Advances in Economics, Business and Management Research, volume 215
Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

Neural Machine Translation Applied in English Chinese: Turn Grammar Connotation into Words

Tongyao Diao *

High School Affiliated to Nanjing Normal University, Nanjing, Jiangsu 210003, China
**Corresponding author. Email: diao_tongyao@outlook.com*

ABSTRACT
Recent advances in Neural Machine Translation (NMT) have largely improved the quality of translations in a variety of language pairs, which helps promote the efficiency in communication and information exchange. English to Chinese translation, however, still suffers from lower accuracy and intelligibility partly due to their inequivalent language structures. This work compares the performance of three frequently used commercial NMT systems and aims to propose a new method of turning English grammar connotations, participles in particular, into Chinese lexical words. Hopefully, with more data analyzed, the performance of NMT in the language pair of English Chinese can be improved to meet the needs in different contexts.

Keywords: *Neural Machine Translation, English Chinese, participles, grammar connotation, lexical words*

1.INTRODUCTION

NMT has made obvious advancements in recent years with Google introducing the system of Transformer [1], with which previous difficulties such as rare words accuracy and long-distance dependencies have been largely solved. However, while the performance of translation between other language pairs is satisfying, that of English to Chinese remained to be improved, especially in terms of the translations of grammar connotations.

Chinese language does not share similar grammar features as English does, among which the most troublesome during translation are non-finite verbs, especially the participles. In English, participles are economically embedded with information suggesting messages that only lexical words can convey in Chinese. Liu Miqing, who proposed different strategies concerning English-Chinese translation in terms of the rewriting of participles, which do not exist in Chinese, brilliantly predicated that the most troublesome part in it will be the differentiation of the theme and secondary messages [2].

Work had been done to examine the performance of Google NMT (GNMT) before the adoption of the new system, back then the most prominent flaw was the inaccuracy when dealing with long sentence dependencies [3]. Now as GNMT adopts a more advanced model which improves accuracy and intelligibility, the translation of participles, however, is the problem remained to be tackled.

This paper aims to locate the barrier in participle-translation and try to find solutions to improve the readability of the Chinese output by comparing the performance of three different NMTs, namely, GNMT, Youdao Translate (YT), and Systran. Two pieces of source text in English, from a news report and a medical case report respectively, are chosen to be studied. The first is the same sentence examined in my prior work, whose translation quality was largely improved yet the participle remained untranslated in current technology. The second is a typical example of participles indicating time sequence with passive voice in a medical context, where accuracy is expected yet not always met. With the analysis of the advantages of each NMT models, two possible solutions are proposed in the end.

2.THE PERFORMANCE OF NMTS

2.1GNMT

GNMT announced its state-of-the-art hybrid model: Transformer encoder-RNN (Recurrent Neural Network) decoder in 2018 [1]. Compared with the older model, the novel Transformer makes many advances. Yet in terms of putting English participles into Chinese more attention is required in the process. To better demonstrate the

progress and the remaining problems in GNMT, the example source text and the back translations of GNMT's output before and after 2018 are listed below.

Source text (S1): Triggering an elaborately choreographed sequence, she is set to see Donald Tusk, the European Council president, on Friday and Jean-Claude Juncker, the European Commission president, for dinner on December 4 [4].

GNMT (2017): She is set to on Friday with the European Commission president Jean-Claude Juncker on December 4 dinner see the European Council president Donald Tusk's dinner.

GNMT (2022): Triggering a carefully choreographed sequence, she is scheduled to on Friday with European Council President Donald Tusk and European Commission President Jean-Claude Juncker on Dec. 4 to have dinner.

Obvious improvements can be seen from the comparison of the back translations of the same source text. The new model has solved the inherit drawback in the RNN model, which failed to capture the whole adverbial phase due to the limitation of "a fixed-length vector" [5].

The participle of "triggering", however, stayed untranslated in Chinese. As is mentioned before, Chinese language conveys messages through words instead of grammar structures, and thus what has been stated in the original text loses and that causes ambiguity. To be more specific, the hidden subject of the adverbial clause started with "triggering" is "she", and from the context and the common usage of adverbial clauses, readers can tell the sequence of logic order. That is to say, "she" first set to see Donald Tusk, and then "she" is set to see Jean-Claude Juncker for dinner. Furthermore, the sequence of the two separate meetings is carefully planned by her. All the mentioned information economically indicated by the participle must be translated into lexical words in Chinese, otherwise, readers may not be able to solve the complicated meaning. That could be the lost in translation.

