www.cadanda.com Computer-Aided Design AND APPLICATIONS™

Building a Parallel Corpus for English Translation Teaching Based on Computer-Aided Translation Software

Wenwen Xie[1] and Xiaolu Wang[2]

[1]College of Sports Science and Technology, Wuhan Sports University, Wuhan, Hubei 430205, China, lwp1225@163.com
[2]College of Sports Science and Technology, Wuhan Sports University, Wuhan, Hubei 430205, China, wang-xianming@hotmail.com

Corresponding author: Wenwen Xie, lwp1225@163.com

Abstract. This paper conducts an in-depth study on the construction of parallel corpus for English translation teaching through computer-aided translation software, and this study adopts a combination of corpus statistics and analysis to portray and study the system. In the corpus statistics, we carried out a detailed data analysis of the translated text and the original text with the help of corpus statistics software to find out the linguistic characteristics of the translated text and the areas that can be improved from the perspective of data statistics. After that, based on frequency, mean value, grading value, and multiplication analysis, we carry out qualitative analysis to study the movie subtitle translation strategy and its causes from multiple perspectives. This thesis aims to make research on film and television translation from a new angle with the help of corpus translation and to make up for the gap in the research direction of film and television translation. Firstly, through the corpus analysis methods such as frequency value, mean value, graded value and multiplier value, a quantitative analysis of film texts is carried out from three levels, namely, high frequency, medium frequency and basic, and it is found that Chinese films have many distinctive characteristics in terms of long words and sentences, words and sentences, beginning and end of sentences, idioms and the use of four-character words.

Keywords: computer assistance; English translation software; corpus construction.
DOI: https://doi.org/10.14733/cadaps.2021.S3.12-22

1 INTRODUCTION

With the dramatic increase in computing power and the rapid development of Internet technology, the study of natural language processing has been intensified. Against this background, corpus linguistics has gradually grown in strength. A corpus, also known as a corpus or material, is a warehouse or database for storing all kinds of linguistic materials and is a database of linguistic materials collected and scientifically organized, which is very helpful for language research [1]. Especially in recent years, with the rise of deep learning, most natural language processing tasks are trained on a corpus. As a result, the construction and study of the corpus have gradually become a hot topic [2]. Their main function is to transmit objective information, and they follow the principles of accuracy, conciseness, and clarity in language expression to ensure the objectivity and logic of the content. Like the objects, purposes, and audiences of the two are different, there are certain differences in vocabulary and grammar between them. For example, in terms of vocabulary diversity, they both have their specialized vocabulary, and the richness of vocabulary is completely different [3]. Of course, English for scientific literature and English for journalism also share many common properties [4]. In journalistic English, the subject is mainly objective facts and events, and the passive voice attracts readers' attention more. Finally, on the syntactic level, both scientific literature and journalistic English tend to use long and complex sentences. Some complex scientific concepts are difficult to explain in simple sentences, so the use of long and complex sentences in the scientific literature can fully reflect the thoroughness, rigor, and structure of scientific style. The purpose of using complex sentences in English news reports is to expand the sentences to carry as much information as possible and to give full play to their functions. From the above comparative analysis, we can see that English for Journalism and English for Scientific Literature have both similarities and differences, and each of them has its characteristics, which cannot be replaced by the other.

Jiménez-Crespo proposed a maximum-entropy-based lexical annotator, whose main function is to enrich the information sources used for annotation to achieve excellent performance [5]. The maximum-entropy-based model is a log-linear model, where maximum-likelihood estimation or regularized maximum-likelihood estimation is performed on the model given the training data so that the model can better handle capitalization of unknown words, eliminate verb tense forms, and disambiguate words from prepositions and adverbs. The highest accuracy achieved by the annotator on the Penn Tree bank dataset was 96.86%, and 86.91% for previously unseen words. Cadwell and others proposed Maximum Entropy Markov Models (MEMM), which directly learns conditional probabilities [6]. Agarwal found two problems with the implicit Markov model, transforming it from a generative model to a discriminant model, allow the use of more complex features [7]. Moorkens and others proposed Conditional Random Fields Conditional Random Fields (CRFs) for segmenting and labeling the CRFs and MEMM are discriminant models, which can determine the probability of labeling by the properties of the word itself, and also improve the problem that HMM is a generative model [8]. Since CRFs were proposed, good results have been achieved in the lexical labeling task, and CRFs, HMM, and MEMM are the most commonly used statistical models. Akbari first proposed the BiLSTM+CRFs model for the lexical labeling task and achieved significant results [9]. Tao compared BiLSTM with traditional lexical tagging methods on tasks with different languages and data sizes [10].

