Korean Journal of English Language and Linguistics, Vol 24, April 2024, pp. 257-276
DOI: 10.15738/kjell.24..202404.257

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474
http://journal.kasell.or.kr

A Corpus-based Multilingual Comparison of AI-based Machine Translations*

Cuilin Liu · **Se-Eun Jhang** (National Korea Maritime & Ocean University)
Homin Park (Electronics and Telecommunications Research Institute) · **Hyunjong Hahm** (University of Guam)

This is an open-access article distributed under the terms of the Creative Commons License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: January 20, 2024
Revised: February 27, 2024
Accepted: March 25, 2024

Cuilin Liu (First author)
PhD student
National Korea Maritime & Ocean University
Claire1182904043@outlook.com

Se-Eun Jhang (Corresponding author)
Professor
National Korea Maritime & Ocean University
jhang@kmou.ac.kr

Homin Park (Co-author)
Researcher
Electronics and Telecommunications Research Institute
hominpark@etri.re.kr

Hyunjong Hahm (Co-author)
Associate Professor
University of Guam
hhahm@triton.uog.edu

* The first draft of this paper was presented orally at the KACL-KASELL Summer Joint Conference, Korea University on June 3, 2023. Some parts of the first draft were also presented orally at the Joint Forum between Korea Maritime & Ocean University and The University of Kitakyushu held at The University of Kitakyushu, Japan on July 27, 2023.

ABSTRACT

Liu, Cuilin, Se-Eun Jhang, Homin Park and Hyunjong Hahm. 2024 A corpus-based multilingual comparison of AI-based machine translations. *Korean Journal of English Language and Linguistics* 24, 257-276.

The present study aims to investigate whether, and to what extent, the corpus linguistic technique type-token ratio (TTR) is valid in identifying the quality of translation productions produced by different AI-based machine translation (MT) systems. Specifically, this study examined the discourse-level discrepancies of MT outputs generated by Google Translate, DeepL and ChatGPT 3.5 on the discourse level utilizing a self-complied multilingual corpus of English translations for the short story Eveline in Korean and Chinese. For this purpose, we calculated the TTR separately for different text segments within a moving span of running word-tokens and visualized the results with a two-dimensional approach. In addition, to verify the validity of this TTR method in predicting the discrepant qualities of the three MT systems, we took a comprehensive reference of three metrics (Bilingual Evaluation Understudy, BLEU; Metric for Evaluation of Translation with Explicit Ordering, METEOR; Recall-Oriented Understudy for Gisting Evaluation, ROUGE) that are commonly used to evaluate the quality of MTs. The paper demonstrated the validity of TTR graphs in assessing the quality of a particular MT system. The findings corroborate the argument in previous studies that AI-based MT produced less lexical diversity and information density.

KEYWORDS

type-token ratio (TTR), span, text structure, machine translation, Google Translate, DeepL, ChatGPT 3.5

© 2024 KASELL All rights reserved

1. Introduction

Recent years have seen a proliferation of automatic translation systems, many of which are available to the general public. Older Statistical Machine Translation (SMT) systems, have given way to more advanced Neural Machine Translation (NMT) systems, which use artificial intelligence (AI) techniques, particularly deep learning neural networks, to automate the translation process. These new systems are designed to be smarter and to excel across a wide range of texts and genres, even when faced with unfamiliar words or contexts. They are often touted as faster and better, with some claiming to have achieved human-like performance, especially for well-supported languages. NMT systems have been continuously improved through more parallel data collection, advances in neural network architectures, and fine-tuning techniques, making them state-of-the-art MT systems. Systems such as Google Translate have undergone iterative updates to improve translation quality. Google Translate transitioned from its previous SMT system to NMT in late 2016 with notable improvements, and continues to update and enhance its service. Likewise, DeepL Translator, introduced in 2017, is an online translation service that also utilizes an NMT system. In 2022, ChatGPT, developed by Open AI, attracted a lot of attention from various fields, including its incorporation of a neural language model. Although the primary purpose of ChatGPT is to provide assistance on a wide range of topics, it has the ability to generate text in multiple languages. NMT systems offer the advantages of generating fluent, contextually appropriate, and accurate translations. However, the implementation and optimization of NMT models and systems can vary. This study aims to explore translation quality discrepancies among NMT (AI-based MT) systems, including Google Translate, DeepL, and ChatGPT 3.5. We chose ChatGPT 3.5 Playground for the translation task due to its user-friendly interface. The Playground allowed for efficient generation and analysis of translations, with the capability for immediate adjustments to prompts and parameters. This approach enabled us to closely monitor and control the quality of inputs and outputs, which was essential for the comparative analysis conducted in our research. We opted for ChatGPT 3.5 over version 4.0 to ensure consistency throughout our study. Our initial purpose is the comparison between most widely used AI-based MT systems on a snapshot of time. At the outset, ChatGPT 3.5 was widely used within both academic circles and the broader community, making it an ideal baseline for comparing the most commonly utilized AI-based MT systems at that time.