Another mistake GNMT 2022 makes is the loss of the verb "see", which might be ascribed to the default of the fixed-length vector as well.

2.2Youdao Translate

Another NMT model which has been prevalently applied in English to Chinese translation contexts is YT. According to its chief scientist Duan Yitao, it adopts the Transformer architecture comprehensively [6]. The back translation of S1 is as follows:

YT: She will meet Donald Tusk, president of the European Council, on Friday and a dinner with Jean-claude Juncker, president of the European Commission, on December 4, which triggered a carefully choreographed arrangement.

There is no missing information, which might be attributed to the self-attention function in the Transformer model proposed by Google, which reduced potential errors in long-range dependencies [1]. The difference between the application of GNMT and YT is that GNMT chooses to apply Transformer as the encoder and leave RNN as the decoder for "higher quality, more training stability, and lower latency" [7] while YT seems to be equipped totally with Transformer, which has been indicated in an interview with the leader of the tech team, Duan Yitao [6].

The shared problem is the translation of the participle "triggering", instead of leaving it untranslated, YT turned it into a non-restrictive attributive clause, where "which" refers to the two sequential matters, and hence lead to an inverted cause-and-effect relationship.

2.3The hidden problem in the translating process

It is obvious that the translation of the participle is the most noticeable problem in this case, so more sentences with participles are examined to locate the specific deficiency. After running the translation of a variety of sentences with participles embedded in NMTs, another repeated error is narrowed down to the participles indicating time sequence. The following is a translation example in a medical context, where precise translation of logic and time sequence is expected [8]. Below demonstrates the source text with a chart (see figure 1) illustrating the timely sequence of the events described in the sentence:

Source text (S2): A 34-year-old bus driver presented in September 1995, with a 24 h history of right-sided headache preceded by teichopsia and variable scotomata [9].

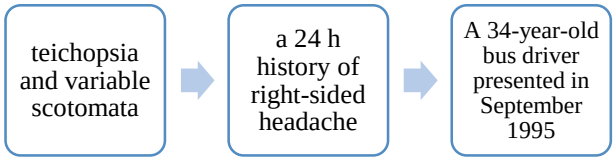

Figure 1 Timely Sequence of S2

GNMT: A 34-year-old bus driver presented in September 1995 with a 24-hour history of right-sided headache, preceded teichopsia (untranslated) and variable scotoma. (If the Chinese output produced by GNMT is put into the source text box again, what it comes up with will be "…, preceded by teichopsia…", but it is not the case in Chinese.)

YT: In September 1995, a 34-year-old bus driver had a 24-hour history of right-sided headache with missing angles and variable blind spots.

Despite mistranslations of terminologies, the time sequence stated in the original sentence, where teichopsia and variable scotomata occurred first, and then the symptom of sided headache appeared, has been altered differently by the two systems. In this case, GNMT gets the sequential symptoms backwards, while YT makes them happen simultaneously. The problem behind might be the failure to turn the grammar connotation of the passive voice embedded in the participle "preceded" into Chinese lexical words.

2.4Systran

Interestingly, when put the same texts into Systran, another NMT system, the participles have been more precisely translated, albeit with more other errors. Systran puts "triggering" in the first sentence into "because triggered …", which is more readable albeit with the nuance emphasizing the deliberate intention with the casual conjunction, "because". That could distract readers' attention from the theme message in the main clause. In the second case, "preceded" is put into "before that there are", which caters to Chinese language style.

The reason of the higher accuracy in participle translation remained unknown, but one of the hypotheses would be its adoption of greedy search algorithm instead of a beam one [10]. Another possibility could be its access to "language specific tokenization", "word segmentation models" for Chinese in particular [11]. Besides the technology basis, looking at the output of Systran, one thing clear is that lexical items can be added to Chinese translations to make up for the loss of grammar connotations in English inputs.

To make the translation more target-reader-friendly in the first source text, the subject in the adverbial clause can be revealed in the target text, for example, "she triggers an elaborately choreographed sequence". As is stated before, participles do not exist in Chinese, let alone the hidden subject, which prevents target readers from comprehending the adverbial clause and its implied suggestion. One way to solve the problem is to reveal the hidden subject in the first place, and Chinese does not follow a very strict SVO structure as English does [12], so the subject in the main clause can be omitted for the sake of conciseness as that has already been indicated before.