Firstly, pre-process the teaching raw data, reasonably process the teaching formulas in sentences, prevent the phenomenon of messy formulae affecting the labeling results, and partition the data into words, and convert them into standard data format. Secondly, we build a neural network framework, train a lexical labeling network model using the existing pure news corpus, label the original teaching data using the model, select the data, not below the threshold value, add them to the training set, test set, and validation set, and remove the news data with the same number of sentences as a unit. Then the training set, test set, and validation set with the added instructional data are used to train a new model and keep repeating the iteration. The iterations are repeated until the trained model decodes the instructional data and the percentage of sentences not below the threshold is small (0.0704 in the text). For sentences with a low probability of being labeled after the completion of the iteration, there are about 2000 words in total, which we can label manually. This paper covers the preparatory knowledge, including the definition of lexical labeling, word vector model, conditional random fields, and deep learning related topics.

2 BUILDING A PARALLEL CORPUS FOR TEACHING ENGLISH TRANSLATION WITH COMPUTER-ASSISTED TRANSLATION SOFTWARE

2.1 English Translation Design for Computer-assisted Translation Software

So far, the existing English lexical annotation corpus is about news materials, but there is no corpus about the scientific literature. In this paper, we propose an approach to lexical annotation of the teaching professional literature, using data from two databases, Springer, and Elsevier, which are part of the full-text teaching literature, about 7000 teaching professional literature. As we choose the teaching professional literature, in the literature, there will be many teaching formulas, and these formulas encoding and ordinary text are not the same, in the txt text will appear when the phenomenon of garbled code [11]. The garbled formulas will affect our classification and prediction of the text, which is detrimental to our study of the text. At the same time, we found that most of the formulas in the teaching text do not affect the structure and meaning of the sentences, so we can delete the formulas. However, some of them have more formulas than others, which would take a lot of time to process manually, and the loss is often not worth the effort. On the other hand, our goal is to study the collocation and association of words and the use of specialized vocabulary in the teaching literature. The abstract is a highly condensed version of the article, which extracts the main ideas and describes the content and specifications of the article in a concise manner. Through manual research on many teaching articles, we found that the abstracts in teaching articles are as rich as the main text in terms of language refinement, formulae, and vocabulary, so we choose the abstracts in teaching literature as the research object in this paper, as shown in Figure 1.

In this paper, we use a news lexical tagging corpus (CCLL2000), which uses the same tags as PennTreebank in its set of lexical tags, with a total of 48 types of tags. However, it should be noted that the tag "PP$" is changed to "PRP$" in the corpus, which means the tags of all the pronouns, such as my, his, and so on. In this paper, the proposed algorithm for constructing a corpus of lexical labels for teaching professionals is based on 48 types of labels, such as CCLL2000. In this paper, the proposed algorithm for constructing the corpus is based on the same kind of 48 classes of labels as CONLL2000. To facilitate our later operation with a neural network, the format of the abstract data is the same as that stored in CCONLL2000, i.e., after dividing a word, for each word, we separate it with space, and add any label from the set of lexical labels, separated by a space, and then save the data into a text document for our later experiments. In this paper, we add the noun label "NN" after each word, and the unit is a sentence, separated by blank lines between sentences.

In the stage, the word vector embedding dimension is set to 50, the number of iterations is set to 200, and the hidden layer dimension is 200. We set the learning rate decay value to 0.05, mainly to prevent the learning rate is too large, in convergence to the optimal value when swinging back and forth, so that the learning rate with the increase in the number of training and continue to decline exponentially, so set the gradient down the learning step to prevent overfitting. In the feature extraction, we will BiLSTM set to three layers, in the later experiments we will prove that the setting of the hyperparameter is reasonable, we set the number of network training iterations is 150 times, the value of the bath_size is 10. The initial word vector of dimension8 lexical list is the word vector obtained from the GloVe model training, in which there are 400,000 words in total, and each word is represented as a continuous vector of 100 dimensions. However, due to the variability of English words, there will always be some unregistered words that are not in the lexicon, so for the initial word vector representation of such words, we choose the same method as the character vector, just initialize it randomly.

