Recognition and Segmentation of English Long and Short Sentences Based on Machine Translation

https://doi.org/10.3991/ijet.v15i101.10182

Tiehu Zhang
Xi'an Aeronautical University, Shaanxi, China
tiehu_zhang@126.com

Abstract—With the advent of the information age, long sentences which include many words and have more complex structures. The translation of long sentences in English-Chinese machine translation has always been the focus of research. In this study, 400 long sentences were randomly selected from NTCIR-9 patent corpus for testing the recognition and segmentation effects of regular match method and error-driven method, and the accuracy rate of the translation was compared on Baidu Online Translation Platform. The results demonstrated that the regular matching method was effective in recognizing and segmenting long sentences, nevertheless there were many defects; the error-driven method was more effective in recognizing and segmenting long sentences; the former increased by 4.8% of the BLEU value of the translated text on Baidu Online Translation Platform and the latter increased by 12.1%, which showed that the error-driven method was more effective in machine translation.

Keywords—Machine translation, long sentence, regular match, error-driven method.

1 Introduction

With the advent of the information age exchanges between different countries have become increasingly close, including a large number of academic exchanges [1]. However, due to the differences in languages between different countries, language translation is needed in communication. As academic exchanges usually contain a large number of long sentences which have complex structures, requirements on language translation has dramatically increased. The workload of manual translation is high, and it is difficult to meet the convenience of communication as it consumes too much time and energy. With the emergence and popularization of computers, high-efficient machine translation came into being. The earliest machine translation system was indeed far faster than manual translation. However, it was difficult to guarantee the fluency and rationality of the "one-to-one correspondence" translation, especially the translation of long sentences. Machine translation quality has progressed with time, although it is still not satisfactory in the translation of long sentences [2]. To enhance the translation quality of long sentences, many scholars have made relevant studies. Sennrich et al. [3] realized the open vocabulary translation of Neuro-Machine

Translation Model (NMT) by coding rare and unknown words into subunit sequences and found that the model could effectively improve the regression baseline of dictionaries in English-German and English-Russian translation tasks. To solve the over translation problem of NMT model, Tu et al. [4] put forward coverage based NMT model and maintained a coverage vector to keep track of the attention history and found that the method had higher translation quality than the attention based NMT model. In a study of Shen et al. [5], minimum risk training was put forward for end-to-end neural machine translation. Different from maximum likelihood estimate, the method directly optimized model parameters according to evaluation indicators. The experimental results suggested that the method was more effective than neural machine translation system which applies maximum likelihood estimate. In the present study, 400 random long sentences were selected from NTCIR-9 patent corpse and used for testing the long sentence recognition and segmentation effects of regular match method and error-driven method, and the accuracy of the translated text was compared on Baidu online translation platform.

2 Machine Translation

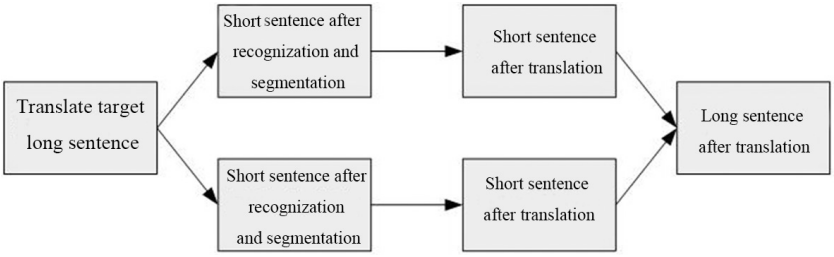

Fig. 1. The machine translation process of English sentences

At present, the mainstream English machine translation systems in the market follow the process [6] shown in Fig. 1 when translating long sentences. Firstly, long English sentences are identified and divided into independent short sentences according to the pre-set method, thereafter the short sentences are translated according to the lexicon. After the translation, the system combines the short sentences into long sentences according to the semantic rules of the target language and checks the translated long sentences repeatedly according to the grammatical rules before the final output.

Lexicon and language model are also important parts of the machine translation system. Lexicon is equivalent to the bilingual dictionary usually used in manual translation, while language model is used to estimate the structure of natural language to make the long sentences more natural after translation.

