

Received April 30, 2020, accepted May 11, 2020, date of publication May 14, 2020, date of current version June 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994551

Application of Data Mining Methods in Internet of Things Technology for the Translation Systems in Traditional Ethnic Books

YUJING LUO^{1,2} AND YUETING XIANG³

¹Institute of Southwest Minzu Research, Southwest Minzu University, Chengdu 610041, China

²College of Translation and Interpreting, Sichuan International Studies University, Chongqing 400031, China

³School of Foreign Languages and Culture, Xichang University, Xichang 615000, China

Corresponding author: Yujing Luo (sisulyj@foxmail.com)

This work was supported by the Innovative Postgraduate Research Project of Southwest Minzu University through a Research on the Current Situation and Methods of Translating and Introducing Ethnic Classics in Southwest China under the Background of “Going out” Strategy, under Project CX2020BS15.


ABSTRACT In order to translate the ethnic classics, based on the research on the Internet of things, machine learning, and translation technology of ethnic classics, the log-linear model is combined with the national corpus scale and the grammatical structure characteristics, and the phrase statistical machine translation is used to establish a discontinuous phrase extraction model. Then, the translation technology is studied from the three aspects of model definition, training, and decoding. Finally, the algorithm is compared with the traditional phrase extraction algorithm to verify its effectiveness. The results show that the extraction number of discontinuous phrase extraction model is significantly higher than that of traditional phrase extraction model, and the model can extract more phrases, handle larger and more complex text, and score higher in translation fluency. From the evaluation indexes scores of Bilingual Evaluation Understudy (B.L.E.U.) and National Institute of Standards and Technology (N.I.S.T.), it can be found that the B.L.E.U. and N.I.S.T. values of the traditional phrase extraction algorithm are lower than those of the discontinuous phrase extraction model algorithm. The discontinuous phrase extraction algorithm can not only extract the regular continuous phrase, but also obtain the discontinuous text, and the translation effect is better. In conclusion, the combination of Internet of things and machine learning can be used in the translation of ethnic classics to achieve high-quality translation of discontinuous phrases, which is of guiding significance for the study of machine translation.

INDEX TERMS Internet of Things, traditional national classic, Corpora, translation, data mining.

I. INTRODUCTION

With the rapid development of communication and Internet technologies, the translation technique of ethnic classics based on machine learning has become an important means to understand the culture of a nation. Ethnic classics are generated under the specific cultural background of the ethnic groups. They are the reflection of the ethnic culture. With unique cultural forms and characteristics that are different from other cultures, ethnic classics reflect the achievements and contribution of ethnic minorities in a certain field [1]. The traditional ethnic classics are in various forms, which have high cultural and research values. The different culture backgrounds indicate that ethnic-cultural classics have different cultural sources and are corresponding

to different cultural values. The translation of ethnic classics enables people to learn from other ethnic cultures and feel the broadness and uniqueness of culture [2]. However, with the overall development of computer technology, the data volume is getting larger and larger, especially in the field of Internet of things, and the data information redundancy results in a decline in user experience. In such a situation, translation relying only on translators has no longer been able to meet people's demands for high efficiency, low cost, and massive data processing capacity. Therefore, the study of machine translation (M.T.) is of great significance. Machine learning (M.L.) is a branch of computer science that enables computer systems to “learn” with data thereby gradually improving the performance of specific tasks [3]. The translation of classics requires the comprehensive utilization of data information in the ethnic corpus, and the translation can be more effective using M.L.. When evaluating the quality

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen .

of M.T., it mainly refers to the quantitative evaluation of the quality of the translated text generated by the given translation system. Then the problems in M.T. system can be found and improved accordingly. For the translation of ethnic classics, the lack of appropriate training data is the most important factor affecting the development of M.T. of ethnic classics.

Traditional M.T. is mainly based on contiguous phrase translation. Although phrases contain certain contextual information, such incomplete information is far from enough for the translation of the whole sentence. The translation quality of the text is low, and the translation effect is determined by the extraction quality of contiguous phrases. In order to improve this limitation, M.T. based on machine learning is proposed to identify and translate non-contiguous phrases without considering phrase information, which is especially suitable for the translation of minority classics with a small corpus [4]. Compared with the translation of Chinese or English where translation techniques are quite mature, the translation of ethnic language has unique structures and contexts. The non-contiguous phrase extraction uses the search algorithm on the model to find the translation with the highest probability. First, a probabilistic computing model based on corpus and data-driven is constructed, then the parameters of the mathematical model are obtained using corpus training. Finally, based on the model and parameters, the target statement with the highest probability is searched for any source language sentence input [5]. The corpus stores the real text materials in the actual use of language and carries the basic resources of language knowledge. The corpus is constantly supplemented dynamically and needs to be processed before it becomes a resource in the corpus with computer as the carrier. As an important carrier for inheriting and disseminating ethnic cultures, ethnic cultural classics is also the record of the historical and cultural achievements of ethnic minorities, which contain the experience and wisdom of people of the ethnic minority. The M.L. under the Internet of Things (I.o.T.) can create corresponding individualized applications. M.L. has been applied to the I.o.T. platforms. Therefore, the discussion of these technologies is of great significance to the development of M.L. under I.o.T. in different fields in the future.

Based on this, to translate the ethnic classics, based on the research on the Internet of things, machine learning, and translation technology of ethnic classics, the log-linear model is combined with the national corpus scale and the grammatical structure characteristics, and the phrase statistical M.T. is used to establish a discontinuous phrase extraction model. It is compared with the traditional phrase extraction algorithm to verify the effectiveness of the proposed algorithm. This study is about the translation of traditional ethnic classics, which is the collation and summary of ethnic cultures. They have independent cultural forms and are different from other cultures in their particularity, so they are of high cultural and research value. Through the translation of traditional ethnic classics, it is possible to learn from other ethnic cultures

and feel the universality and uniqueness of culture. At the same time, the discontinuous phrase extraction model based on machine learning designed in this study can solve the limitations of the traditional M.T. which is mainly based on the translation of continuous phrases, and has higher applicability, which can provide theoretical value for the translation of discontinuous phrases.