The rapid advancement of MT systems gives power impetus to research in computational linguistics and corpus linguistics. Studies on quality evaluation metrics for MT system outputs have resulted in powerful automated metrics such as BLEU, METEOR, and ROUGE. However, these metrics do not always align with human judgment, implying that strong performance on automated metrics does not guarantee high translation quality from MT systems. Similar to corpus-based studies on the universals of human translations (referred to as 'translationese' by Baker, 1993), MTs are being examined for their systematic and recurring tendencies, termed 'machine translationese'. Yet, the findings and conclusions so far appear to be inconsistent and even contradictory. In addition, researchers have raised concerns about MT, such as the potential loss of lexical variation in the target language (Roberts et al. 2020), loss of lexical richness (Vanmassenhove et al. 2019), potential lexical impoverishment of the target language, strange invented translations, all of which might have side effects on language learners through post-editing learning (Brglez and Vintar 2022, Kruger 2012, Shin and Chon 2023).

The motivation for the present study stems from two pivotal observations in the domain of MT. First, recently more effort has been directed towards identifying 'machine translationese' (MT universals) across different systems and languages. One such feature is the simplification of outputs produced by MT in comparison with human translated texts, which has been observed and measured using the type-token ratio (TTR)—a statistical measure of lexical diversity. Nevertheless, the application of TTR as a universal metric raises questions when we

© 2024 KASELL All rights reserved

consider variables like language specialization, translation directionality, and the influence of various registers and genres on translation. Second, regarding MT evaluation, most of the evaluation metrics, such as the three above-mentioned prevalent metrics, rely on superficial correspondence between reference and candidate translations, lacking in depth. For example, Shin and Chon's (2023) recent study highlights mistranslations—misrepresentations of source text meaning—as the most prevalent error and reveals potentials for multiple errors within a single sentence. Obviously, the present evaluation metrics cannot sufficiently account for such errors, underscoring the need for more refined MT evaluation tools which can catch the pace of the advancement of MT systems and provide a more exhaustive and detailed error analysis.

The two observations inspired the present study from the following two aspects. Firstly, TTR or TTR-based measures have proven effective in exploring the lexical diversity of machine translations (MTs), a characteristic that may vary across different language pairs and directions of translation. To our knowledge, no existing research has extensively applied TTR to more in-depth analyses, such as discourse analysis of translated texts. Secondly, prevailing MT evaluation tools often fail to adequately assess the quality of AI-based MT outputs, as they rely on shallow comparisons and evaluations at the sentence level. To fill these gaps, the present study aims to assess the validity of a TTR-based measure in revealing quality discrepancies among the three above-mentioned AI-based MT systems. This paper initially focused on the Korean and Chinese human-translated versions of Eveline and investigated whether the visualized TTR curves for the AI-based MT outputs, specifically Google Translate, DeepL, and ChatGPT 3.5, are similar to those for the Human Translations (HT). It is hypothesized that the TTR concept applied to English texts can predict patterns in their corresponding Chinese and Korean translations. Furthermore, since AI-based MT systems are trained on extensive human-translated texts, a higher similarity between their TTR lines and HT TTR lines should imply higher translation quality. Therefore, the main questions for the present study are:

Question 1. Do the TTR curves for the translated Chinese and Korean texts effectively detect the three discourse boundaries claimed by literary critics?

Question 2. Do the TTR graphs for the outputs of the three AI-based MT systems identify any discrepancy among them per se?

Question 3. How can we judge which is better or worse?

The primary question is the extent of the difference between machine translation and human translation in terms of lexical diversity and text structure.

2. Literature Review

2.1 Theoretical Considerations on Lexical Diversity and Information Low

Lexical richness measures are commonly used in applied linguistic research to assess the vocabulary diversity and complexity of a text, such as lexical density or lexical sophistication (Laufer and Nation 1995, O'Loughlin 1995). TTR is undoubtedly the most widely used measure to assess the extent of linguistic simplification in translations, measuring how the diversity of vocabulary in the original language is maintained or reduced. Essentially, TTR measures the ratio of unique words (types) to the total word count (tokens) in a text. However, a widely acknowledged issue is that TTR values are influenced by text length. They tend to decrease in longer texts due to increased word repetitions (Brezina 2018). This is because extended texts are more likely to contain repeated usage of the same words, resulting in lower TTR scores. To overcome this problem, several alternative measures have been developed. The Mean Segmental TTR (MSTTR) proposed by Engber (1995) calculates the mean TTR of consecutive text segments of equal length. Guiraud's index measures the number of types over the square root of the tokens, thus reducing the influence of token length (Broeder et al., 1993). Yule's K (Yule 1944) takes into account the frequency distribution of word lengths in a text and quantifies how different the distribution of word lengths is from what would be expected if lengths were uniformly distributed. A measure developed specifically for child language acquisition is the D-measure (Malvern et al. 2004), which models the rate at which new words are introduced in increasingly longer text samples through a curve-fitting procedure using a single parameter, the parameter D. Other TTR transformations include measures such as Textual Lexical Diversity (MTLD) (McCarthy 2005), Herdan's Index, or Uber's Index (Vermeer 2000).

