Scientific Research Publishing

Open Journal of Modern Linguistics, 2020, 10, 765-772
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

Comparing and Analyzing Cohesive Devices of SMT and NMT from Chinese to English: A Diachronic Approach

Jiao Liu

Qinghai Normal University, Xining, China
Email: liujiaofls@126.com

How to cite this paper: Liu, J. (2020). Comparing and Analyzing Cohesive Devices of SMT and NMT from Chinese to English: A Diachronic Approach. *Open Journal of Modern Linguistics, 10,* 765-772.
https://doi.org/10.4236/ojml.2020.106046

Received: November 2, 2020
Accepted: November 23, 2020
Published: November 26, 2020

Abstract

This work presents a detailed comparison and analysis of the usage of cohesive devices by three Machine Translation systems from Chinese to English, in both SMT and NMT situations. By both a general analysis of sentence length as well as cohesive devices and detailed analysis of a sentence translation in SMT and NMT with human translation as a reference, it is shown that, compared with SMT, NMT system is better at handling cohesive ties such as additive, adverbs and pronouns; however, both SMT and NMT underperform at dealing with demonstratives and lexical cohesion. This suggests an evidence of improved translation quality and the necessity of pre-editing and post-editing cohesive devices in MT translations.

Keywords

Statistical Machine Translation, Neural Machine Translation, Cohesive Devices

1. Introduction

Machine Translation (MT) can be regarded as the pearl on the crown of Artificial Intelligence (AI), as people's first practice of automatic translation from one natural language to another exists even earlier than the invention of digital computer. Major developments of MT went through three phrases: Phrase-based Machine Translation (PBMT), Big-data-based Statistical Machine Translation (SMT), Deep-learning-based Neural Machine Translation (NMT). SMT and NMT are both trained with parallel corpus and their input and output are word sequence, but differences lie that SMT needs word-alignment and n-gram language model, while NMT has a single, large neural network, conducted on real values without symbols.

The new method NMT can greatly improve quality of automatic translation. The evaluation of machine translations adopts either human-metrics or automatic metrics to view the overall performance of MT system, famous ones including BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and Meteor (Denkowski & Lavie, 2011), which judge the quality of machine translations in terms of fluency, fidelity, and so on.

However, in this paper, we try a different way and carry out a case study on a text from academic genre by diachronically identifying and comparing cohesive devices of the text, translated by human and three MT systems, i.e. Google Translate, Baidu Translate, and Bing Translator, in 2016 and 2020, relying on SMT and NMT respectively. Cohesive devices, including reference devices and conjunction devices, are one kind of grammatical errors that can be easily found in machine translations and can greatly influence the coherence of target text. Knowing the linguistic differences between English and Chinese language, the author also tries to outline possible strategies for MT pre-editing and post-editing.

The following two aims are targeted:

1) to compare overall use of cohesive devices by human translator, and three MT systems both in SMT and NMT, identifying particular strengths in terms of cohesive devices for the NMT approach compared with SMT approach, and weakness of both NMT and SMT in handling cohesive devices compared with human translator;

2) to examine if the NMT approach can better translate cohesive devices than SMT approach and examine if human is better at dealing with cohesive devices, hence giving suggestions for machine translation post-editing and pre-editing in terms of achieving cohesion of translated text.

MT quality varies a great deal across different language pairs, genres, and domains. Chinese and English are two distinct languages, and automatic translation between the two language pairs underperforms compared with other language pairs such as English to French or Spanish to English. The significance of comparing cohesive devices of Chinese to English translations by different MT systems based on both SMT and NMT approaches outweighs, thus benefiting the development of SMT and NMT systems as well as MT evaluation.

2. Related Works

Limited studies were carried to compare SMT and NMT between Chinese and English in detail, especially through a diachronic way. There was an empirical analysis on Chinese to English news translation done by Microsoft AI & Research center revealed that NMT was at "human parity" compared to professional human translations; however, errors including incorrect words, ungrammatical, missing words, named entity were identified after human analysis, indicating big room to improve the quality even of NMT, the best-state-of-the art at present (Zhou, 2019).

Other analyses between NMT and PBMT approach were carried in (Bentivogli et al., 2016; Toral & Víctor, 2017). Motivated by the advantage of statistical approach, a pre-translation technique was also given to combine PBMT and NMT on the English to German translation task, which uses PBMT to pre-translate source text and generates target text using NMT and increases MT quality measured in BLEU by up to 2 points (Niehues et al., 2016).