II. LITERATURE REVIEW

M.L. and I.O.T. are now very popular concepts, which are widely used in production and life for constant improvement and update. The words in ethnic classics should have specific meanings or objects, and the I.O.T. technology can be used to connect a meaning or object with other languages. The application technology and architecture construction of language translation have been researched by various scholars. Dahan *et al.* (2019) proposed that the homophones should be semantically processed with semantic network technology; an ontology-based Arabic-English M.T. model Nan was proposed to simulate the translation mode of humans, and the eventual translation results would make the translated texts of T.L. be correct to some extent and be semantically more similar to the artificial translation of non-Arabic natives and non-English natives [6]. Ting *et al.* (2019) proposed that M.T. could use deep learning technology to establish a context prediction model and matrix decomposition model for phrases. The feature of sentence vector was extracted. The quality of M.T. is improved. Its correlation with manual evaluation was estimated. The method extracted the language features of non-contiguous sentences using the context word prediction model based on machine learning technology, and the performance statistics was always better than the traditional translation quality estimation method Quest (Quality estimation shared task) in continuous space language. The former was superior to the latter in both quality and fitness of M.T. [7]. Koponen *et al.* (2018) introduced an M.T.-based language system for editing after content output [1]. Ive *et al.* (2018) proposed that the M.T. could be improved through human-computer interaction; a pre-editing protocol was set up to correct the output [8]. Karimova *et al.* (2018) verified the application of M.T. in the field of patent translation and offline M.T. for different spoken and written languages [9]. Yu *et al.* (2017) combined information technology and MT technology. The Internet-based M.T. system accessed English Chinese bilingual parallel information through the Internet and completed the translation through manual assistance, thereby breaking the dependence on the machine [10]. Sze *et al.* (2017) proposed that in the context of big data of the Internet of things, machine learning technology can quickly extract meaningful information from existing databases, understand and apply such data information to identify application or take immediate action. M.T. is an efficient processing and transformation based on limited data by machine learning technology of the Internet of things [11]. The traditional phrase extraction algorithms need to limit the length of the extracted phrase; otherwise, the number of extracted

phrase pairs will be quite large. Traditional phrase extraction algorithms are often used in phrase-based statistical M.T. (S.M.T.) systems because of their simplicity and accuracy. However, it has the following disadvantages: it is dependent on the quality of word alignment; the strict qualifications will discard information; no non-contiguous phrases are extracted; and when the language differences are significant, the effect is not good. Chinea-rios *et al.* (2018) used the discriminant ridge regression (D.R.R.) method to estimate the log-linear weights of statistical M.T. systems. D.R.R. can provide comparable translation quality and reduce computing costs compared to estimation methods with prior art. In addition, the experimental results were consistent between different corpora and language pairs [12]. Kazemi *et al.* (2017) proposed a reordering model (R.M.) for hierarchical phrase statistical M.T. (hpb-S.M.T.). This model had semantic characteristics, so that the reordering model can be generalized to pairs not found in the training but with the same meaning [13]. Mirjam *et al.* (2017) applied Slavic language to statistical M.T. based on phrases, and found that the interest of the community in studying more difficult languages was increasing, and the translation quality of these languages would reach the level of practical use in the near future [14]. Arcan *et al.* (2018) extracted and integrated the automatically aligned bilingual term into the S.M.T. system, and compared the two term injection methods of X.M.L. markup and caching model. It was found that compared with the widely used S.M.T. system, the model showed a significant improvement from the benchmark S.M.T. system 2.23 to 6.78 B.L.E.U. points, and a significant improvement from 0.05 to 3.03 in the X.M.L. markup method [15]. Passban *et al.* (2017) proposed two different methods to associate complex words with complete sentences in multiple words or even simpler languages in the S.M.T. model. In the first model, the factor S.M.T. engine was enriched by introducing new morphological factors that depend on sub-word perceptive word embedding. In the second model, the focus was on language modeling components to explore the sub-word-level neural language model (N.L.M.) to capture sequences, words, and sub-word-level dependencies. It showed that N.L.M. produced better scores for conditional word probability approximations, so the decoder produced smoother translations [16].

In summary, many researchers have applied M.T. techniques to language translation, and some studies have shown that the translation quality of discontinuous phrases based on M.T. is better than that of continuous phrases. Traditional phrase extraction algorithms need to limit the length of extracted phrases and are often used in phrase based S.M.T. systems. However, it does not extract discontinuous phrases, and when the language is very different, the translation is not good. Some of the existing researches have applied statistical M.T. in phrase translation, but most of them focus on language translation, and there is little research on traditional national classic translation. to translate the ethnic classics, based on the research on the Internet of things,

machine learning, and translation technology of ethnic classics, the log-linear model is combined with the national corpus scale and the grammatical structure characteristics, and the phrase statistical M.T. is used to establish a discontinuous phrase extraction model. It is compared with the traditional phrase extraction algorithm to verify the effectiveness of the proposed algorithm. It is expected that through the translation of traditional ethnic classics, people can learn from other national cultures, and feel the universality and uniqueness of cultures, which provides theoretical value for the translation of discontinuous phrases.

III. PROPOSED METHOD

A. ML UNDER I.O.T

ML is a way for machines to improve the performance of the system through Learning experience by imitating human beings through exposure to new knowledge and calculation. M.L. is a simulation of human behavior. Machines improve their system performance through learning experiences, the process of gaining experience from computing and acquiring new knowledge. M.L. is more about adaptation than learning. M.L. does not describe the behavior of the system by using a physics-based model; instead, it derives a model of the system from the data [17]. M.L. is a branch of artificial intelligence, and its goal is to enable computers to learn on their own. The learning algorithm of the machine enables it to identify patterns in the data, thereby building models that interpret the data and predicting objects without explicit pre-programmed rules and models. The steps are as follows. The computer machine derives the “initial model” algorithm from the massive data and feeds back the empirical data obtained through the learning algorithm to the computer; then, the machine generates a new model based on the new data. Eventually, once new data appears, the computer can help people make corresponding judgments based on the generated model [18]. Machine learning covers a wide range of disciplines. It is the study of how computers simulate human learning behavior to acquire new knowledge or skills and reorganize the existing knowledge structure so that it can continuously improve its own performance. Fig. 1 visually shows the operation process of machine learning for reference in classic book translation. The concept of machine learning