Some researchers make good use of the text length dependence of the TTR. If the text is considered as a flow of linear words, it can be seen that the TTR decreases as the text gets longer because the number of word-tokens keeps increasing while the number of word-types increases more and more slowly. Some researchers have related this phenomenon to the discourse frame of information structure (Chafe 1994). Tuldava (1998), for example, argues that speaker-writers must constantly choose between "old" and "new" words, either consciously or under the influence of grammatical rules. Youmans (1991) extended Chafe's (1987) three categories of informativeness—given, accessible, new information—by analogy with vocabulary management in text. According to Youmans (1991), repeated words could be classified as "given information", new function words as "accessible information", and new content words as "new information". In order to visualize the information flow, i.e., to signal the changes of topics, he proposed a new method called Vocabulary Management Profiles (VMPs), a more sensitive quantitative indicator than type-token curves. Instead of visualizing the relation of types to tokens, the adapted calculation is focusing on △y/△x, where △y equals the number of new types, and △x equals the number of new tokens in the interval, to observe the rate of change over a finite interval. Based on the adapted algorithm in his study, he visualized ebb and flow of new information in a text and argued that major valleys on VMPs correlated very closely with the boundaries between major constituents of discourse. In well-edited narrative stories, he found rhythmic alternations between new and repeated vocabulary. Similarly, Stubbs (2001) showed that speaker-writers have to consider two opposing issues: one is that new words are needed to develop and expand the topic, and the other is that old words are needed to make the text cohesive. These arguments emphasized that speaker-writers make choices from the available vocabulary by alternating between repeating old words and introducing new words, which influence not only the whole texts but also the smaller sections of the texts. Stubbs also emphasized that a marked increase in the frequency of new words towards the end of a text was very likely to indicate a significant boundary. The underlying cognitive reason is that as the story progresses, the number of new words is expected to steadily decrease due to the repetition of "old information". If the cluster of new words appears late in the text, new information or, in this case, new story episodes are expected. He applied this knowledge and used the method proposed by Youmans (1991) to analyze the overall text structures of *Eveline* written by James Joyce, and he examined the sensitivity of TTR to identify significant boundaries within the story, where he found the agreement of the quantitative results with the literary critics[1] made by Hart (1969).

[1] Hart (1969) identified three phases in the Eveline story: a long phase in which Eveline reflects on her past, present, and possible future (lines 1-110); a second phase in which she "reaffirms her decision to choose life" (lines 111-132); a third phase culminating in her "psychological failure" (lines 133-158).

2.2 Automatic Evaluation Metrics for MTs

When it comes to evaluating the quality of MTs, several approaches are typically used, including automatic quality metrics such as BLEU or METEOR, as well as ROUGE. BLEU, introduced by Papineni et al. (2002), compares the n-grams of the MT with the n-grams of the HT. It also takes into account the brevity penalty to avoid favoring excessively short translations. The intuition behind BLEU is that a good translation should have similar n-gram sequences to the reference translations. Compared to BLEU, METEOR, developed by Banerjee and Lavie (2005), captures not only word overlap but also fluency and appropriateness. It incorporates synonyms, word order, stemming, and other factors to provide a more comprehensive evaluation. METEOR has been shown to correlate well with human judgments and can provide a broader perspective on translation quality. As for ROUGE (Lin 2004), it is not primarily used to evaluate machine translation (MT), some researchers have adapted and applied it to evaluate MT output by considering the machine-generated translation as the summary and the reference translations as the reference summaries. In other words, the metrics compare an automatically generated summary or translation with reference (high-quality and human-generated) summaries or translations.

Recently, TTR has been a very useful tool in MT research. Bentivogli et al. (2016) used lexical diversity, measured by TTR, as an indicator of vocabulary size as well as topic diversity in a text. They compared statistical MT with NMT, and the results suggest that NMT is better at dealing with lexical diversity than statistical MT. Vanmassenhove et al. (2019) conducted a study on the lexical richness of MT systems. They analyzed the output of 12 different systems using original and back-translated data. The researchers observed a decrease in lexical richness, with more frequent words appearing more often and less frequent words appearing less often. They also compared phrase-based systems with neural systems and found that phrase-based systems had greater lexical variety. They proposed that neural systems experience an even greater loss of linguistic variation due to their tendency to favor the most likely solutions and overlook rarer words. In another experiment focusing on algorithmic bias in the training of MT systems, Vanmassenhove et al. (2021) compared the training data (HTs) with the outputs of MT systems trained on the same dataset but using different architectures: a phrase-based statistical system, a long short-term memory (LSTM) neural network, and neural transformers. The researchers found that phrase-based systems produced the least diverse translations. However, unlike the previous study, the neural systems used byte pair encoding (BPE), which allowed them to translate rare or unseen words by segmenting them into smaller pieces. As a result, the neural transformer models consistently showed higher lexical diversity compared to neural LSTM and phrase-based statistical models. The researchers also found, on average, a positive correlation between diversity metrics and translation quality metrics.

Although some linguists have studied MT using the TTR technique, there seem to be few studies that use TTR by calculating it separately, and even fewer that test this method on texts translated by either human or MT systems. The present study will extend the TTR-based measure to detect the discourse flow of the translated Korean and Chinese story, which are entirely different languages from English, with the purpose of revealing the discrepancies among the Korean and Chinese outputs from the three earlier-mentioned AI-based MT systems. In addition, the present study will not only compare the translation variants produced by different MT systems by replicating Stubbs' method through self-coding and revising to detect the discourse segments in the Korean and Chinese translated texts, but it will also compare the degree of similarity of the TTR graphs representing different MT outputs by using Dynamic Time Warping (DTW), a technique used to measure the similarity between two sequences that may vary in length or speed.