3. Cohesive Devices

Cohesive devices are both, however differently, used in English and Chinese language, and many scholars discussed the issue. Guo (2006: p. 188) has compared the translation of textual conjunctive devices between English and Chinese through statistical research, and proved that Chinese emphasizes logical and semantic conjunction, while English focuses on form and structure. Wang (2006: p. 254) indicates in his research A Comparison and the Translation of Grammatical Cohesion between English and Chinese, that cohesive devices are a valued subject in discourse linguistics, while in translation it is paid less attention to because it causes few difficulties for an experienced translator. However, it may bring difficulties to MT systems. Wang (2006: p. 255) holds the view that common existence of cohesive devices in both English and Chinese gives the possibility of translating between them, while differences in their frequency of using cause minor changes when they are translated. He summarizes some differences of cohesive usage with examples, which translators should pay attention to. For example, Wang (2006) thinks that some pronouns should be substituted by other types of cohesive devices like lexical cohesion and ellipsis because pronouns are far more frequently used in English but content words are prominent in Chinese, though we can find their equivalents in Chinese except for some exceptions like “the”, “one”, and “one’s”. He also warns that we do not have to follow form when translating cohesion devices; instead we translate in terms of function (Wang, 2006). Therefore, cohesive devices are regarded as one major source of difficulties in translation.

Cohesion in this paper refers to a series of obvious and language specific resources, which link text together at the global level. These resources include five categories, which are “reference, conjunction, substitution, ellipsis and lexical cohesion” realized by both grammar and vocabulary as in Halliday and Hasan (1976). Reference, substitution and ellipsis are under the term of grammatical cohesion. Conjunction is both grammatical and lexical. Lexical cohesion belongs to the lexical cohesion.

Of the five types, reference includes pronouns that are further classified into “personal pronouns”, “possessive determiners” and “possessive pronouns”, demonstratives such as “this”, “that”, “these” and “those”, definite article “the”, comparatives such as “same”, “similar”, “equal”, “other”, “different”, and adverbs such as “here”, “there”, “now” and “then”. Generally, pronouns in Chinese are simpler in form than in English.

Substitution in English can be nominal (achieved by the use of "one/ones" or "the same" in place of a noun phrase, as in "We have no coal fires; only wood ones"), verbal (realized with the help of "do"/"did" in place of a verb, as in "No one can accomplish this task better than I do"), and clausal (realized through the use of "so" and "not", when they replace an entire clause, as in "Is there going to be an earthquake?" "It says so.") (Halliday & Hasan, 1976: pp. 91-130). According to Hu (1994: pp. 73-74), "这么着", "来" and "干" in Chinese can function as verbal substitution. Chinese does not have as many substitutions as English. However, some lexical cohesion and ellipsis should be replaced by substitutions when translating from Chinese to English (Wang, 2006: p. 269).

Ellipsis occurs when an item is omitted and no tangible substitution happens.

Conjunction can be additive (e.g. "and" and "besides"), adversative (e.g. "but", "yet", "however", "although" and "nevertheless"), causal (e.g. "so", "thus", "hence" and "therefore"), and temporal (e.g. "then", "thereupon" and "later") according to Halliday and Hasan (1976). Conjunctions are more frequently used in English than in Chinese, and therefore, MT systems' ability to process source Chinses text and add proper conjunctions within sentences are of concern in this paper.

The last category, lexical cohesion includes reiteration and collocation. The former refers to the direct repetition of lexical words or the repetition of their synonyms, and collocation means "a word that is in some way associated with another word in the preceding text", including superordinates, hyponyms, and antonyms (Halliday & Hasan, 1976: p. 318). According to Hoey (1991: p. 9), lexical cohesion contributes to probably more than 40% of all cohesive ties in Halliday and Hasan's text samples.

4. Data Collection and Research Methods

In January 2016, when SMT was still adopted, a Chinese-English parallel corpus of 239,504 words including academic, literary, and news texts translated by human and three online MT systems: Google Translate, Baidu Translate, and Bing Translator were collected and complied. Through a quantitative analysis on the corpus, differences of MT systems and human in dealing with cohesive devices were found, e.g. in abstract corpus, MT uses more definite article in Baidu Translate, and more additive devices, i.e. "and", "besides", in Baidu Translate and Bing Translator, compared with human.

As in the latter part of 2016, NMT began to be deployed for users and developers. In 2020, one of the academic texts was retrieved and re-translated by the three MT system, which now both adopt NMT. And in this paper, a diachronic and comparative case study was carried out based on the text, an abstract of a dissertation which includes 1637 words, in 21 short or long sentences. For this time, a qualitative analysis was performed. Having the data, we are able to compare solutions of dealing with cohesive devices between SMT and NMT in three MT system by first holistically overviewing six machine-translated texts, then tracking specific cohesive ties with FileLocatePro in each text, and finally comparing individual sentences translated by MT systems. In this way, we hope to target our research aims.

5. Analysis and Discussion

5.1. Sentence Length of Each Translation

As shown in Table 1, for the same piece of source text of academic genre, 21 original Chinese sentences were remained in the same segments for both three MT systems supported by SMT; however, quite different sentences segmentation for NMT, comparing that human translator tended to separate some long sentences into smaller ones, having 34 sentences in total, which shows that NMT more flexibly deal with sentence length than SMT.