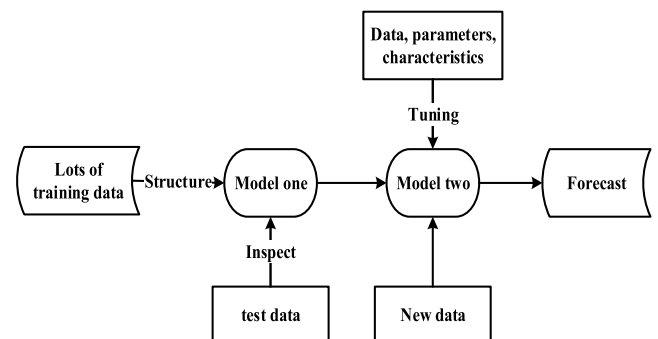


FIGURE 1. The operation mode of ML.

is embodied in the way it works. The first step was based on the initial model of the language training data framework provided to the computer. The test data were used to check the correctness and practicability of the original scheme 1 in step 2. Then, the well-trained scheme was updated and improved, which was denoted as model ii. In other words, it was equivalent to obtaining further optimization performance on the basis of dynamic enrichment of the existing knowledge structure. Eventually, the new data is provided to model two, and people can use the computer and the constructed model two to make new judgments and predictions. The learning algorithm of M.L. can be further improved after getting more data and different data features.

The Internet of things can connect and embed computing devices in everyday objects through the Internet, enabling them to send and receive data. In the search and translation of ethnic classic database, the use of the Internet of things can improve the corpus of ethnic classic works. I.o.T. data describes all aspects of user lives and makes it easier than ever to understand user needs, wishes, history, and preferences. It makes I.o.T. data become the perfect data and can create personalized applications based on the personality of users [19]. The Internet of things is mainly used for the mapping of phrase related corpus in ethnic classics. The words in the ethnic classics all have corresponding meanings or objects. The Internet of things technology was used to associate a meaning or an object with another language, so as to realize the association of national phrases with mandarin or other foreign languages. Moreover, as the Internet of things improves the efficiency of work and life by collecting highly personalized data and providing highly personalized applications and services, machine learning under the background of the Internet of things can effectively solve the translation of ethnic classics in this research. The rational application of the Internet of things and machine learning will improve the translation of ethnic classics.

B. THE ETHNIC CULTURE BASED ON I.O.T

The first step is to establish a strategy of design and development for informational non-material products derived from the ethnic culture of the I.o.T. system. It includes situational animation, electronic fiction, visual communication products displayed through a network platform, and various non-material information products, such as a web platform for cultural communication and communication [20].

The second step is to introduce the construction of a complete industrial chain by the diversified ethnic culture products and the application of the I.o.T. system in the ethnic culture and creative industries, as well as the construction of a strategic emerging cultural industry model characterized by the I.o.T. era [21].

The third step is to realize the construction of a network information platform for the ethnic culture and creative industries. The establishment of the platform will help the ethnic culture resources transfer from creativities to products. Despite the forms of the products, such as material

and non-material, its characteristic as a product is necessary to eventually create the economic and humanistic values based on the needs of people until the product-associated processes are completed, such as promotion, production, and sales. Thus, it provides a relatively complete platform for information exchange and processing, which assists in the industrialized process of ethnic culture and creative products.

The construction of the national culture of the Internet of things is conducive to the collation and statistics of ethnic classics on the platform of the Internet of things, and the in-depth development and utilization of the national cultural corpus [22].

C. TRANSLATION TECHNIQUE OF ETHNIC CLASSICS

Ethnic classics are essential carriers of distinctive cultures. They have recorded the great courses of the ethnic groups, continued the spirit of the ethnic groups, and should always continue to collect and organize, thereby forming documentaries and ancient ethnic books [23].

Classics are the general name of important ancient documents, and ethnic classics are the important carrier of the history and culture of minority nationalities. There are more than 3,000 kinds of classics of ethnic minorities in China, such as the biography of king Gesar, the edition of Buluotuo Jing Shi, and the wisdom of happiness, etc. The ethnic classics of minorities are important carriers for inheriting ethnic cultures. The correct translation of ethnic classics is an important basis for understanding ethnic cultures of minorities and is the most effective way to spread the culture of minorities to the world. In terms of translation, the unique historical features of ethnic minority languages and the features of regional cultural attributes must be strengthened, the cultural values and semantics of words in ethnic classics must be correctly understood, and the effective translation must be carried out. The translation of minority languages in China is divided into two general directions. One is the translation based on phrases, and the other is the translation method of hierarchical sentences and syntactic semantics by means of new methods. Compared with the research results of English Chinese translation, the translation of ethnic classics is still relatively backward. The technology of translation by machine is still primitive, and the sentence analysis and grammatical structure in the translation process are more difficult due to the complexity of ethnic regions. Therefore, the MT of ethnic classics requires constant optimization and efforts.

Because of its simplicity and accuracy, the traditional phrase extraction algorithm is often used in the statistical M.T. system based on contiguous phrases. But it also has the following defects: first, the traditional phrase extraction algorithm relies excessively on the quality of word alignment; second, strict restrictions throw away information; third, non-contiguous phrases cannot be extracted; fourthly, when the language difference is large, the translation effect is not good. In the traditional phrase extraction process, if the target language phrase only contains the word aligned to N.U.L.L. or the source language phrase contains the word aligned

beyond the target language phrase, the phrase pair cannot be extracted.

The S.M.T. considers that the translation of the source sentence to the target sentence is a problem of probability. MT is to find the sentence with the highest probability. The higher the probability is, the more likely it is to be the correct translation. Decoder is an indispensable part of a translation system. Given the source language sentence, the translation system uses the model and search algorithm to obtain the most possible translation, which is completed by the decoder. S.M.T. is the cross-fusion of M.L. and linguistics, i.e. the language translation process is regarded as the process of M.L.. The application direction of M.L. in M.T. is full of various aspects, ranging from parameter adjustment to the model establishment. The log-linear Model is widely used in the field of machine learning and is the most widely used modeling tool. For example, the Naive Bayes model is based on a log-linear model. The mathematical expression of the log-linear model is as follows, where λ_m represents the weight of the feature function, h_m represents the transformed feature function, each h corresponds to a λ , and p represents the probability. Log-linear model can assign different weights to different submodels to improve the translation effect, and can also conveniently add more beneficial features to improve the translation effect. The principle of log-linear model translation is to pre-process the S.L. sentences and solve them in the model. In the process, a large number of special functions h are added, or different weights λ are given; then, the target sentence with the largest $P(x)$ is found, i.e. the most likely translation is searched.

$$P(x) = \exp \left[\sum_{m=1}^M \lambda_m h_m(x) \right] \quad (1)$$

The process of translation is to convert the S.L. into T.L. by means of a model algorithm. Therefore, S.M.T. can be divided into 3 aspects, i.e. model, training, and decoding. The model is to establish a probabilistic model for M.T.. That is to define the method of calculating the translation probability from the source language sentence to the target language sentence. The training is to utilize the corpus to get all the parameters of the model. Based on the given models and parameters, the decoding is to find the translation with the highest probability for any input S.L. sentences.