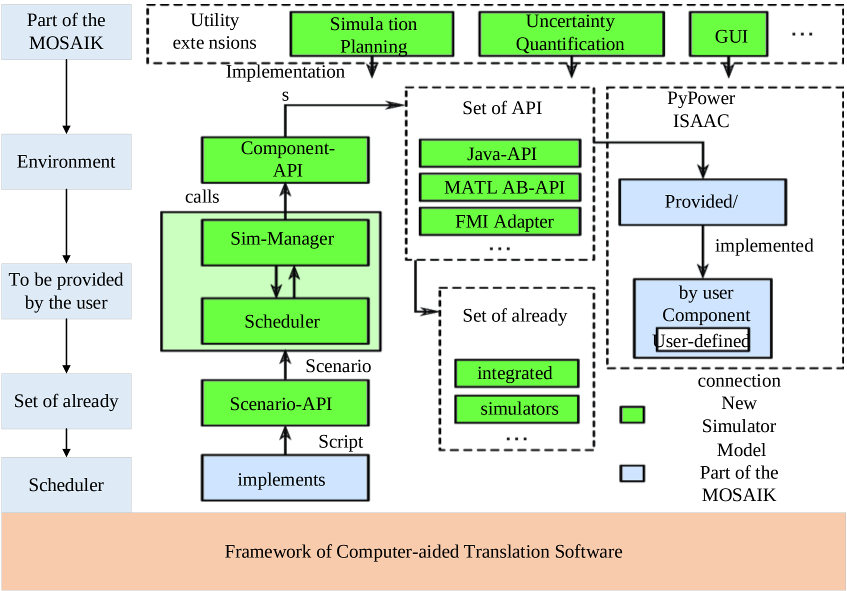

Figure 1: Framework of Computer-aided Translation Software.

2.2 Teaching Parallel Corpus Construction Design

The collection and processing of corpus is a crucial step, and while it is in use it needs to be continuously collected and processed before being added to the corpus. In the subsequent improvement of the corpus, it is possible to add to the corpus written and spoken corpus in speech form, using data collected by the English-Chinese teachers themselves and some British mother tongue corpus translated by the author. The purpose of the corpus is to serve both teachers' teaching and students' learning, and the results of corpus collection and processing will have a direct impact on the quality of teaching.

A quality corpus should not only be a text mover but should also be more prominent in that it allows users with different needs to find the information they need quickly and easily. Furthermore, influenced by the teaching properties of the corpus, different user interfaces, and eye-catching search bars can be set up so that people with different needs can retrieve the target corpus more quickly. The management system was developed by computer engineers with the aim of better serving both teachers and students, so they have designed it with the features of both target audiences in mind as much as possible to prepare scientific management solutions and the management professionals who should be hired. The corpus has two separate interfaces for teachers and students, namely a "student page" and a "teacher page", with different modules added according to their attributes, which is not only simple but also effective. The creation of the corpus is a huge task, and there are inevitably some errors or flaws in the early stages, which must be gradually improved at a later stage. Besides, the corpus is constantly being updated, so the corpus should inevitably be kept up-to-date as well, and improvements and adjustments should be made to the corpus following the ever-changing language reality. Also, the website system may become unstable due to network or technical problems, so it is very important to maintain the website afterward, as shown in Figure 2.