5.2. Overall Usage of Cohesive Devices

To have a general view of cohesive ties used in each text, we used FileLocatePro to track both the human-translated text and 6 MT texts and about 72 cohesive ties were classified as references and conjunctions under the aforementioned classification. The results were shown in Table 2.

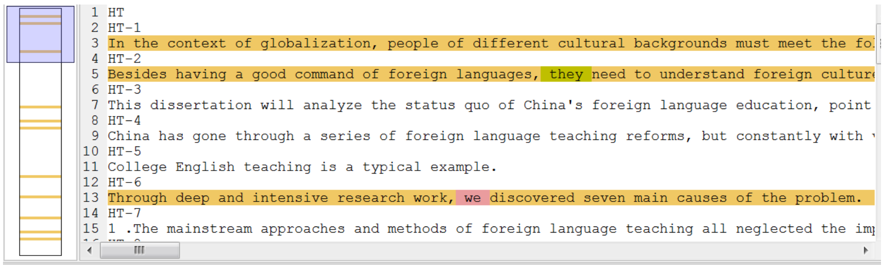
For example, when we searched personal pronouns of text translated by human, all 11 results were located and source sentence with high lightened key words "they" and "we" were clearly shown in the following Figure 1.

Table 1. Sentence length of three MT systems between SMT and NMT.

System	Google Translate	Baidu Translate	Bing Translator
SMT	21	21	21
NMT	37	30	20

Table 2. Reference and conjunction traced in human translation and MT translations.

Cohesive Ties		Human Translator	Google-SMT	Google-NMT	Baidu-SMT	Baidu-NMT	Bing-SMT	Bing-NMT
Reference	Personal pronouns	11	12	11	7	10	10	9
	Demonstratives	11	3	5	3	3	4	6
	Definite article	56	47	50	73	55	39	46
	Comparatives	4	8	7	6	7	8	9
	Adverbs	3	1	2	3	3	3	3
	Personal pronouns	11	12	11	7	10	10	9
Sub-total		**85**	**71**	**75**	**92**	**78**	**64**	**73**
Conjunction	Additive	44	37	47	42	45	41	38
	Adversative	3	4	3	3	3	3	3
	Causal	1	0	0	2	0	0	0
	Temporal	1	0	0	1	0	0	0
Sub-total		**49**	**41**	**50**	**48**	**48**	**44**	**41**
Total		**134**	**112**	**125**	**140**	**126**	**108**	**114**



```
HT
HT-1
In the context of globalization, people of different cultural backgrounds must meet the fo
HT-2
Besides having a good command of foreign languages, they need to understand foreign cultur
HT-3
This dissertation will analyze the status quo of China's foreign language education, point
HT-4
China has gone through a series of foreign language teaching reforms, but constantly with
HT-5
College English teaching is a typical example.
HT-6
Through deep and intensive research work, we discovered seven main causes of the problem.
HT-7
1 .The mainstream approaches and methods of foreign language teaching all neglected the im
```

Figure 1. Location and source sentence of personal pronouns in human translation.

The general calculation of cohesive ties suggests that both references and conjunctions are incorporated in texts translated either by the three MT systems and human translators. However, minor differences can be spotted: 1) demonstratives such as "this", "that", "these" and "those" were less frequently used in both SMT or NMT systems compared with human translator, indicating the incapability of MT systems to add enough demonstratives in machine translations; 2) definite article has 73 hits in text translated by Baidu SMT system, significantly larger than other systems and human translator, for example "*Based on **the** theory of intercultural communication, **the** ideal goal of teaching a foreign language is to let **the** students use **the** language in **the** cultural context of the target language to meet each other's cultural habits of communication is to cultivate **the** students ability of cross culture communication*" (Baidu-SMT-Sentence 18) uses too many definite articles before nouns making the sentence less fluent, however, this feature was less predominant in Baidu NMT; 3) the total classifications of cohesive ties were found more diversified in human translation, as most of MT systems lack of casual or temporal conjunctions; 4) compared with SMT, the number of total cohesive ties tracked above was larger in both Google NMT and BING NMT, still smaller than human translator, which probably showed that NMT was more capable of incorporating cohesive devices into sentences.