IV. EXPERIMENTS

A. DEFINITION OF DISCONTINUITY AND MODEL FORMALITY

In order to optimize the limitation of contiguous phrase translation in M.T. and provide better intelligent support for the translation technology of ethnic classics, a non-contiguous phrase translation model based on Internet of things machine learning is proposed. Based on the contiguous phrase M.T., the non-contiguous source phrase translation model improves the phrase extraction algorithm, and the corresponding decoding mechanism is designed.

The following are the main three complete M.L. translation techniques from model defining and model training to corresponding decoding algorithms.

In the traditional extraction algorithm of contiguous phrases, such phrases are more granular than words and can contain richer context information. In the process of translation, the internal structure of the phrase is transparent and there is no problem of word order adjustment in the phrase. To some extent, the ambiguity of translation can be overcome and most grammatical phrases and idioms can be well translated. The traditional contiguous phrase pair satisfies the consistency condition: the words in the source language phrase can only be aligned to the words in the corresponding target phrase, and the source language phrase and the target language phrase must have at least one word aligned. Non-contiguous phrase pairs can be derived from traditional consistency conditions. The length of the source language sentence in the non-contiguous phrase pair is 1 to the boundary division of the source language phrase. The length of the target language sentence is 1 to the boundary of the target language phrase. A non-contiguous phrase consists of a boundary set of source language phrases and a boundary set of target language phrases.

First, the phrase-based translation uses word-groups as the basic translation unit of translation, which may not have linguistic significance. Non-contiguous phrases are not contiguous in word position, with gaps between words. It is supposed that c denotes the target linguistic sentence with the number of words h , d denotes the S.L. sentence with the word length n , and J denotes the set of s consecutive phrase pairs. Then, the rational expression of the contiguous phrase to J_s is as follows:

$$J_s = (h_s, k_s, n_s), \quad 1 \leq s \leq S \quad (2)$$

In (2), h_s represents the position of the last word of the s T.L. phrase; k_s and n_s respectively represent the positions of the first word and the last word of the corresponding SL phrase.

A non-contiguous phrase can be regarded as a set that contains a plurality of consecutive phrases, a definition of a contiguous phrase pair is obtained by the contiguous phrase (2), and the equation of a non-contiguous SL phrase and a contiguous T.L. phrase to J_s is as follows:

$$J_s = (h_s; \tilde{O}_s), \quad 1 \leq s \leq S \quad (3)$$

In (3), O_s represents a set of front and rear boundaries (k, n) of non-contiguous phrases of the S.L. Based on a non-contiguous phrase and a log-linear model, (4) indicates the model definition.

$$\hat{h}_{best} = \arg_{c_1^h, J_1^s} \max \left\{ \sum_{m=1}^M \lambda_m h_m \left(d_1^n, c_1^h, J_1^s \right) \right\} \quad (4)$$

In (4), h_{best} is the best translation or the best target phrase, \arg_{max} is the process of searching for the best translation, S.L. sentence d , candidate target sentence c , and

non-contiguous phrase pair j are three parameters of feature function h . A non-contiguous phrase number penalty function, such as (5), is used to calculate the number of non-contiguous phrases in the S.L.

$$f_e \left(d_1^n, c_1^h, J_1^s \right) = \sum_{s=1}^S (Z_s > 1) \quad (5)$$

In (5), f_e represents the number of non-contiguous phrases, Z_s represents the number of all non-contiguous sub-phrases; $Z_s > 1$ indicates a false value of 0, and true value is 1. The non-contiguous phrase discontinuity length penalty function is shown in (6) to calculate the total discontinuity length in the source language.

$$f_g \left(d_1^n, c_1^h, J_1^s \right) = \sum_{s=1}^S \sum_{z=1}^{Z_s-1} (k_{s,z+1} - n_{s,z} - 1) \quad (6)$$

In (6), f_g represents the non-contiguous length of the non-contiguous phrase, $(k_{s,z+1} - n_{s,z} - 1)$ indicates the length of the non-contiguity between the two preceding and following sub-phrases in the same non-contiguous phrase.

B. THE EXTRACTION OF DISCONTINUOUS PHRASES

In the case of non-contiguous phrase extraction, the boundary partitioning set of the S.L. phrase is \tilde{O}_d , and the boundary partitioning set of the T.L. phrase is \tilde{O}_c ; then, for a non-contiguous phrase pair $(\tilde{O}_d, \tilde{O}_c)$, $1 \leq \tilde{O}_d \leq n$; $1 \leq \tilde{O}_c \leq h$.

$$\forall (h, n) \in A : h \in \tilde{O}_c \leftrightarrow n \in \tilde{O}_d \quad (7)$$

$$\exists (h, n) \in A : h \in \tilde{O}_c \wedge n \in \tilde{O}_d \quad (8)$$

Equation (7) represents that for a word of position n in any S.L. phrase aligned to a T.L. word with a word position of h ; Equation (8) represents that there is a word alignment that satisfies $h \in \tilde{O}_c, n \in \tilde{O}_d$.

The word alignment-based extraction algorithm first extracts contiguous phrases; then, it verifies the consistency, expands the phrase pairs by breaking into a loop, and finally extracts the matching non-contiguous phrase pairs.