© 2024 KASELL All rights reserved

3. Method

3.1 Data Collection, Processing and Calculation

We obtained the original English story *Eveline* as well as its Korean and Chinese professional HTs. The short story *Eveline* including 1,836 words was retrieved from the e-book Dubliners written by James Joyce, which was downloaded from the website “Project Gutenberg”. Both the Korean and Chinese human-translated versions of *Eveline* were selected from the translated Dubliners book by professional translators, which were then carefully evaluated to determine the most suitable ones for the method of the present study. Among the accessible digital sources of the translated Dubliners, the Korean version translated by IL-Dong Han and the Chinese version translated by Fengzhen Wang were selected because they are the most widely recognized by the local translators. Apart from the HTs, both the Korean and Chinese translated versions generated by the three above-mentioned AI-based MT systems were also collected.

All obtained translations were initially tokenized using Part of Speech (POS) tagging, and these tokenized translations were then computed as word-tokens for the purpose of making comparisons with the original English text. And for English, the present study identifies the words as graphic tokens—word divided by spaces; only hyphens and apostrophe are included (Francis and Kučera 1982). TagAnt 2.0.5 was used to tag the Chinese translated texts and KiwiGui v0.15.0 was used for the Korean translated texts. Human inspection and revisions were conducted by both the Chinese and Korean researchers of the present study, in case of some mistakes during the automatic POS tagging. The primary objective was to guarantee consistent POS classification between the human-translated texts and their corresponding machine-translated versions. Punctuations were removed to eliminate noise in word-token counting. The Jupyter tool within the Python programming environment was used to develop the computational program and generate TTR graphs for the analysis.

The TTR was calculated separately within a segment of the text rather than for the entire text. As mentioned earlier, the span is of great importance to get the result of interest for the purpose of the present study. The length of the span predetermined to calculate the TTR was an intuitive decision aimed at obtaining interesting results that could reveal the three significant boundaries of the story[2]. Put simply, if the span was too short, the line turned out to be rather jagged, making it challenging to observe the three phases. Conversely, if the span was too long, it became too smooth, and no boundaries could be detected, as illustrated in Figure 1 below.

During the procedure, it was found that the suitability of the span is related to the total length of the text. Roughly, the longer the text (more tokens), the longer the span should be. But it should be careful to generalize this experience to even larger dataset since both the English texts and translated texts were within three thousand words. We followed the previous study to set a span of 151 for English original text, and then moved the span token by token through the text, from word 1 to 151, from 2 to 152, and so on. For each span, the program calculates new types. The ratio of new types/span (△y/△x) was stored and later displayed as the value of the y-axis, referred to as TTR value. The word-tokens were counted individually and displayed on the x-axis with the value span/2 as the starting point and increasing sequentially. The primary goal of this algorithm was to observe the dispersion of newly introduced words by the author, which will help to identify clusters of new information and infer new story episodes. Different spans were set for Korean and Chinese translated texts (for Korean, span=251; for Chinese,

[2] This suggestion came from Dr. Michael Stubbs in a personal email on May 19, 2023. We would like to thank him for these ideas.

span=131), with the MI-based machine-translated texts being the same as human-translated texts. The identical procedure was then followed.

To examine the quality of these MT productions and make comparisons with the results obtained from the TTR analysis, we also relied on a set of comprehensive MT evaluation metrics (BLEU scores, METEOR, and ROUGE Scores). All three metrics require human-generated translations as a standard baseline/reference for comparison. BLEU scores are typically reported as BLEU-1, BLEU-2, BLEU-3, etc., depending on the n-gram order considered. BLEU-4, which considers 4 grams, is a common choice in MT evaluation and was used as the measure in the present study. For the METEOR score, the calculation involves several steps. First, the algorithm matches words and phrases in the machine-generated output with those in the reference translations. It then computes several measures, such as unigram precision, unigram recall, and alignment score, which capture the degree of similarity and alignment. ROUGE (the ROUGE Score) which includes three main metrics ROUGE-N, ROUGE-L, and ROUGE-S, has carved its niches in Natural Language Processing (NLP) and one of the pivotal fields is MT. In the present study, ROUGE-L[3] was adopted with beta set to 1 (F1). Different from ROUGE-N, ROUGE-L calculated the longest common subsequence (LCS) of the MT and the reference translation, in this case HT. Focusing on LCS allows ROUGE-L to evaluate sentence-level coherence and the order of information, which is crucial for the quality of translation. Correspondingly, for ROUGE-L, the composite factors are LCS-based recall and precision, i.e., ROUGE-L recall, and precision were calculated by calculating their agreement with respect to the LCS. The present study follows the equation of ROUGE-L proposed by Lin in 2004, hence, the ROUGE-L F1 score is computed as follows:

$$\text{ROUGE-L F1-score} = (2 \times P_{LCS} \times R_{LCS}) / (P_{LCS} + R_{LCS}) \quad (1)$$

In NLP, precision (P), recall (R), and the F1 score are critical metrics for evaluating model performance, particularly in tasks like text summarization, MT, and information retrieval. For MT, precision quantifies how many of the words or phrases produced by the translation model (candidate translation) are correct or relevant when compared to a set of reference translations. Recall assesses how many of the words or phrases in the reference translations are successfully captured by the candidate translation. The F1-score, essentially a machine learning metric, is the harmonic mean of precision and recall, providing a single metric that balances the two by considering both the model's ability to retrieve relevant information (recall) and its ability to exclude irrelevant information (precision). The scores produced by these metrics are typically normalized to a scale between 0 (no overlap or similarity) and 1 (a perfect match or similarity) for ease of interpretation.