Participles indicating the sequence of time in the second sentence can also be turned into lexical words, exactly in the way Systran attempts. Or, word order can be altered to meet Chinese readers' preference, for instance, "teichopsia and variable scotomata as a portent of a 24 h history of right-sided headache". Either way, it is necessary to turn to lexical items to help smooth the language barrier.

Inspired by YT's three-step approach of Chinese Grammatical Error Correction (GEC) [13], it might be helpful to examine the performance of NMTs in participle translation from English to Chinese, identify the errors, replace them with correct ones with the aid of NMT models, and re-rank the sentence. More economically, the solution used to tackle open vocabulary, the technique of subword translation, could be applied in this case, since they share a similar trait that such translations are based on subword units such as morphemes [14]. Take participles in S1 and S2 as examples (see Table 1&2):

Table 1 Chinese Equivalence of English Participles in S1

S1	trigger	-ing
English	lexical word (an action)	grammar connotation (done actively by the subject)
Chinese	lexical word (trigger)	lexical word (she)

Table 2 Chinese Equivalence of English Participles in S2

S2	precede	-d
English	lexical word (a condition)	Grammar connotation (passive voice)
Chinese	lexical word (precede)	lexical word ("bei", a Chinese lexical equivalence of passive voice)

3.CONCLUSION

Albite the development of NMT, the performance of the language pair of English-Chinese has long been suffered from the inequivalent language structures with different grammatical features. Tackled prior problems, the next block to be removed could be the accuracy in translating English grammar connotations. One way that may not have been tried yet is to turn English grammar connotation into Chinese lexical words applying encoding and decoding techniques based on subword units. Hopefully, readability can be increased with precise expression of nuances, which play an important part in information exchange. This paper shows the inaccuracy in translation in terms of English participles and proposes potential solutions. Mainly motivated by prior work with limited examined sentence range, this study will conduct research on NMT-based GEC and NMT models training based on parallel corpus to better address the problem in the future.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin. (2017) Attention is all you need. Advances in Neural Information Processing Systems. pp. 6000–6010

[2] Lui, M. (2006) The Guidance of English-Chinese Translation Techniques. China Translation & Publishing Corporation, Beijing.

[3] Vandeghinste, V. (2010) Scaling up a hybrid MT system: From low to full resources. Evaluation of Translation Technology, 8, pp.65-80.

[4] Barker, A., and Parker, G. (2017) EU and UK aim to strike Brexit divorce deal within 3 weeks. https://www.ft.com/content/525c4bd8-cede-11e7-9dbb-291a884dd8c6

[5] Bahdanau, D., Cho, K., & Bengio, Y.. (2014). Neural machine translation by jointly learning to align and translate. Computer Science.

[6] Jiqizhixin. (2018) Beat two internationally renowned translation engines. Analysis of NetEase Youdao Neural machine translation model. https://www.jiqizhixin.com/articles/2018-12-25-22

[7] Caswell, C., Liang, B. (2020) Recent Advances in Google Translate. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html?m=1

[8] Zhang, J., Li, K., Gong, Y. (2015) The Study of the Strategies of Medical Translation and comprehension under the Perspective of Skopos Theory. Chinese Journal of Health Laboratory Technology (11), 3.

[9] PF Chinnery, DM Turnbull, TJ Walls, PJ Reading. (1997) Recurrent strokes in a 34-year-old man - The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)05005-8/fulltext.

[10] Klein, G., et al. (2020) Efficient and High-Quality Neural Machine Translation with OpenNMT. https://www.aclweb.org/anthology/2020.ngt-1.25

[11] Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., & Senellart, J., et al. (2016). Systran's pure neural machine translation systems. arXiv preprint arXiv:1610.05540.

[12] Shen, J. (2017) Is There A “SV” Structure in Chinese. Modern Foreign Languages (Bimonthly) 40.1:13.

[13] Fu, K., Huang, J., & Duan, Y. (2018). Youdao's winning solution to the nlpcc-2018 task 2 challenge: a neural machine translation approach to chinese grammatical error correction. Springer, Cham.

[14] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. Computer Science. arXiv:1508.07909 [cs.CL]