5.3. Comparison of Cohesive Devices in Sentence

[***Source***]: *到目前为止，比较系统化的语言教学法流派不下二十种，其中最具影响力的流派有五种：翻译法、直接法、听说法、认知法、交际法。*

[***Human Translator***]: *Up to **now**, **there** are no **less** than twenty systemized approaches **and** methods in language teaching, five of which are **most** influential. **They** are Grammar-translation Method, Direct Method, Audio-lingual Approach, Cognitive Approach **and** Communicative Approach.*

[***Google-SMT***]: *So far, **more** systematic language teaching methods no **less** than twenty kinds of genres, including **the most** influential genre has five: translation method, direct method, **I** heard French, cognitive method, communicative approach.*

[***Google-NMT***]: *So far, **there** are no **less** than 20 schools of **more** systematic language teaching methods, of which five are **the most** influential: translation, direct method, listening **and** speaking, cognitive, **and** communicative methods.*

Taking the seventh sentence translated by Google SMT and NMT for example, human translator divides it into two sentences and incorporates 7 cohesive ties of both 3rd person pronouns, comparatives, adverbs, and additive conjunction which was tracked as above, and some lexical cohesive ties such as "twenty" and "five of which" were also noticeable, while in Google SMT, the number of cohesive ties was only 4, lacking adverbs, proper personal pronouns, though including a first pronoun "I", was used improperly however, as well as necessary additive conjunctions. It seems Google NMT was better at dealing with cohesive ties because the sentence contains enough of them, making the sentence basically a cohesive one. However, Google NMT still failed to handle phrases such as "比较系统化", which was a common usage in Chinese, remaining a comparative "more" in the sentence but obviously violating English grammar. Experienced human translator could easily decide to omit the comparative in this case.

6. Summary and Outlook

In this paper, we had conducted a detailed comparison of cohesive ties between SMT and NMT for Chinese to English language pair. The targets were to identify some strength and weakness of the two systems and raise possible suggestions for pre-editing and post-editing cohesive ties for MT translations. Our findings are:

1) NMT system is better at handling a) additive devices to make English sentence a cohesive one, b) adverbs, c) and pronouns compared with SMT, suggesting an evidence of improved translation quality.

2) Compared with human translator, both SMT and NMT underperform at dealing with demonstratives and lexical cohesion, as Chinese language usually lacks those kinds of cohesive devices, however, easily noticed by experienced translators; and both SMT and NMT are hard to decide where a definite article is needed.

Therefore, to better deal with cohesive devices in Chinese to English translation by MT, it was suggested we incorporate covert cohesive ties in pre-edit, making them easier to be traced by machines, check the use of those ties carefully after translated by MT systems, properly correct them if they were misunderstood or add enough of cohesive ties to ascertain the post-edited machine translations a coherent one in the post-editing process.

It was believed that our analysis would benefit both development of MT system and MT evaluation methods. A more linguistic approach to MT quality, possibly as the analysis presented in this paper, would bring MT quality into a higher stage.

Acknowledgements

The paper is under the project of An Empirical Research on Machine Translation Post-editing in Age of Artificial Intelligence Based on Qinghai Cultural Corpus (Project No. 2020QSK004), sponsored by Qinghai Normal University Scientific Research Foundation for Youth and the Middle-aged. All the project

members made their contributions to the paper.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Bentivogli, L., Arianna, B., Mauro, C., & Marcello, F. (2016). Neural Versus Phrase-Based Machine Translation Quality: A Case Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, 1-5 November 2016, 257-267. https://doi.org/10.18653/v1/D16-1025

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation,* Edinburgh, 30-31 July 2011, 85-91.

Guo, F. Q. (2006). Critical Thinking and Translation Study on English-Chinese Textual Cohesive Ties. In D. F. Wang (Ed.), *Functional Grammar and Translation Study* (pp. 188-200). Guangzhou: Sun Yat-sen University Press.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hoey, M. (1991/2000). *Patterns of Lexis in Text*. Oxford: Oxford University Press, Shanghai: Shanghai Foreign Language Education Press.

Hu, Z. L. (1994). *Cohesion and Coherence of Text*. Shanghai: Shanghai Foreign Language Education Press.

Niehues, J., Eunah, C., Thanh-Le, H., & Alex, W. (2016). Pre-Translation for Neural Machine Translation. *Proceedings of the 26th International Conference on Computational Linguistics,* Osaka, 11-17 December 2016, 1828-1836.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* Philadelphia, 6-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* Cambridge, 8-12 August 2006, 223-231.

Toral, A., & Víctor, M. S. (2017). A Multifaceted Evaluation of Neural versus Statistical Machine Translation for 9 Language Directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* Valencia, 3-7 April 2017, 1063-1073. https://doi.org/10.18653/v1/E17-1100

Wang, D. F. (2006). Comparison and Translation of English-Chinese Grammatical Cohesion. In D. F. Wang (Ed.), *Functional Grammar and Translation Study* (pp. 254-277). Guangzhou: Sun Yat-sen University Press.

Zhou, M. (2019). *The Bright Future of ACL/NLP.* https://www.microsoft.com/en-us/research/uploads/prod/2019/08/ACL-MingZhou-50min-ming.v9.pdf