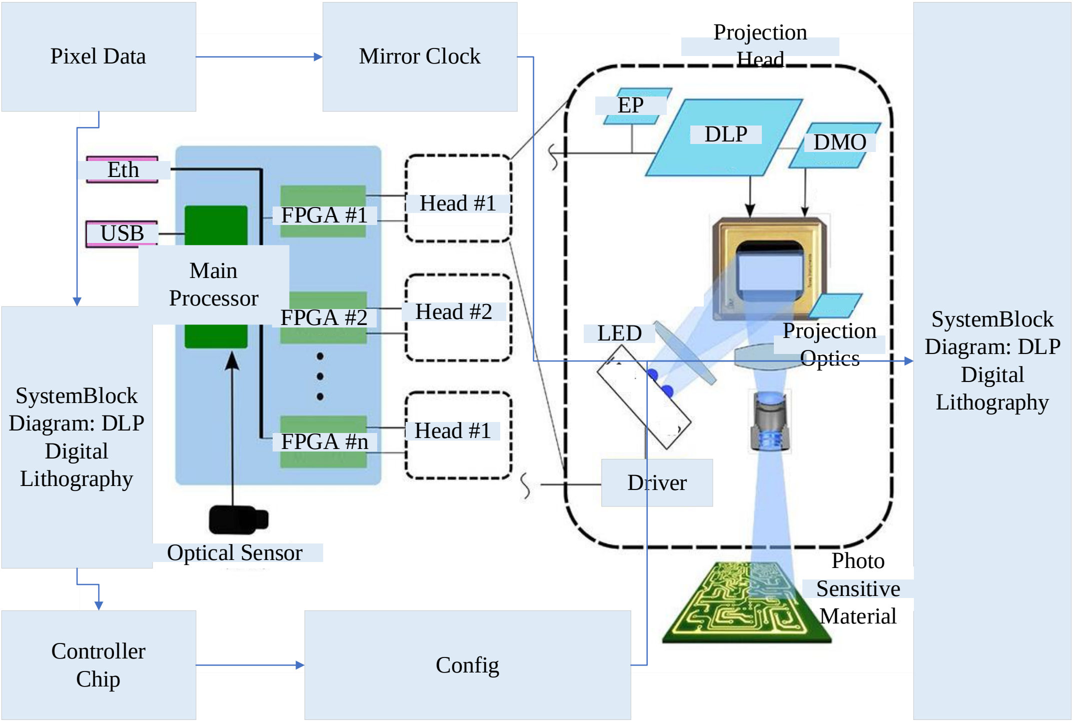

Figure 2: Teaching parallel corpus construction design.

The rule-based translation method is the earliest, and its working principle is to parse the source input text, analyze the grammatical structure of the input text, convert the input statement into an intermediate machine-readable code, and then convert them into the target language; Statistical method-based machine translation is the more popular method, and its working method is to use a very large parallel corpus and monolingual corpus to train the statistical model. It will automatically find the statistical correlation between the input text and the target text, and then find the target text with the highest probability based on the input text. And in recent years, due to the breakthrough progress of deep learning, a series of end-to-end learning methods have been proposed one after another, constantly breaking the record of machine translation-related tasks, and becoming a powerful tool in the task of machine translation.

Grammar is a higher level of abstraction of language, and some grammatical errors are unavoidable in daily life. The traditional Chinese grammar error diagnosis task only deals with four types of grammatical errors: redundant words (R), missing words (M), wrong word selection (S), and disordered words (W). Thus, once surface errors (e.g., spelling errors) are removed, it becomes relatively easy for the model to learn to recognize them. The statistical language model is used to eliminate most of the spelling errors, while the neural network model is used to identify and correct other grammatical errors. The statistical language model is used to eliminate most of the spelling errors, while the neural network model is used to identify and correct other grammatical errors. Finally, the two types of models are combined to generate the final output. The error identification sub-task is performed by using the language model to score the words in the statement, and the low scores are considered as the locations to be corrected. The location to be corrected is then combined with the contextual combinations for a dictionary search, and if all the combinations do not appear in the dictionary, the location is a typo. Contrast research is an extremely important method in textual research. Contrast research relies more on scientific statistics and scientific analysis. Before we develop an artistic appreciation of the text, it is very necessary to present the linguistic features of the translated text employing corpus analysis including frequency, mean, graded, and multiplied values.

To make the reference and comparison texts more reasonable and scientific, this paper combines these works so that the number of characters is closer to the number of characters in English film translations, and also combines three original film collections and divides them into three subbases, which facilitates equal and overall comparisons, tries to avoid coincidence, and increases the persuasive power.

3 ANALYSIS OF RESULTS

3.1 Computer-assisted Translation Software Results Analysis

The spelling error correction algorithm based on the N-Gram language model is the method proposed for solving spelling errors in this thesis and does not deal with deeper grammatical errors, so this section only sets up a set of comparison experiments, i.e., it is compared only with the results of the Computational Linguistics Research Center, the results of which are shown in Figure 3.