The non-contiguous phrase decoding algorithm adds a loop that extends the existing translation hypothesis as a translation option if the non-contiguous phrase reaches the sum of the source phrase lengths and the phrase header position. Finally, the extended paths that generate the same translation hypothesis are merged together; whichever path has the highest probability. Since the extended path covers the same S.L. and the same T.L. is generated, the path with a high probability of high score is selected as the final extended path during the reorganization. The improved maximum probability translation of the non-contiguous phrase decoding algorithm is as shown in (9).

$$DBP_n \left(f_1^i, e_1^i, A \right) = \left\{ \left\{ f_{j_1,1}^{j_1,2} \# \dots \# f_{j_{n+1},1}^{j_{n+1},2}, e_{i_1}^{i_2} \right\} \mid (1 \leq i_1 \leq i_2 \leq I) \wedge \right. \quad (9)$$

Equation (9) represents all the phrase pairs under the generalized definition extracted from a sentence pair. First, the

word alignment-based extraction algorithm extracts the contiguous phrases; then, it verifies the consistency and expands the phrase pairs by breaking into a loop; finally, it extracts the matching non-contiguous phrase pairs. The discontinuous phrase extraction process and algorithm implementation are shown in Fig. 2 and Table 1 respectively.

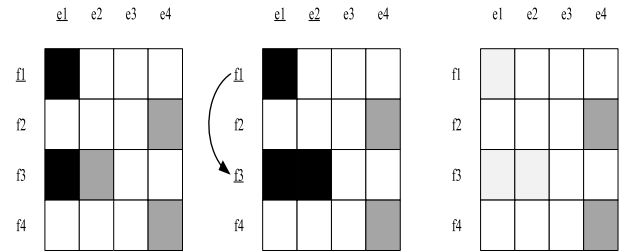


FIGURE 2. The extraction process of non-contiguous phrases.

C. DECODING OF DISCONTINUOUS PHRASE ALGORITHM

The non-contiguous phrase decoding algorithm adds a loop that extends the existing translation hypothesis as a translation option if the non-contiguous phrase meets the sum of the source phrase length and the phrase header position. Finally, the extended paths that generate the same translation hypothesis are merged together; whichever path has the highest probability. Since the extended path covers the same S.L. and the same T.L. is generated, the path with a high probability of high score is selected as the final extended path during reorganization. The improved maximum probability translation of the non-contiguous phrase decoding algorithm is as shown in (10).

$$\hat{S} = \max \left(S \left(\{1 \dots N\}, e, \tilde{N} \right) + S_{LM} (\#/\tilde{e}) + S_{DM} (n_1, n_2) \right) \quad (10)$$

In (10), S represents the probability that the translation hypothesis (c, e, n) extends from the null hypothesis to the best path of another hypothesis, c denotes the set of covered phrases, e denotes the language model value, and n denotes the covered tail position.

D. EVALUATION INDEXES

B.L.E.U. (Bilingual Evaluation Understudy) is one of the most widely used M.T. Evaluation indexes. Its basic principle is to measure the similarity of accuracy of n -element grammar matching between the reference text and M.T.. B.L.E.U. penalizes M.T. statements that are shorter than the text. B.L.E.U. addresses word order problems by allowing B.L.E.U. to use multiple references. B.L.E.U. scoring interval $(0,1)$, the more n -element grammars there are, the higher the score will be, that is, the better the translation quality will be [24].

The N.I.S.T. index evaluation is based on the B.L.E.U. evaluation standard, which is an improvement on B.L.E.U. The N.I.S.T. score is a real number greater than 0, and a

TABLE 1. Discontinuous phrase extraction algorithm.

Input: source language sentence f_1^J , target language sentence e_1^I . The words are aligned with A, and the upper limit of the number of intervals is N
Output: a set of discontinuous phrase pairs DBP
1. $S_0 = \{ \}$ // the initial phrase pair set is empty
2. FOR $j_1 = 1$ TO J DO // 2-9 lines, extracting the continuous phrase and storing it in the first set, S_0
3. IF j_1 alignment is empty THEN CONTINUE // there is no alignment for the current word, skip
4. $i_1 = \text{length}(e_1^I)$; $i_2 = 0$
5. FOR $j_2 = j_1$ TO J DO // extending a word backwards, looking for successive pairs of phrases
6. IF j_2 alignment is empty THEN CONTINUE // skip
7. $i_1 = \text{MIN}\{i_1, \text{MIN}\{i (i, j_2) \in A\}\}$ // the first word of the target phrase
8. $i_2 = \text{MAX}\{i_2, \text{MAX}\{i (i, j_2) \in A\}\}$ // the word at the end of the target phrase
9. $S_0 = S_0 \cup \left\{ \left\langle d_{j_1}^{j_1}, c_{i_1}^{i_2} \right\rangle \right\}$ // extracting pairs of phrases that conform to the traditional consistency
10. FOR $n = 1$ TO N DO // 10-21 lines, extracting a discontinuous phrase
11. $S_n = \{ \}$
12. FOR $\langle f_{j_{n,1}}^{j_{n,2}} \# \dots \# f_{j_{n,1}}^{j_{n,2}}, e_{i_1}^{i_2} \rangle$ IN S_{n-1} DO // iterating over the phrases in the previous set
13. FOR $j_1 = j_{n,2} + 2$ TO J DO // spacing a word and then expanding the phrase
14. IF $\exists (i, j) \in A : j_{n,2} < j < j_1$ THEN CONTINUE
15. IF j_1 alignment is empty THEN CONTINUE // there is no alignment for the current word, skip
16. $i_1' = i_1$; $i_2' = i_2$
17. FOR $j_2 = j_1$ TO J DO // extending a word backwards, looking for continuous subphrase alignment
18. IF j_2 alignment is empty THEN CONTINUE // skip
19. $i_1' = \text{MIN}\{i_1', \text{MIN}\{i (i, j_2) \in A\}\}$
20. $i_2' = \text{MAX}\{i_2', \text{MAX}\{i (i, j_2) \in A\}\}$
21. $S_n = S_n \cup \left\{ \left\langle f_{j_{n,1}}^{j_{n,2}} \# \dots \# f_{j_{n,1}}^{j_{n,2}} \# f_{j_1}^{j_2}, e_{i_1'}^{i_2'} \right\rangle \right\}$ // new discontinuous phrases
22. FOR $n = 0$ TO N DO // picking the right phrase
23. FOR phase pair dbp IN S_n DO
24. IF dbp conforms consistency THEN
$DBP_n = DBP_n \cup \{dbp\}$
25. $DBP = DBP \cup DBP_n$

high score means the translation is of good quality. N.I.S.T. index evaluation is an improvement on B.L.E.U. index evaluation standard, and the translation quality determines the rating level. Although N.I.S.T. is indeed based on B.L.E.U., B.L.E.U. is still more widely used than N.I.S.T. [21].