3 Regular Match Method

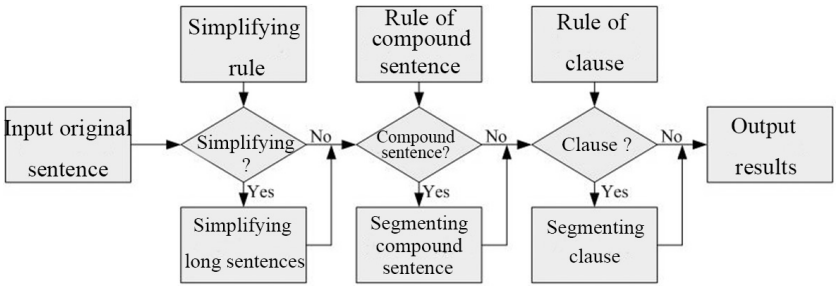

Fig. 2. The segmentation process of long sentences using regular match method

In the machine translation of English long sentences, the sentences need to be segmented into short sentences for subsequent translation. Regular match method is used for segmenting long sentences and its process is shown in Fig. 2. There are three steps: simplifying long sentences, segmenting compound sentences and segmenting subordinate clause.

- Simplification of long sentences [7] means to merging the same part in long sentences, i.e., reducing the number of "words" in the aspect of part of speech. Following is an example of regular expression of simplifying long sentences.

$$\{NN[SP]?\backslash s\}\{2,\} \Rightarrow 1X0 \tag{1}$$

Equation (1) means merging successive nouns. The part of speech of the last word is regarded as the part of speech after merging. The premise of the rule is on the left side of the regular expression, and the execution part of the rule is on the right side of the expression. The word class reduces after a long sentence is shortened by the regular expression, which makes the subsequent matching easier.

- Segmentation of compound sentences [8] means segmenting compound sentences in English long sentences according to categories of sentences. There are two kinds of compound sentence in English long sentences, two clauses connected by punctuation and two clauses connected by conjunction. As the use of grammar of Chinese and English is different, the logical order in compound sentences after translation is also different. Following is an example of regular expression of segmenting compound sentences:

$$\text{^.*?}NN[SP]?\backslash sVB[PDZ].+?,.+?NN[SP]?.+?VB[PDZ]?\backslash s.+?\$ \Rightarrow P0 \tag{2}$$

The left part of equation (2) is corresponding to the compound sentence of the whole long sentence, and the main content of the sentence is expressed by part-of-speech tagging. As to the two parameters on the right side, the former one stands for the connection category of coordinate clause, & stands for conjunction, P stands for

punctuation used for connection; the latter one stands for whether two coordinate clauses need order adjustment, 0 for no need and 1 for need.

- Segmentation of subordinate clauses [9] is a method used for processing English long sentences which include no coordinative constituents or which are still complex after segmentation of compound sentences. Segmenting subordinate clauses can make the structure of long sentences easier. The key of segmenting subordinate clauses is to search for the antecedent of a clause. Following is an example of regular expression of segmentation of subordinate clauses.

$$^{\wedge}.+?VB[PZ?.+?NN[SP]?\backslash sWDT\backslash sVB[PZ]?[^{\wedge}MPWZ]+\Rightarrow WD \tag{3}$$

In equation (3), the left is corresponding to the whole long sentence, and the two parameters on the right stand for categories of the guide word of subordinate clause. Categories of guide word include IW (preposition + guide word of subordinate clause), WD (determiner guide word of WH clause) and WP (pronoun guide word of WH clause) and WC (comma + guide word).