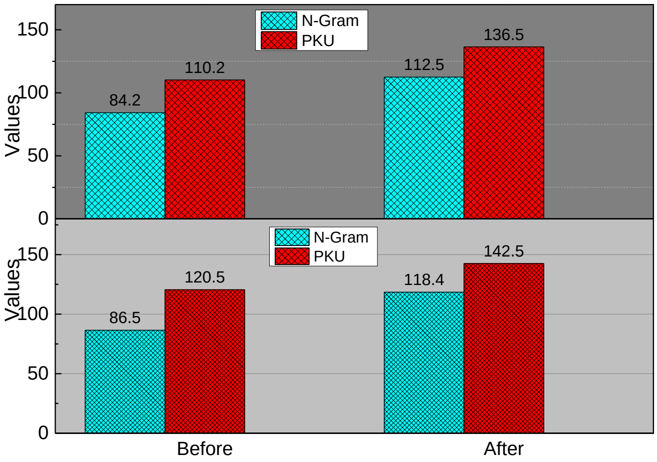

Figure 3: Results of the language model experiments.

By observing Figure 3, it can be seen that the N-Gram language model scores low, and the individual spelling error model works poorly. Also, by observing the output results of the N-Gram model, it is found that the longer the sentences are, the poorer the N-Gram model is, which proves that as the sentence length increases, the model will have cumulative errors. The main introduction is a spelling error correction algorithm based on the N-Gram language model, firstly, the language model is elaborated step by step from a probabilistic point of view, and to solve the problem of too many parameters of the language model, the Markov hypothesis is introduced, which formally leads to the N-Gram language model. Then, to solve the problem of zero probability in the language model, the smoothing technique is introduced, focusing on the additive smoothing method and the Goode-Turing smoothing method. Then, how to identify and correct text spelling errors based on the N-Gram language model is described in detail, focusing on the two subtasks of error identification and error correction, and the training steps of the N-Gram language model, the processing steps of the spelling error correction algorithm of the N-Gram model, and the flowchart are given. Finally, the evaluation index of the language model - confusion level - is introduced, and the experimental design of the N-Gram language model is described and the results are analyzed.

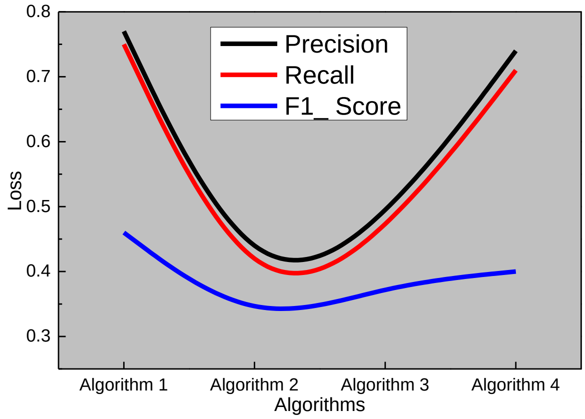

Figure 4: Convergence diagram.