[3] Reference *https://thepythoncode.com/article/calculate-rouge-score-in-python* where a brief explanation of the advantages of ROUGE-L compared to ROUGE-N as well as examples of python code to calculate ROUGE-L score is available.

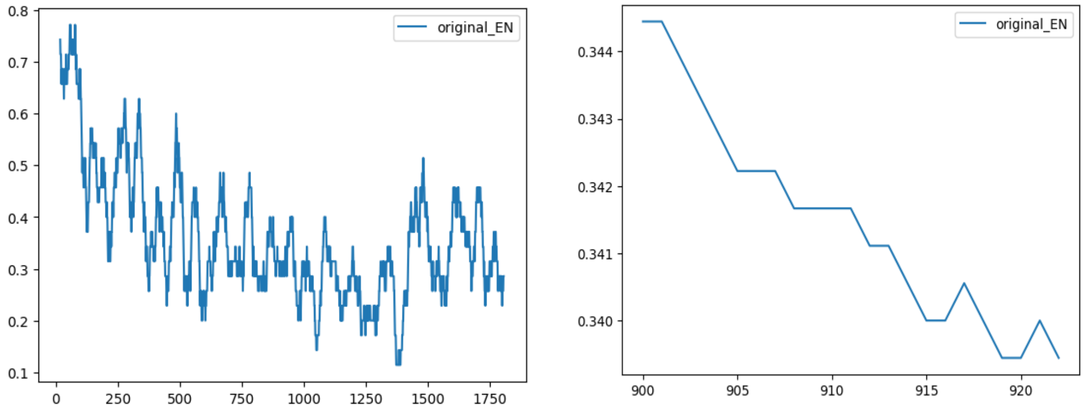

Figure1. Original English *Eveline* with Different Spans (left, span=151; right, span=1800)

3.2 Procedures of the Analysis

The present study mainly used quantitative analysis and investigated the MT discrepancies mainly based on the formal features of the text. In the first step, the TTR lines of the original English text were drawn and compared with the human-translated texts in both Korean and Chinese to examine whether the significant boundaries of the translated texts could be identified. As mentioned earlier, Stubbs (2001) has obtained results through quantitative analysis that are consistent with certain professional literary critics. Thus, this step seeks to investigate the applicability of such a TTR method to Korean and Chinese translated texts. Given that the last two distinct phases are near the ending part of the story (Hart 1969, an effective strategy for identifying significant boundaries in the TTR graph is to locate the lowest TTR value towards the story's end. A notable increase following this low point often signals the start of a new episode. Observation should then shift to the next significant low point prompted by a marked decrease following that increase, to pinpoint further significant boundaries.

In the second step, the discrepancies between the three different systems were investigated by comparing the TTR graph of the HTs with that of each AI-based MT system for Korean and Chinese languages, respectively. DTW (Sakoe and Chiba 1978) was utilized to evaluate the degree of similarity between the HT and the corresponding MT outputs. The primary goal of DTW is to measure the similarity between two sequences that may vary in time or speed. In the context of this study, DTW is proper to compare the distance or rather measure the similarity between sequences of different lengths, in this case, the comparison between the values of HT TTR graph and each MT graph. This evaluation was done by calculating the average score across the three segments of the story. It is worth noting that the three phases of the story are characterized by uneven proportions within the overall narrative, so the relative weights and proportions of these phases were carefully considered in calculation. Another important point is that since the HT was considered the reference, the choice of HT would affect the outcome of the comparison.[4]

In the third step, the scores of the Korean and Chinese outputs generated by the three AI-based MT systems were calculated utilizing the three above-mentioned MT evaluation metrics. The ranking of the scores was then

[4] We repeated the research for two different HTs of Chinese and found consistent results. In other words, the quality ranking of the three MT systems remained the same when examined on two different HT benchmarks.

© 2024 KASELL All rights reserved

compared with the results obtained from the analysis of TTR. Regarding the degree of similarity between the TTR graph of each MT system and the HT, we implemented the DTW algorithm and the DTW distances were transformed into values ranging from 0 (completely dissimilar sequences) to 1 (identical sequences) using the similarity equation:

similarity = 1 / (1 + DTW distance) (2)

The transformation that equation (2) defines is based on the idea that as the DTW distance approaches 0 (indicating very similar or identical sequences), the similarity measure approaches 1. Conversely, as the DTW distance increases (indicating dissimilarity), the similarity measure approaches 0. ‘Distance’ in the DTW includes point to point distance, cumulative distance and optimal path distance (Sakoe and Chiba 1978). In the present study, DTW distance is computed using a custom function named dtw_distance, which is implemented in Python utilizing the NumPy library for numerical operations.