4 The Error-Driven Method

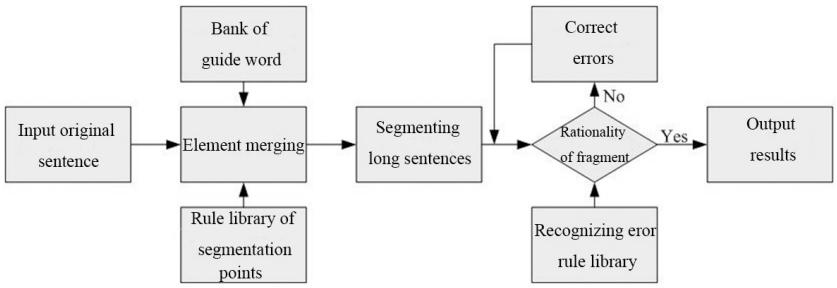

Fig. 3. The segmentation process of long sentences using the error-driven method

The segmentation process of long sentences using error-driven method [10] is shown in Fig. 3. Firstly, elements which will affect segmentation of clauses in long sentences are merged under the assistance of segmentation point rule base and guide word library. These elements include fixed collocations and segmentation points containing guide words. Then they are segmented according to natural segmentation points. After segmentation, the fragments are detected, and errors in the fragments are corrected referring to the recognition error rule base.

4.1 Element merging

Element merging [11] means merging and eliminating elements which will affect segmentation in long sentences. An English long sentence usually contains multiple natural segmentation points, but not all segmentation based on segmentation points is

reasonable. Therefore, unreasonable segmentation points need to be merged as a word. To achieve that target, collocations in segmentation points can be merged using method of exhaustion and referring to the set of collocations in corpus first, and then matching was performed using the method of simplifying long sentences mentioned in the regular match method. The specific rule is:

$$(\backslash b[\hat{}\,ILvM\backslash s\backslash *]\{2,4\}\backslash b\backslash s,\backslash s)\{2,\}(CC\backslash s[\hat{}\,ILvM\backslash s\backslash *]\{2,4\}\backslash s)?\rightarrow 0 \tag{4}$$

$$".\{1,25\}"\backslash s\rightarrow 1 \tag{5}$$

$$VB[PZDNG]?\backslash sRB\backslash s\rightarrow 0 \tag{6}$$

The left side of equation (4) stands for merging coordinate parts which are connected by comma, and parameters on the right side means that the part of speech after merging became the part of speech of the original first word. The left side of equation (5) means that the part inside the quotation mark is a word, and parameters on the right side means that the word is an appositive. The left side of equation (6) means that the verb and adverb which is connected with the verb is merged as a verb, and parameters on the right side means that the part of speech after merging became the part of speech of the original first word.

4.2 Segmentation of long sentences [12]

Elements which will affect reasonable segmentation are eliminated after the processing of the last step, and the remaining natural segmentation points can be normally used for segmentation of long sentences. Therefore, the following procedure is relatively simple. The specific rule is as follows. Firstly, sentences before and after punctuation are segmented. Secondly sentences before and after coordinate conjunction or the guide word of clause is segmented. Thirdly sentences before and after “and” or “or” is segmented. The translation will be easier after segmentation based on the above rule, but the rationality of fragments cannot be ensured after segmentation. Thus, error correction is required.

4.3 Error correction [13]

It is known from the last section that error correction is needed because of the rationality problem of fragments after segmentation. There are four kinds of rationality error, i.e., irrationality of segmentation based on punctuation marks, irrationality of segmentation based on “and” or “or”, irrationality of segmentation based on conjunction and irrationality of segmentation based on the right margin of clause. The reason for the first, second and third error is that the margin is determined only based on natural segmentation points; the solution to such kinds of error is merging the front part to the latter part or merging the latter part to the front part. The fourth error is special, and it can be divided into two situations. If the clause is also segmented, the solution is the same; if the clause is not segmented, then the solution is to search the

right margin of the clause and segmenting at that point. The basis of determining rationality of short sentence segmentation in this study includes the length of sentences, whether punctuation is at the end of a sentence, whether "and" or "or" is in the beginning of a sentence, whether predicate verb exists or not, whether there is conjunction in the beginning and whether there is guide word in the beginning.

5 Experiment on Segmentation of Long Sentences

5.1 Experimental configuration

Four hundred long sentences which were randomly selected from MTCIR-9 patent corpus [14] were taken as experimental materials to verify the recognition and segmentation effects of regular match method and error-driven method. NET Framework3.5 SP provided the procedure code of regular expression, and C++ of NET Framework was used in the writing of code. The whole system program operated on the server of a laboratory which was equipped with Windows7 system, I7 processor and 16 G storage.