The training process of a deep learning model is essentially a fitting process to the training data, and the lower the data complexity, the better the model is. Conversely, the higher the complexity of the data, the lower the effectiveness of the model. It is now assumed that there are discrete points generated based on functions, and models are built based on these discrete points respectively. Subjectively, the convergence speed and accuracy of the latter model should be better than the former model. Experiments are then conducted, and the results prove that the latter model converges faster and with higher accuracy, and the convergence results are shown in Figure 4. Therefore, when we build a deep learning model, we need to process the data to reduce the data complexity as much as possible and build a suitable model so that the model fits better with the data and speed up the convergence speed and effect of the model. Effective data augmentation can not only increase the number of training samples but also increase the diversity of training samples, which is helpful to avoid overfitting on the one hand and to improve model performance on the other. Data augmentation neither increases the network capacity nor increases the complexity of model calculations and tuning workload, which is of great significance in practical applications. Data augmentation is usually used in computer vision to add data by folding, rotating, or mirroring a picture and not worrying about its markers changing. However, in the field of natural language processing, the situation is completely different, as changing a word or even punctuation can change the meaning of an entire sentence. Therefore, data augmentation is generally implemented in three ways in the domain of natural language processing. Data augmentation is achieved by four simple operations: synonym substitution, random insertion, random exchange, and random deletion, but in the domain of natural language processing, this method is less used because simple add/drop operations are highly likely to damage the model performance, as shown in Figure 5. With the same data, the Transformer model performs better than the traditional Encoder-Decoder model with an F0.5 value of 1.78 percentage points from 20.14 to 21.92, confirming that the Transformer model is better at feature extraction than the traditional cyclic neural network; masked sequence to sequence pre The training strategy can effectively improve the performance of the model by improving the F0.5 value by 2.35 percentage points from 21.92 to 24.27. Also, the performance of the model trained on character-level Chinese Wikipedia corpus was significantly better than the model trained on word-level Wikipedia corpus, which indicates that the pre-training strategy is indeed beneficial to the model performance by indirectly increasing the size of the dataset. This shows that the pre-training strategy does improve the model performance by increasing the size of the dataset indirectly. Data pre-processing is used to edit the distance and data augmentation of the HSK corpus to improve the model performance, followed using sub-word level models to avoid the rare and unknown word problem. The reasons for constructing Transformer models are then explained, and a course learning strategy and a masked sequence-to-sequence pre-training strategy are introduced in the error correction task, and both strategies are described in detail. This is followed using multi-modal integration to improve model performance.

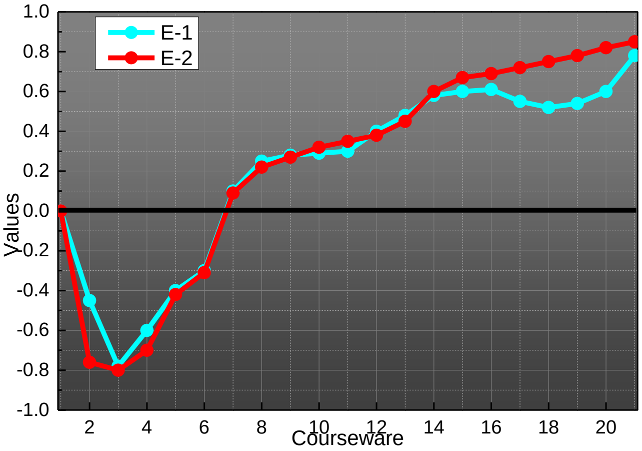

Figure 5: Accuracy results.

3.2 Analysis of Corpus Construction Results

To test the merits of these models, we select a new part of the mathematical data and pre-process it as the test data, called Data 1. To ensure the feasibility of the experiments, we selected a portion of the remaining mathematical data with low probability values as data two. Then we used these models to decode data one and data two, and the final labeling results are shown in Figure 6.

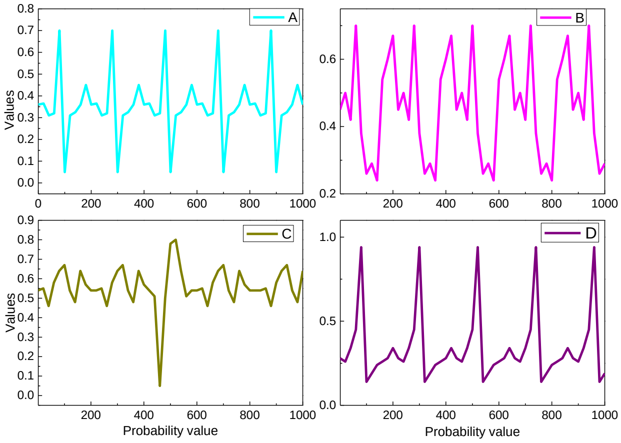

Figure 6: The decoding results of different BiLSTM layer network models for data one.

To analyze the decoding efficiency of the model obtained from different BiLSTM layers, we also used the proportion of sentences with probability values above 0.9 in the total number of sentences as the evaluation criterion. As shown in Figure 7, as the number of BiLSTM layers increases, the proportion of correctly labeled data (sentences with a probability above 0.9) to the total number of sentences increases, reaches the highest in the third layer, and then decreases in the fourth layer.