4. Results

In this section, we initially compare the English original text with the human-translated texts in both Korean and Chinese to gain insights into the first question. The findings from this initial comparison lay the groundwork for addressing the feasibility of the method in addressing the second and third questions. Subsequently, quantitative calculations were employed to compare the outputs of different AI-based MT systems.

4.1 Comparison between English Original Text and HT in Korean and Chinese

In this section, we analyze the results of comparing TTR graphs of the English text with its Korean and Chinese translations and examine the applicability of this TTR-based method to identify text structure in translations. In the following, “English” will be abbreviated as “EN”, “Korean” as “KR”, and “Chinese” as “CN” when discussing or comparing the translations.

In Figure 2, three lines of TTR were plotted, with the English text having a total of 1,820 tokens, the KR human-translated text having 2,794 tokens, and the CN human-translated text having 1,807 tokens. The TTR line for English text shows the three phases of the narrative story, with the blue vertical lines marking the corresponding boundaries, which is consistent with the result of Stubbs’ 2001 study. The first phase ended when the number of tokens reached about 1,313, where the lowest point occurred with a coincident TTR of less than 0.2. The second phase ended when the number of tokens was about 1,571 with a coincident TTR of about 0.3, higher than the previous part. Two subsequent peaks indicated the “lexical frenzy” (O’Halloran 2007) associated with the introduction of new episodes near the end of the story. Moreover, the peak of the last phase was even higher than that of the second phase.

© 2024 KASELL All rights reserved

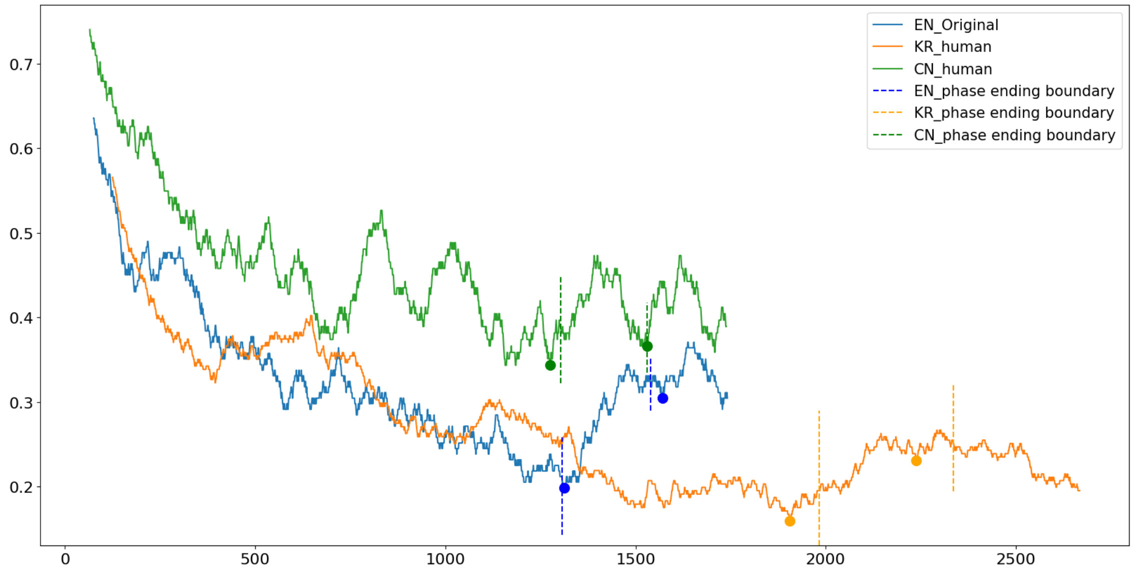

Figure 2. English Eveline, Korean and Chinese HT Graphs (EN, span=151; KR, span=251, CN, span=131)

For CN and KR human-translated texts, two significant boundaries within the text were also correctly and unambiguously identified. For KR, the first phase ended when the number of tokens reached 1,905, with a corresponding TTR value of about 0.16, and the second phase ended at the 2,238th token, accompanied by a TTR value of about 0.23. Subsequently, three nearby peaks were observed. However, it is noteworthy that the last phase begins after the first peak, which suggests that the Korean human translated text for the last phase lack lexical richness in comparison to the English source text. In the case of CN, the lowest TTR value is around 0.35 and the corresponding number of tokens is around 1,275. Additionally, two peaks appeared in the late part of the graph, corresponding to the second and third phases of the story. The second phase ended at the point where the number of tokens is about 1,530, with a TTR value of about 0.37, higher than the end point of the first phase. To verify the correctness of the two distinct points, the corresponding translated Korean and Chinese tokens near the two significant boundaries were retrieved, and the results of both languages showed good agreement with their TTR graphs (not listed here). To make the three TTR graphs comparable, the three curves were normalized and plotted with the minimum length of the TTR line (Chinese TTR line) as the reference, as shown in Figure 3 below.