5.2 Evaluation criteria

The segmentation effect of the two methods was evaluated using accuracy rate, recall rate and F value [15]. The formula of accuracy rate is: accuracy rate = the correctly recognized number/the total recognition number × 100%. The correctly recognized number refers to the number of sentence elements or structures which are correctly matched using the recognition and segmentation method mentioned in the last section. The total recognition number refers to the number of sentence elements or structures which are matched using the recognition and segmentation method.

The formula of recall rate is: recall rate = the correctly recognized number/the actual existing number × 100%. Actual existing number refers to the number of actually existing sentence elements or structures which ought to be recognized.

The formula of F value (comprehensive evaluation index): F value = 2 × accuracy rate × recall rate/(accuracy rate + recall rate). The corpus selected in this study included the correct Chinese translation, which was used for evaluating the effectiveness of the two methods proposed in this study. BLEU value was taken as the evaluation index, which could reflect the translation quality; the higher the score was, the more accurate the translation was. The index could be used for evaluating the quality of machine translation which did not adopt segmentation and quality of machine translation which adopted segmentation based on the method proposed in this study. The scoring of the index was based on the number of consecutive words in the target text and the similarity and difference of machine translation text and correct text. It could avoid subjective assume of artificial scoring.

5.3 Experimental results

The segmentation effect of regular match method: As shown in Table 1, 400 English sentences were matched and segmented using the regular match method. According to the segmented number shown in the table, the element of clause accounted for a small proportion in the 400 English sentences, but it occupies less proportion in the total sentence. Compound sentences were more than clauses, but its proportion was not signficant.

Table 1. The number of recognized and segmented number in different procedures of the regular match method

Step	Recognized number	Correctly recognized number	Actually existing number
Simplifying long sentences	2871	2514	2769
Segmenting compound sentences	521	439	495
Segmenting clauses	156	127	141

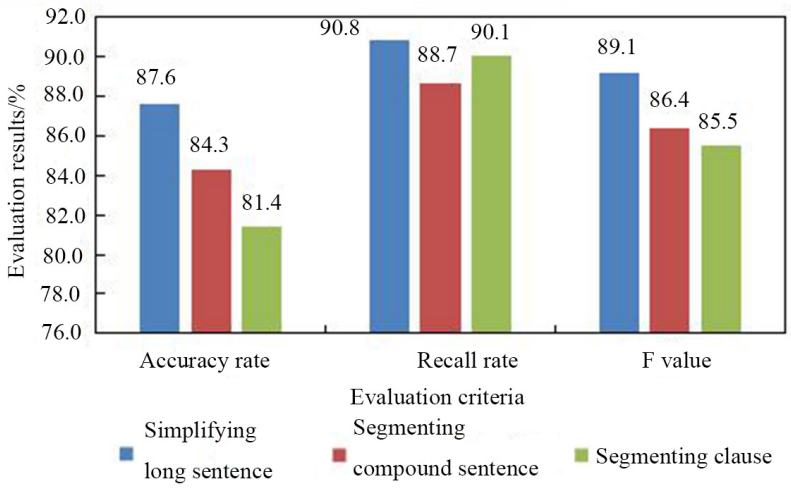

Fig. 4. The segmentation effect of different steps of the regular match method

The segmentation results in different steps of the regular matching method are shown in Fig. 4. The accuracy rate, recall rate and F value of simplifying long sentences were 87.6%, 90.8% and 89.1% respectively; the corresponding data of segmenting compound sentences were 84.3%, 88.7% and 86.4% respectively; the corresponding data of segmenting clause were 81.4%, 90.1% and 85.5% respectively. The experimental results showed that the segmentation method was effective in segmenting English long sentences. However, there were some recognition errors in each step. The error included:

- The wrong labeling of part of speech during simplification of long sentences could result in recognition error. The reason was that the rule of long sentence simplifica-

tion was based on the pattern structure of sentences, and the part of speech of words was a part of sentence pattern in long sentences. If the part of speech was wrongly labeled, then the preset rule would be unable to correctly recognize and merge words. The reason for wrong labeling was that one word usually had two or more parts of speech, and the part of speech was difficult to be precisely determined on the premise of not contacting the context.