Confusion index is the most commonly used measure to evaluate a M.T. [25]. It is the probability of test data calculated according to the language model. Confusion, or the probability of a sentence being graded, is determined by the first word of the sentence, the length of the sentence, or the number of words. In M.T., confusion is usually used as a measure of sentence fluency. The smaller the confusion, the better the order of words in the translated text is in line with human language. The magnitude of the confusion is related to the type of language and text. The confusion of the language model calculation is about 50 to 1000. The lower the value is, the smoother the text is. However, most automated evaluation methods usually score between 0 and 1, and a higher value is considered better. In order to keep up with the habits of most automated evaluation methods, the reciprocal of the translational confusion is defined as the fluency score of the translation, which is derived to obtain (11).

$$\text{score}_{\text{fluency}} = \frac{1}{P(s)} = \left(\prod_{i=1}^n p(w_i | w_{i-2} w_{i-1}) \right)^{1/n} \quad (11)$$

V. RESULTS

Under this model, the M.L. ethnic classics translation based on I.o.T. can deal with the problem of the small-scale corpus. The reason is that based on the extensive data in the Internet of Things, the translation of traditional ethnic classics can be solved by using machine learning extraction model algorithm of non-contiguous phrase, which can thereby solve the translation difficulties brought by small-scale corpus. The extraction and decoding based on non-contiguous phrases can affect the quality of translation. In the previous section, the problem has been mathematically modeled and the algorithm is designed for solutions. In this experiment, the effectiveness of the performance of the algorithm needs to be verified.

The experimental data of this study include the dictionaries of minority language and the parallel corpus of minority language and mandarin, in which the minority languages refers to Uighur language. Also, the Mongolian to mandarin, Tibetan to mandarin, and Uighur to mandarin are the three major evaluation projects of China Workshop on M.T. (C.W.M.T.). The Uighur dictionary is from the Internet. The Uighur parallel corpus is the corpus data provided by Xinjiang University, which contains more than 100,000 pieces of Uighur sentences. The experiment classifies the words, and the number of categories is 40. The words with the highest number of occurrences in the 40 categories are extracted from the corpus, which constitutes the corpus dictionary. In addition, 500 sentences are extracted from the

TABLE 2. Information on test data.

	Training Set	Test Set	New Set
<i>Gesar</i> , a Tibetan classic	10000	5000	3000
Chinese Corpora (Sentences)	10000	5000	3000
Tibetan-Chinese Phrases (sentences)	3000	2000	1000

During the experiment, the algorithms are compared through M.T. simulation.

corpus as the development set, and the test set is independent of the corpus, which contains the unregistered words.

Thus, the Tibetan long heroic epic *Gesar* is taken as the object to randomly select sentences for translation. The comparison is performed by a conventional algorithm, i.e. the phrase extraction, and the improved algorithm, i.e. a non-contiguous phrase algorithm for M.L. in the context of the I.o.T..

As shown in Table 2, the ‘‘Training Set’’ is used for model training of the system, the ‘‘New Set’’ is used for tuning training of the system, and the ‘‘Test Set’’ is used for evaluating the translation quality of the system.

As shown in Fig. 3, the number of extractions of traditional algorithms is significantly lower than that of the extraction model algorithm of non-contiguous phrases designed in this study. The extraction algorithm based on non-contiguous phrases can extract more phrases. The main reason is that the new algorithm can broaden the extraction conditions of traditional algorithms. The phrase that does not conform to the traditional algorithm extraction condition can be recognized by the new non-contiguous phrase extraction algorithm. Machine learning technology in the context of the I.o.T. uses big data analysis to achieve better translation quality.

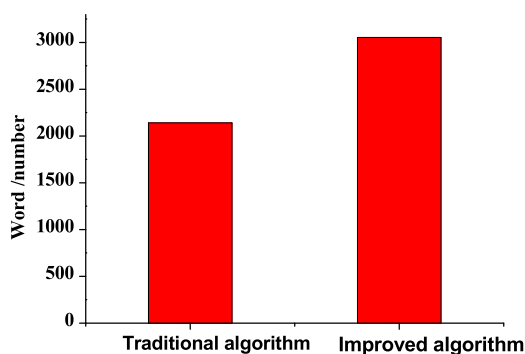


FIGURE 3. The comparison of extracted phrases.

As shown in Fig. 4, the fluency score of the proposed algorithm is always better than that of the traditional algorithm. The main reason is that traditional algorithms do not rely on the Internet of things technology and can only extract contiguous words, thus, the fluency score decreases as the number of text increases. The machine learning translation model algorithm in the context of the Internet of things is able to handle larger and more complex text volumes, and has a higher translation fluency score.

B.L.E.U. refers to Bilingual Evaluation Understudy. The B.L.E.U. indicator evaluates the influence of word order and considers the number of n-variable grammars that match the length of the translation produced by M.T. with the reference translation. The higher the number is, the higher the score is. As shown in Fig. 5, compared with the scores that the extraction model algorithm of non-contiguous phrases gets, the B.L.E.U. indicator score of the traditional algorithm is lower. The non-contiguous phrase extraction algorithm proposed in this study can obtain non-contiguous text options in addition to the regular contiguous phrases, which has a good translation effect on the tested sentences. The B.L.E.U. indicator evaluates the influence of word order and considers the number of n-variable grammars that match the length of the translation produced by M.T. with the reference translation. The higher the number is, the higher the score is.

N.I.S.T. refers to National Institute of Standards and Technology. The N.I.S.T. indicator evaluation is an improvement of the B.L.E.U. indicator evaluation standard, in which the quality of the translation determines the level of the score. As shown in Fig. 6, compared with the traditional algorithm, the N.I.S.T. indicator score of the extraction model algorithm of non-contiguous phrases designed in this study is increased. The M.T. model algorithm of machine learning under the background of I.o.T. can obtain non-contiguous phrases and has better translation quality. The N.I.S.T. indicator evaluation is an improvement of the B.L.E.U. indicator evaluation standard, and the quality of the translation determines the level of the score. It can be seen that the extraction model algorithm of non-contiguous phrases designed in this study can obtain more effective translation results.