Figure 3 made it easier to identify the three boundaries for all three TTR lines and it also clearly illustrates the differences in text structure recognition between English and its translations. It is worth noting that the third peak in both the Korean and Chinese TTR graphs is not significantly higher than the second peak, whereas in English it is. This suggests that human translation may not accurately reflect the lexical density of the original story in the third new phase. Another notable observation is the presence of two distinct peaks in the Chinese TTR graph, absent in both the English original and the Korean translation.

© 2024 KASELL All rights reserved

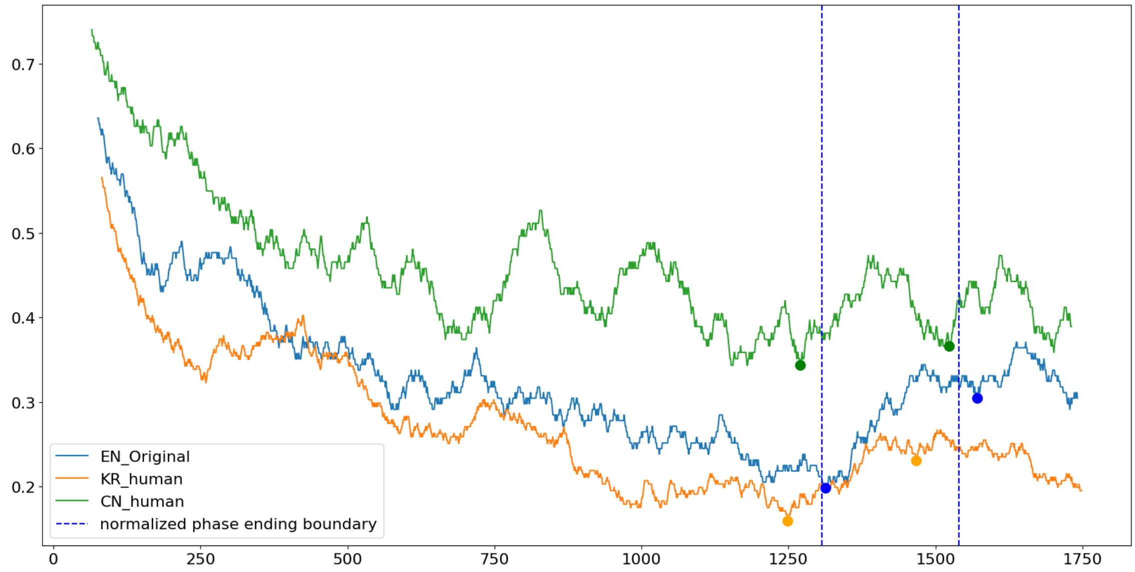

Figure 3. English Eveline, Korean and Chinese HT Graphs for Normalized Comparison

The above results indicate that the separately counted TTR is applicable in detecting the overall text structure of both Korean and Chinese translated texts. However, it seems to be sensitive to different linguistic features. It also sheds some light on the hypothesis of 'universals' of translated language, which typically include simplification, explication, normalization, and shining-through (Ilisei 2012). The "loss" of lexical frenzy at the end of the story in both translated versions is likely to indicate the tendency to simplify the language used in translation, and the two distinct obvious peaks shown in the middle of the graph of the Chinese translated text may be related to the tendency to exaggerate features of the target language and conform to its typical patterns (Teich 2003).

4.2 Discrepancies of Different AI-based MT Systems

After verifying the applicability of this TTR-based method for the present study, this section is dedicated to the analysis of translation quality differences between three AI-based MT productions: ChatGPT 3.5, Google Translate and DeepL. The HT TTR graphs of KR and CN served as the benchmark to measure their quality discrepancies, with the results discussed in 4.2.1 and 4.2.2. Three dimensions were considered when observing the results: the overall text structure of the MT productions, the overall TTR value, and the similarity between HT and the output of the MT systems.

4.2.1 Comparison between Korean HT and outputs of AI-based MT systems

To better identify the characteristics of each TTR curve, the comparison is made separately for each MT system. The TTR curves for a specific MT system and the reference Korean HT are graphed as Figures 4, 5, and 6, respectively. The result of DTW analysis for the three MT systems is shown in Table 1.

In Figure 4, the Google Translate TTR curve displayed two distinct rises and even more marked peaks than HT TTR curve in the third phase. This suggests that Google Translate delineates the three narrative phases of the story with marked clarity. In Figure 5, the ChatGPT TTR curve steadily decreased as the span moved token by token although the last part showed a slight rise, indicating that ChatGPT failed to identify the three significant story boundaries. In Figure 6, DeepL TTR curve showed a prominent increase in the second part and rose again in the last part, but with a lower peak in comparison to that of the second phase. This indicates some distortion in the third phase of the story. With regard to the overall TTR value shown by the entire TTR line for each MT system, it is important to note that the TTR line of either ChatGPT or Google Translate is below that of HT for most parts, whereas about half of the TTR line of DeepL is above that of HT. This result indicated, to some degree, that the lexical diversity of the Korean translated text by DeepL was closer to that of HT, whereas the other two MT systems showed lower lexical diversity compared to HT. This finding was consistent with some previous studies that argue that MT tends to have less lexical diversity than HT (Castilho et al. 2019, Toral 2019).

