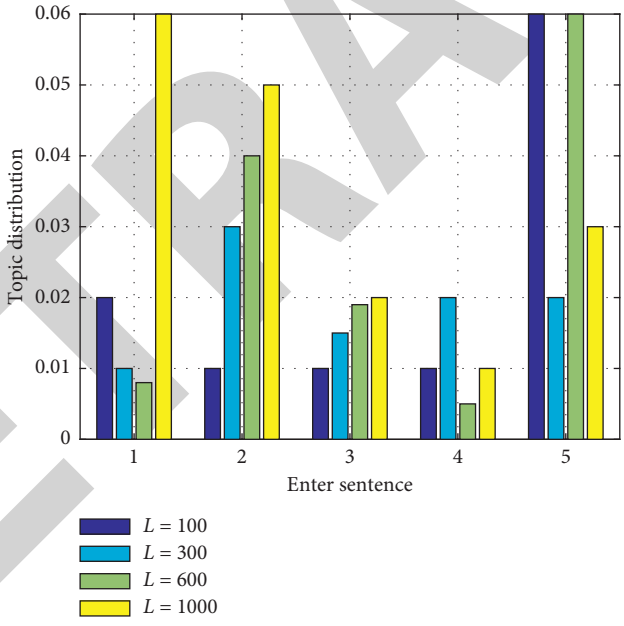

Figure 9: Topic distribution information of three sentence candidates.

fused. The category with the highest probability in the probability vector is selected as the final judgment category, and this judgment category is used as the statistical basis for the accuracy rate and the recall rate and the F-1 value. Four subexperiments are performed. The results are shown in Table 7. The results show that using only the domain knowledge tagger will cause misjudgment and omissions due to the lack of self-built domain knowledge base feature words, so the score is not high, but the judgment efficiency is high; using only the deep learning domain tagger requires a lot of the training data belonging to mining tacit knowledge and continuous features, but training is slow and does not combine prior knowledge; simple linear fusion models linearize the probability vectors of the first two and so on.

Proportional weighting combines explicit and tacit knowledge but this simple fusion is difficult to improve most of the misjudgments and missed judgments and has not been greatly improved; the final neural network deep fusion model is deepened through multilayer neural networks. The integration, giving full play to the advantages of the two, greatly reduced the phenomenon of misjudgment and omission and significantly improved.

Table 7: Effect of each labeling method on labeling performance.

	Domain knowledge	Deep learning	Simple linear fusion model	Deep fusion model
Accuracy	0.85	0.9185	0.898	0.947
Recall	0.76	0.95	0.886	0.9345
F-1 value	0.7633	0.9531	0.8936	0.942

5. Conclusion

When the training corpus is small, some pairs of training sentences related to the test text may be filtered out, which will affect the quality of the translation. But when the training data is large enough, such problems will hardly occur. In addition to learning the domain model independently for each domain, different domains share the same general model. Through the method of massive corpus, these models can be combined to make the model learning more accurate. The experimental results show that this method can significantly improve the translation accuracy of multiple fields in large-scale machine translation tasks. In addition, the performance of this joint tuning method is better than independent model migration. At the same time, this result can be easily applied to the online translation system, training different models for a predetermined number of fields, determining the domain according to the input source language text, and selecting the corresponding domain model or general model for translation. The experimental results also show that when there is no such problem, the method of this paper can effectively improve the translation quality of the statistical machine translation system.

The work that needs to be improved in this study is as follows. (1) Consider in the field that the adaptive mechanism is placed in the architecture of human neural machine translation. Calculate different domain vectors, improve the attention mechanism, and make the domain adaptive. (2) How to integrate deep learning methods and prior knowledge to improve the system's performance will be researched in every area of natural language processing in the future. Later, we will try different ways in neural machine translation Chinese and Canadian prior knowledge in the field to improve translation quality for different fields. How to integrate deep learning methods and prior knowledge to improve the performance of the system will be a problem that needs to be studied in the future. After that, we will try to add a priori knowledge in the field of neural machine translation in different ways to improve the translation quality for different fields.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Liu, C. M. Vong, and P. K. Wong, "Extreme learning machine for huge hypotheses Re-ranking in statistical machine translation," *Cognitive Computation*, vol. 9, no. 2, pp. 285–294, 2017.

[2] S. Zhu, Y. Yang, M. I. Chenggang, X. Li, and L. Wang, "Corpus selection for Uyghur-Chinese machine translation based on bilingual sentence coverage," *Journal of University of Science & Technology of China*, vol. 47, no. 4, pp. 283–289, 2017.