- The error appeared in the segmentation of compound sentences was caused by the imperfect recognition rule. Although there are segmentation rules which were summarized based on a large number of long sentences in the compound sentence recognition rule base, the rules could not generalize all the sentences because of the complex change of language, resulting in the omission of segmentation of some sentences, and the wrong part-of-speech tagging described above was also one of reasons.
- The wrong part-of-speech tagging mentioned above was one of reasons for the appearance of errors; on the other hand, and the wrong recognition of the right boundary of a clause. In most cases, a regular expression can find the right boundary of a clause, but it cannot judge the correctness of the right boundary.

The segmentation effect of the error-driven method: Component merging in error-driven method and simplification of long sentences in the regular match method play a similar role in the machine translation of long English sentences. Both of them can reduce the length of sentence pattern. The comparison between the two methods is shown in Figure 5. The accuracy rate, recall rate and F value of the error-driven method were 98.66%, 98.88% and 98.71% respectively, which were much higher than those of the regular match method.

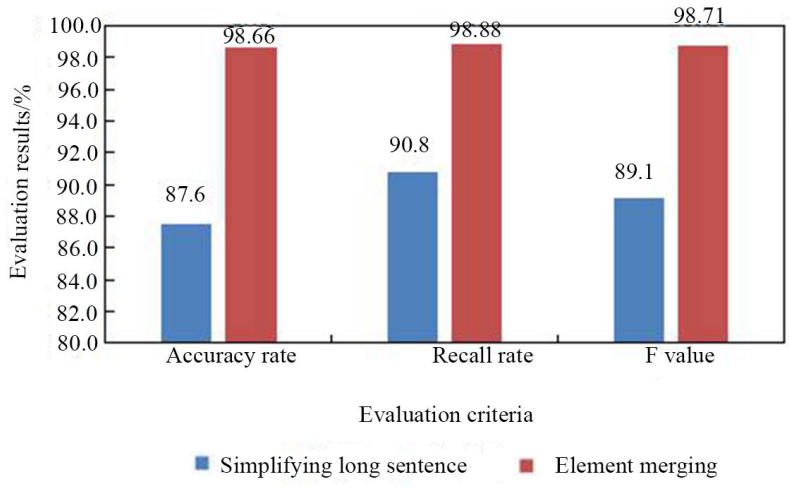

Fig. 5. The effectiveness of simplifying sentence elements with the two segmentation methods

However, there were still some recognition errors in the element merging of the error-driven method, which are caused by the wrong part-of-speech tagging and the wrong recognition of fixed collocations. But in general, the element merging effect of the error driven method was superior in simplifying long sentences than shortening pattern of long sentences.

Table 2. Error correction effect

Number of testing sentence/sentence	Number of fragments before correction/n	Number of fragments after correction/n	Correction times/time	Rationality of fragments before correction	Rationality of fragments after correction
400	2121	1579	511	42.01%	95.11%

After element merging, the system segmented the long sentences according to the natural segmentation points, and then corrected the errors of fragments. The effects before and after the correction are shown in Table 2. The number of fragments before error correction was 2,121, and the rationality was 42.01%. The number of fragments after error correction was 1,579, the rationality was 95.11%, and the correction times was 511. The rationality of the corrected fragments has been greatly improved, but there were still some unreasonable fragments, which were caused by the errors left over by the step of element merging and the special linguistic phenomenon.

Application of the two methods in machine translation: The recognition and segmentation effects of the two methods were tested using Baidu online translation. Firstly, one sentence was randomly selected from the 400 long sentences and directly translated using Baidu online translation. Then the sentence was segmented using the two methods, and the segmentation results were input into Baidu online translation platform. The translation results were merged in order. Finally, BLUE values of the target text obtained by direct translation and the target texts obtained using the two methods were calculated. The final calculation results are shown in Table 3.

Table 3. BLEU values of translated texts obtained by different treatment means

	BLEU value
Translated text obtained by direct translation	0.0745
Translated text obtained after processing of regular match method	0.0781
Translated text obtained after processing of error-driven method	0.0835

It could be seen from Table 3 that the BLEU value of direct translation was 0.0745 and that of regular match method was 0.0781, which was 4.8% higher than that of direct translation; the BLUE value of the error-driven translation was 0.0835, which was 12.1% higher than that of direct translation. Although the translated text obtained by the regular match method was closer to the correct translation than the direct translation, the improvement was not as great as the error-driven method, because of the above-mentioned problems of the regular match method itself and the sequential merging of the translated text.