VI. DISCUSSION

In order to translate the ethnic classics, based on the research on the Internet of things, machine learning, and translation technology of ethnic classics, the log-linear model is combined with the national corpus scale and the grammatical structure characteristics, and the phrase statistical M.T. is used to establish a discontinuous phrase extraction model. The algorithm is compared with the traditional phrase extraction algorithm to verify the effectiveness. The results show that the number of discontinuous phrase extraction model is

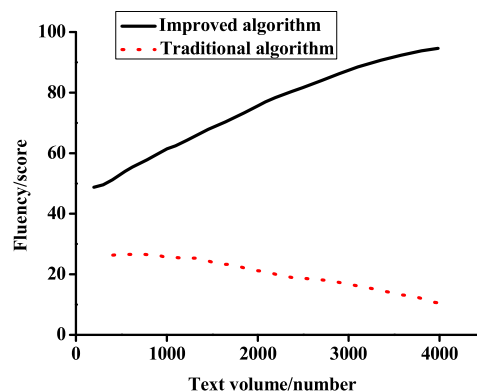


FIGURE 4. The comparison of fluency scores.

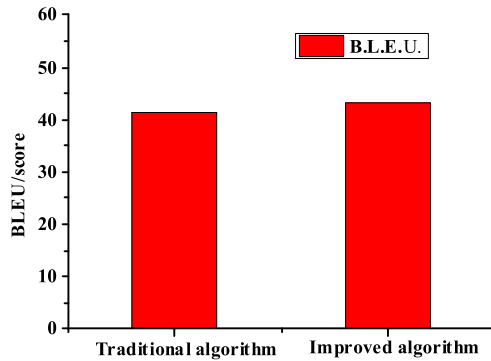


FIGURE 5. The comparison of B.L.E.U. indicator scores.

significantly higher than that of traditional phrase extraction algorithm, and the model can extract more phrases, handle larger and more complex text, and translate more fluently. The B.L.E.U. and N.I.S.T. values of the traditional phrase extraction algorithm are lower than those of the discontinuous phrase extraction model algorithm. The algorithm of discontinuous phrase extraction can not only extract the regular continuous phrase, but also obtain the discontinuous text. The translation effect is better, and the results are consistent with the experimental expectation. Hong *et al.* (2018) converted the short S.I.M.D. command into a discontinuous phrase and translated it, which greatly improved its speed [26]. Miura *et al.* (2016) proposed a method to remember key discontinuous phrases in triangulation stage, and used key language model as additional information source in the transformation stage. Experimental results showed that all tested language combinations have achieved significant improvements [27]. Saeed *et al.* (2018) proposed a phrase dependent tree source reordering method. This method described the dependencies between successive non-grammatical phrases. It can automatically learn to reorder elements from the reordered phrase dependency tree library and use it to generate the source reorder lattice. Finally, S.M.T. based on monotone phrases is used to decode the lattice and translate the source sentences, and the results showed that the translation quality of this method is better [28]. There-

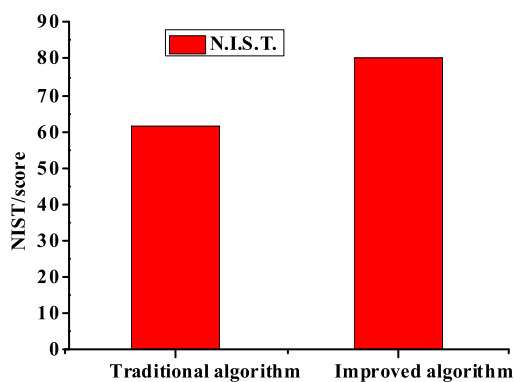


FIGURE 6. The comparison of N.I.S.T. indicator scores.

fore, translation based on discontinuous phrases can achieve better translation quality, which is the same as the goal of this study.

VII. CONCLUSION

In this study, the translation technique of ethnic classics is introduced, the M.L. technology is utilized for comprehensive analysis, and the translation algorithm based on M.L. in the context of I.o.T. is put forward. In addition, the traditional contiguous phrase extraction is optimized, and the extraction model and algorithm of non-contiguous phrases are designed [29]. A complete S.M.T. method is analyzed from the aspects of model definition, model training, and decoding algorithm. In view of the specific language scenarios where the ethnic corpus is insufficient, the phrase extraction method that ignores the non-contiguity of position in traditional phrase extraction is improved, the strategy of extracting and decoding non-contiguous phrases is realized, the small-scale corpus is fully utilized [30]; then, the experimental comparison between the traditional algorithm and the extraction model algorithm of non-contiguous phrases is carried out. The experimental results have shown that the comparison between the proposed algorithm and the traditional algorithm is improved, which means that the proposed model algorithm for extracting non-contiguous phrases is more effective than the traditional model for extracting contiguous phrases in M.T. of ethnic classics. Thus, the technical algorithm can improve the number of phrases extracted, as well as the fluency, quality, and efficiency of translation, which realizes the performance optimization of ethnic classic translations [31].

The size of the ethnic corpus is relatively small, and the grammatical structure and semantics of ethnic language have certain regional, historical, and special features. In terms of the translation of ethnic classics, it is necessary to further develop statistical translation models. The extraction model algorithm of non-contiguous phrases designed in this study improves the effect of ethnic classic translation; however, for longer sentences, limitations can still be found in word order adjustment, i.e., the longer the sentence is, the lower the accuracy of translation is, and the better the effect of short sentence translation is. Thus, the algorithm should be improved in the future, and subsequent optimization should also be carried out in terms of strengthening the deep learning.

REFERENCES

- [1] M. Koponen, L. Salmi, and M. Nikulin, "A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output," *Mach. Transl.*, vol. 33, nos. 1–2, pp. 61–90, Jun. 2019.
- [2] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 371–383, Dec. 2016.
- [3] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, Jan. 2018.
- [4] H.-B. Chen, H.-H. Huang, A.-C. Hsieh, and H.-H. Chen, "A simplification–translation–restoration framework for domain adaptation in statistical machine translation: A case study in medical record translation," *Comput. Speech Lang.*, vol. 42, pp. 59–80, Mar. 2017.