[3] K. Kim, E.-J. Park, J.-H. Shin, O.-W. Kwon, and Y.-K. Kim, "Divergence-based fine pruning of phrase-based statistical translation model," *Computer Speech & Language*, vol. 41, pp. 146–160, 2017.

[4] S. Salami, M. Shamsfard, and S. Khadivi, "Phrase-boundary model for statistical machine translation," *Computer Speech & Language*, vol. 38, pp. 13–27, 2016.

[5] J. Shang, J. Liu, J. Meng et al., "Automated phrase mining from massive text corpora," *IEEE Transactions on Knowledge & Data Engineering*, vol. 30, no. 10, pp. 1–15, 2018.

[6] S. Al-Dohuki, Y. Wu, F. Kamw et al., "SemanticTraj: a new approach to interacting with massive taxi trajectories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 11–20, 2017.

[7] F. S. Al-Anzi and D. Abuzeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent semantic indexing," *Journal of King Saud University—Computer and Information Sciences*, vol. 29, no. 2, pp. 34–45, 2017.

[8] M. C. Swain and J. M. Cole, "ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[9] Y. Yi, Q. Yao, and H. Qu, "VISTopic: a visual analytics system for making sense of large document collections using hierarchical topic modeling," *Visual Informatics*, vol. 1, no. 1, pp. 45–57, 2017.

[10] J. P. A. Ioannidis, "The reproducibility wars: successful, unsuccessful, uninterpretable, exact, conceptual, triangulated, contested replication," *Clinical Chemistry*, vol. 63, no. 5, pp. 943–945, 2017.

[11] M. Rahnemoonfar, G. C. Fox, M. Yari, and J. Paden, "Automatic ice surface and bottom boundaries estimation in radar imagery based on level-set approach," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 55, no. 9, pp. 1–8, 2017.

[12] J. Thomas and A. Zaytseva, "Mapping complexity/human knowledge as a complex adaptive system," *Complexity*, vol. 21, no. S2, pp. 207–234, 2016.

[13] F. Baertling, B. Alhaddad, A. Seibt et al., "Neonatal encephalocardiomyopathy caused by mutations in *VARS2*," *Metabolic Brain Disease*, vol. 32, no. 1, pp. 1–4, 2016.

[14] M. A. Raquet, G. J. Measey, and J. M. Exbrayat, "Annual variation of ovarian structures of Boulengerula taitana (Loveridge 1935), a Kenyan caecilian," *African Journal of Herpetology*, vol. 64, no. 2, pp. 116–134, 2016.

[15] C. Hedbergoldfors, N. Darin, and A. Oldfors, "Muscle pathology in Vici syndrome—a case study with a novel mutation in *EPG5* and a summary of the literature," *Neuromuscular Disorders*, vol. 27, no. 8, pp. 840–858, 2017.

[16] E. X. Fang, M. D. Li, M. I. Jordan, and H. Liu, "Mining massive amounts of genomic data: a semiparametric topic modeling

approach," *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1–15, 2017.

[17] M. Knight, "The accused is entering the courtroom: the live-tweeting of a murder trial," *Journal of Media Practice*, vol. 18, no. 2-3, pp. 1–26, 2017.

[18] K. Hörtnagel, I. Krägeloh-Mann, A. Bornemann et al., "The second report of a new hypomyelinating disease due to a defect in the VPS11 gene discloses a massive lysosomal involvement," *Journal of Inherited Metabolic Disease*, vol. 39, no. 6, pp. 849–857, 2016.

[19] A. Takagi, H. Ozawa, M. Oki, H. Yanagi, K. Nabeshima, and N. Nakamura, "*Helicobacter pylori*-negative advanced gastric cancer with massive eosinophilia," *Internal Medicine*, vol. 57, no. 12, pp. 1715–1718, 2018.

[20] W. Kojima and K. Hayashi, "Changes in the axo-glial junctions of the optic nerves of cuprizone-treated mice," *Histochemistry & Cell Biology*, vol. 149, no. 5, pp. 1–8, 2018.

[21] M. Kala, M. V. Shaikh, and M. Nivsarkar, "Equilibrium between anti-oxidants and reactive oxygen species: a requisite for oocyte development and maturation," *Reproductive Medicine and Biology*, vol. 16, no. 1, pp. 28–35, 2017.

[22] M. Betanzos, M. R. Costa-jussà, and L. Belanche, "Tradares: a tool for the automatic evaluation of human translation quality within an MOOC environment," *Applied Artificial Intelligence*, vol. 31, no. 1, pp. 1–10, 2017.