6 Conclusion

This paper briefly introduced the process of machine translation and explained principles of the regular match method and the error-driven method. Four hundred long sentences which were randomly selected from the NTCIR-9 patent corpus were used for testing the recognition and segmentation effects of the two methods. Moreover, the accuracy of the translated text was compared on Baidu translation platform. The results are as follows.

- The accuracy rate, recall rate and F value of simplifying long sentences, segmenting compound sentences and segmenting clauses were all above 80% in the regular matching method. The accuracy rate, recall rate and F value of the element merging in the error-driven method were all above 98%, which was much higher than those of simplifying long sentences in the regular match method. The rationality of fragments obtained after segmentation and correction improved from 42.01 % to 95.11%.
- The accuracy rate of the translated text which was obtained using the regular match method was 4.8% higher than that of the direct translation; the accuracy rate of the translated text which was obtained using the error-driven method was 12.1% higher than that of the direct translation; and the translated text which was obtained using the error-driven method was closer to the correct translated text.

7 References

[1] Bojar, O., Chatterjee, R., Federmann, C., et al. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. Tenth Workshop on Statistical Machine Translation, 1-46. https://doi.org/10.18653/v1/w15-3001

[2] Firat, O., Cho, K., Bengio, Y. (2016). Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism, 866-875. https://doi.org/10.18653/v1/n16-1101

[3] Sennrich, R., Haddow, B., Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. Computer Science. https://doi.org/10.18653/v1/p16-1162

[4] Tu, Z., Lu, Z., Liu, Y., et al. (2016). Modeling Coverage for Neural Machine Translation, 76-85.

[5] Shen, S., Cheng, Y., He, Z., et al. (2016). Minimum Risk Training for Neural Machine Translation. Meeting of the Association for Computational Linguistics, 1683-1692.

[6] Luong, M. T., Manning, C. D. (2016). Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models, 1054-1063. https://doi.org/10.18653/v1/p16-1100

[7] Bywood, L., Georgakopoulou, P., Etchegoyhen, T. (2017). Embracing the threat: machine translation as a solution for subtitling. Perspectives Studies in Translatology, 25(3): 1-17. https://doi.org/10.1080/0907676x.2017.1291695

[8] Ebrahimi, J., Lowd, D., Dou, D. (2018). On Adversarial Examples for Character-Level Neural Machine Translation.

[9] Marciano, J. P. (2017). Methods and systems for multi-engine machine translation.

[10] Rozovskaya, A., Dan, R. (2016). Grammatical Error Correction: Machine Translation and Classifiers. Meeting of the Association for Computational Linguistics, 2205-2215. https://doi.org/10.18653/v1/p16-1208
[11] Yuan, Z., Briscoe, T. (2016). Grammatical error correction using neural machine translation. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 380-386. https://doi.org/10.18653/v1/n16-1042
[12] Pushpananda, R., Weerasinghe, R., Niranjan, M. Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages. International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, 2015, pp. 545-556. https://doi.org/10.1007/978-3-319-18111-0_41
[13] Germann, U. (2015). Sampling Phrase Tables for the Moses Statistical Machine Translation System. Prague Bulletin of Mathematical Linguistics, 104(1): 39-50. https://doi.org/10.1515/pralin-2015-0012
[14] Jean, S., Firat, O., Cho, K., et al. (2015). Montreal Neural Machine Translation Systems for WMT'15. Tenth Workshop on Statistical Machine Translation, 134-140. https://doi.org/10.18653/v1/w15-3014
[15] Zoph, B., Yuret, D., May, J., et al. Transfer Learning for Low-Resource Neural Machine Translation. Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1568-1575. https://doi.org/10.18653/v1/d16-1163

8 Author

Tiehu Zhang works in the School of Foreign Languages at Xi'an Aeronautical University in China.

Article submitted 2019-01-20. Resubmitted 2019-06-19. Final acceptance 2019-06-19. Final version published as submitted by the authors.