- [5] K. Nguyen, H. Daumé, III, and J. Boyd-Graber, "Reinforcement learning for bandit neural machine translation with simulated human feedback," 2017, *arXiv:1707.07402*. [Online]. Available: <http://arxiv.org/abs/1707.07402>
- [6] N. A. Dahan and F. M. Ba-Alwi, "Extending a model for ontology-based Arabic-English machine translation (NAN)," *Int. J. Artif. Intell. Appl.*, vol. 10, no. 1, pp. 55–67, Jan. 2019.
- [7] L. Tingting and X. Mengyu, "Analysis and evaluation on the quality of news text machine translation based on neural network," *Multimedia Tools Appl.*, pp. 1–12, Apr. 2019.
- [8] J. Ive, A. Max, and F. Yvon, "Reassessing the proper place of man and machine in translation: A pre-translation scenario," *Mach. Transl.*, vol. 32, no. 4, pp. 279–308, Dec. 2018.
- [9] S. Karimova, P. Simianer, and S. Riezler, "A user-study on online adaptation of neural machine translation to human post-edits," *Mach. Transl.*, vol. 32, no. 4, pp. 309–324, Dec. 2018.
- [10] Y. Zhang, "Research on English machine translation system based on the Internet," *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 1017–1022, Dec. 2017.
- [11] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, Dec. 2017.
- [12] M. Chinea-Rios, G. Sanchis-Trilles, and F. Casacuberta, "Discriminative ridge regression algorithm for adaptation in statistical machine translation," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1293–1305, Nov. 2019.
- [13] A. Kazemi, A. Toral, A. Way, A. Monadjemi, and M. Nematbakhsh, "Syntax- and semantic-based reordering in hierarchical phrase-based statistical machine translation," *Expert Syst. Appl.*, vol. 84, pp. 186–199, Oct. 2017.
- [14] M. S. Maučec and J. Brest, "Slavic languages in phrase-based statistical machine translation: A survey," *Artif. Intell. Rev.*, vol. 51, no. 1, pp. 77–117, Jan. 2019.
- [15] M. Arcan, M. Turchi, S. Tonelli, and P. Buitelaar, "Leveraging bilingual terminology to improve machine translation in a CAT environment," *Natural Lang. Eng.*, vol. 23, no. 5, pp. 763–788, Sep. 2017.
- [16] P. Passban, Q. Liu, and A. Way, "Providing morphological information for SMT using neural networks," *Prague Bull. Math. Linguistics*, vol. 108, no. 1, pp. 271–282, Jun. 2017.
- [17] A. Laskovaia, G. Shirokova, and M. H. Morris, "National culture, effectuation, and new venture performance: Global evidence from student entrepreneurs," *Small Bus. Econ.*, vol. 49, no. 3, pp. 687–709, Oct. 2017.
- [18] L. Li, C. P. Escartín, A. Way, and Q. Liu, "Combining translation memories and statistical machine translation using sparse features," *Mach. Transl.*, vol. 30, nos. 3–4, pp. 183–202, Dec. 2016.
- [19] A. Toral and V. M. Sánchez-Cartagena, "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions," 2017, *arXiv:1701.02901*. [Online]. Available: <http://arxiv.org/abs/1701.02901>
- [20] U. Hermjakob, Q. Li, D. Marcu, J. May, S. J. Mielke, N. Pourdamghani, M. Pust, X. Shi, K. Knight, T. Levinboim, K. Murray, D. Chiang, B. Zhang, X. Pan, D. Lu, Y. Lin, and H. Ji, "Incident-driven machine translation and name tagging for low-resource languages," *Mach. Transl.*, vol. 32, nos. 1–2, pp. 59–89, Jun. 2018.
- [21] Z. B. Jiang, "Professional ethics in Chinese traditional folk culture," *J. Ethnic Art.*, vol. 2, no. 4, pp. 43–48, 2019.
- [22] B. B. Wang and Z. Qun, "Study on discipline system of talent training in marine cultural industry based on collaborative innovation of governments, enterprises, colleges, scientific institutions and users," *Climatic Environ. Res.*, vol. 5, pp. 33–36, Oct. 2018.
- [23] B. Steeve and J. E. Skinner, "The invention of Greek ethnography: From Homer to Herodotus," *Amer. Historical Rev.*, vol. 119, no. 5, p. 1755, 2017.
- [24] E. Reiter, "A structured review of the validity of BLEU," *Comput. Linguistics*, vol. 44, no. 3, pp. 393–401, Sep. 2018.
- [25] J. Bajpai, P. Panda, S. Kagwade, M. Govilkar, S. Velaskar, Y. Kumbhavi, S. Gupta, J. Ghosh, and J. Deodhar, "Translation and validation of European organization for research and treatment for cancer quality of life questionnaire-OV-28 module into Indian languages (Hindi and Marathi) to study quality of life of ovarian cancer patients from a tertiary care cancer center," *South Asian J. Cancer*, vol. 7, no. 1, pp. 37–41, 2018.
- [26] D.-Y. Hong, Y.-P. Liu, S.-Y. Fu, J.-J. Wu, and W.-C. Hsu, "Improving SIMD parallelism via dynamic binary translation," *ACM Trans. Embedded Comput. Syst.*, vol. 17, no. 3, pp. 1–27, Jun. 2018.
- [27] A. Miura, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Improving pivot translation by remembering the pivot," *J. Natural Lang. Process.*, vol. 23, no. 5, pp. 499–528, 2016.
- [28] S. Farzi, H. Faili, and S. Kianian, "A preordering model based on phrasal dependency tree," *Digit. Scholarship Humanities*, vol. 33, no. 4, pp. 748–765, Dec. 2018.
- [29] Y. Su, L. Han, J. Wang, and H. Wang, "Quantum-behaved RS-PSO-LSSVM method for quality prediction in parts production processes," *Concurrency Comput.-Pract. Exper.*, vol. 9, p. e5522, Sep. 2019.
- [30] Z. Liu and C. Wang, "Design of traffic emergency response system based on Internet of Things and data mining in emergencies," *IEEE Access*, vol. 7, pp. 113950–113962, 2019.
- [31] C.-W. Shen, M. Chen, and C.-C. Wang, "Analyzing the trend of O2O commerce by bilingual text mining on social media," *Comput. Hum. Behav.*, vol. 101, pp. 474–483, Dec. 2019, doi: 10.1016/j.chb.2018.09.031.

• • •