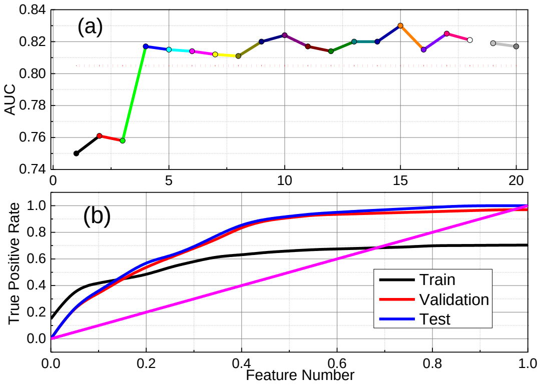

Figure 7: Proportion of correctly labeled sentences for different BiLSTM network models.

Moreover, as the model increases, the number of network parameters increases, and the hardware memory consumption is increasing, the training speed of the network model becomes slower and slower, and the time used is also longer and longer, so we cannot continue to increase the number of BiLSTM layers indefinitely. From the experimental results, we finally choose the three-layer BiLSTM as the network framework for extracting word features.

For this neural network framework, we want to determine the most appropriate optimization algorithm for the network. In this experiment, we control the values of other parameters of the model without changing them and only change the optimization algorithm in the network to determine the appropriate optimization algorithm for the network model. Since the loss function is the mean square error, which is a measure of the accuracy of the model, we change the optimization function to observe the effect of the optimization function on the accuracy of the model. Figure 8(a) shows the accuracy of SGD, AdaGrad, and Adam algorithms for the network model training.

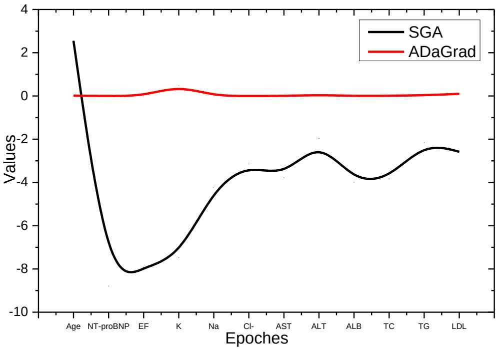

Figure 8: The accuracy of the algorithm.

It can be seen from Figure 8 that when we use the SGD optimization algorithm, the model reaches the maximum accuracy at the 53rd time, which is about 98.36%. When the AdaGrad optimization algorithm is used, the model reaches the maximum accuracy in the 80th time, which is about 98.17%. When the Adam optimization algorithm is used, the model reaches the maximum accuracy at the 26th time, which is about 96.78%. In Figure 9, the accuracy of the model trained with the AdaGrad and Adam optimization algorithms is lower than that of the SGD optimization algorithm, regardless of the number of times the model is trained.

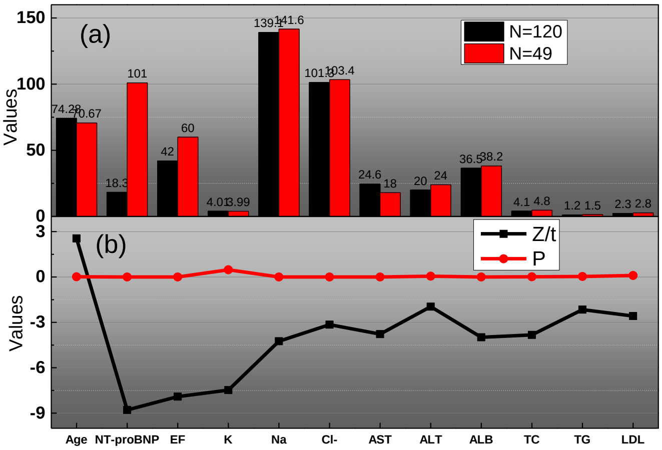

Figure 9: Proportion of correct sentences to total sentences after decoding different models.

In contrast, the Adam optimization algorithm has the fastest convergence rate, but the convergence process is extremely unstable and volatile. On the contrary, AdaGrad and SGD optimization algorithms have similar convergence processes, and both are more stable. To further visualize their convergence, we change the vertical scale and visualize the convergence process in Figure 9. The two networks converge slowly and fluctuate steadily, showing an upward trend, and finally converge to the optimal value. However, the final accuracy of the AdaGrad optimization algorithm was lower than that of the SGD optimization algorithm at 98.17%, which was achieved at the ninth training session with the SGD optimization algorithm. The convergence speed of the AdaGrad optimization algorithm is also much lower than the SGD optimization algorithm, as shown in Figure 9.

As we keep replacing news data with mathematical data to train the model, the number of mathematical data in the training set is increasing, while the number of news data is decreasing, so that the new model learns the distribution of data gradually from the distribution of news data to the distribution of mathematical data, and then slowly replaces them. When we use the optimal model to label pure mathematical data, we achieve the highest proportion of correctly labeled sentences. We propose an algorithm to build a mathematical lexical annotation corpus by combining neural network and conditional random field, comparing the similarities and differences between news data and mathematical data, and conducting a series of experiments.

4 CONCLUSION

With the rapid development of Internet technology, human research on linguistics has been expanding. In particular, in recent years, with the rise of deep learning, the corpus has become the basis for the effective operation of neural network algorithms, and the higher the accuracy of corpus labeling, the larger the scale, the better the performance of neural network models, so the construction and research of corpus has gradually become a hot spot. The construction and development of a corpus play an important role in improving the teaching efficiency and quality of international education, and effective corpus construction can actively promote corpus research and foreign language teaching and research. Effective corpus development can actively promote corpus research and foreign language teaching and research. The trained model is then used to decode the mathematical data, and after several iterations, we finally obtain a model with high efficiency for the lexical annotation of teaching data and a mathematical lexical annotation corpus. In the research process, we propose to combine deep learning and lexical annotation corpus construction, and continuously write algorithms and debug programs to improve our programming ability, as well as a better understanding of the algorithm.

Wenwen Xie, https://orcid.org/0000-0001-7648-9373
Xiaolu Wang, https://orcid.org/0000-0001-8893-1552

REFERENCES

[1] Tian, L.; Zhu, C.: Making Connections Through Knowledge Nodes in Translator Training: On a Computer-Assisted Pedagogical Approach to Literary Translation, International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL), 2(2), 2020, 15-29. http://doi.org/10.4018/IJTIAL.20200701.oa2

[2] Rico, C.: The ePortfolio: constructing learning in translation technology, The Interpreter and Translator Trainer, 11(1), 2017, 79-95. https://doi.org/10.1080/1750399X.2017.1306995

[3] Chen, M.-L.: A corpus-based study on imagery and symbolism in Goldblatt's translation of Red Sorghum. Babel, 65(3), 2019, 399-423. https://doi.org/10.1075/babel.00099.che

[4] Man, D.; Mo, A.; Chau, M.-H.; O'Toole, J.-M.; Lee, C.: Translation technology adoption: evidence from a postgraduate programme for student translators in China, Perspectives, 28(2), 2020, 253-270. https://doi.org/10.1080/0907676X.2019.1677730

[5] Jiménez-Crespo, M.-A.: The role of translation technologies in Spanish language learning, Journal of Spanish Language Teaching, 4(2), 2017,181-193. https://doi.org/10.1080/23247797.2017.1408949

[6] Cadwell, P.; O'Brien, S.; DeLuca, E.: More than tweets: A critical reflection on developing and testing crisis machine translation technology, Translation Spaces, 8(2), 2019, 300-333. https://doi.org/10.1075/ts.19018.cad

[7] Agarwal, C.; Chakraborty, P.: A review of tools and techniques for computer aided pronunciation training (CAPT) in English, Education and Information Technologies, 24(6), 2019, 3731-3743. https://doi.org/10.1007/s10639-019-09955-7

[8] Moorkens, J.: What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise, The Interpreter and Translator Trainer, 12(4), 2018, 375-387. https://doi.org/10.1080/1750399X.2018.1501639

[9] Akbari, A.: Translation quality research: A data-driven collection of peer-reviewed journal articles during 2000–2017, Babel, 64(4), 2018, 548-578. https://doi.org/10.1075/babel.00051.akb

[10] Tao, Y.: Translator training and education in China: Past, present and prospects, The Interpreter and Translator Trainer, 10(2), 2016, 204-223. https://doi.org/10.1080/1750399X.2016.1204873

[11] Bothma, T.-J.; Prinsloo, D.-J.; Heid, U.: A taxonomy of user guidance devices for e-lexicography, Lexicographica, 33(1), 2018, 391-422. https://doi.org/10.1515/lexi-2017-0018