The results for each phase and the overall weighted average DTW similarity score between KR HT and MT systems were shown in Table1. The TTR similarity between Korean HT and DeepL took the lead with an average score of 0.849, followed by Google Translate with a score of 0.813 and ChatGPT with a score of 0.557. This result indicated that DeepL had the best quality in the translation of the narrative story, followed by Google Translate and ChatGPT.

Table 1. DTW Similarity between KR HT and MT Systems

Korean	Google Translate	ChatGPT	DeepL
Part 1	0.794	0.402	0.875
Part 2	0.705	0.841	0.618
Part 3	0.975	0.981	0.907
Average (Ranking)	0.813 (2nd)	0.557 (3rd)	0.849 (1st)

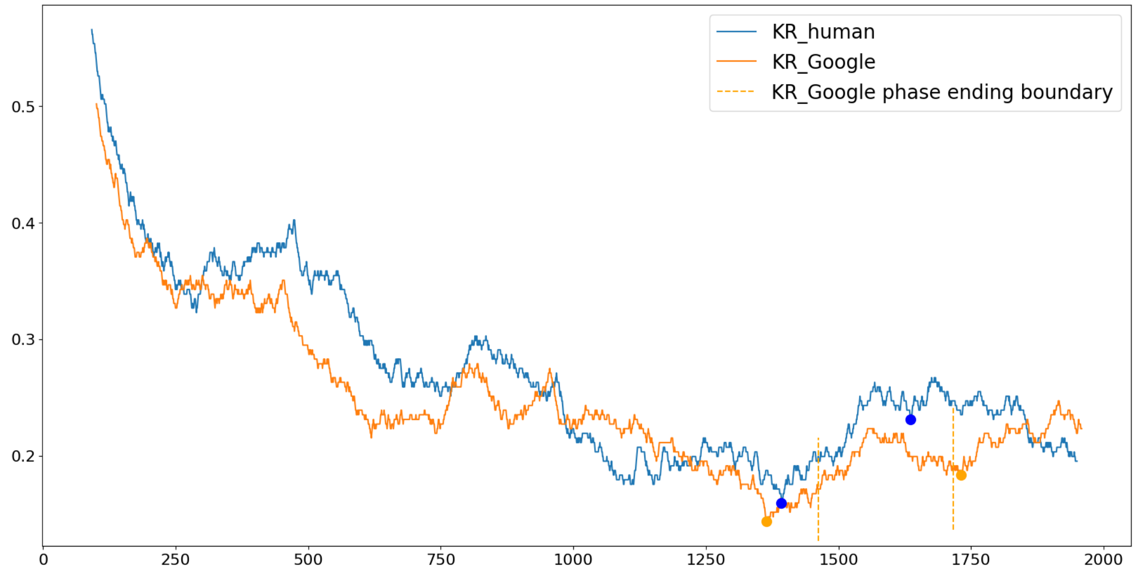

Figure 4. Comparison of KR HT TTR and Google Translate TTR

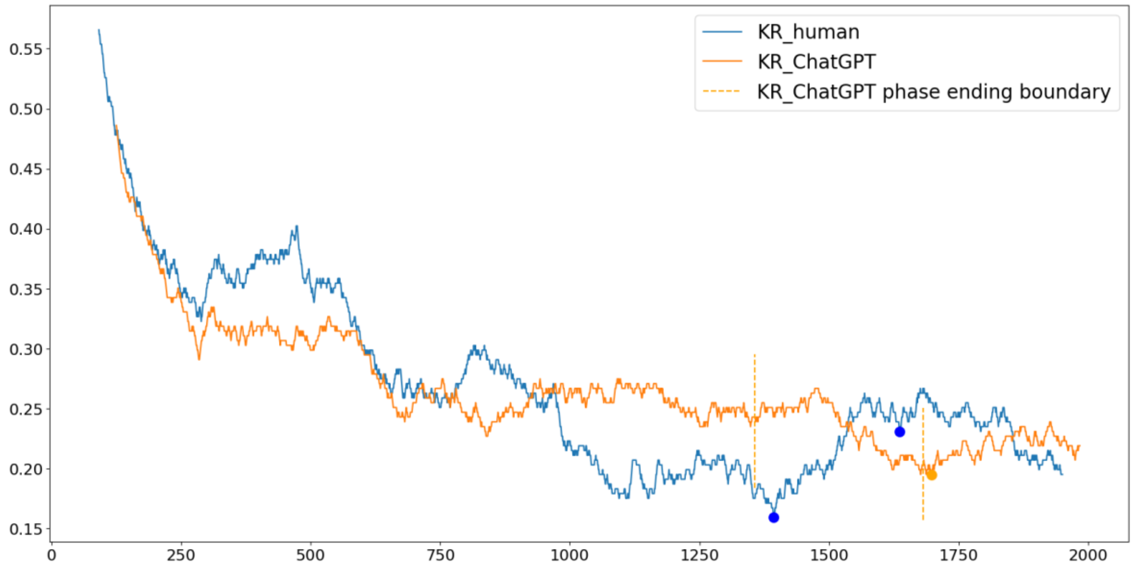

Figure 5. Comparison of KR HT TTR and ChatGPT TTR

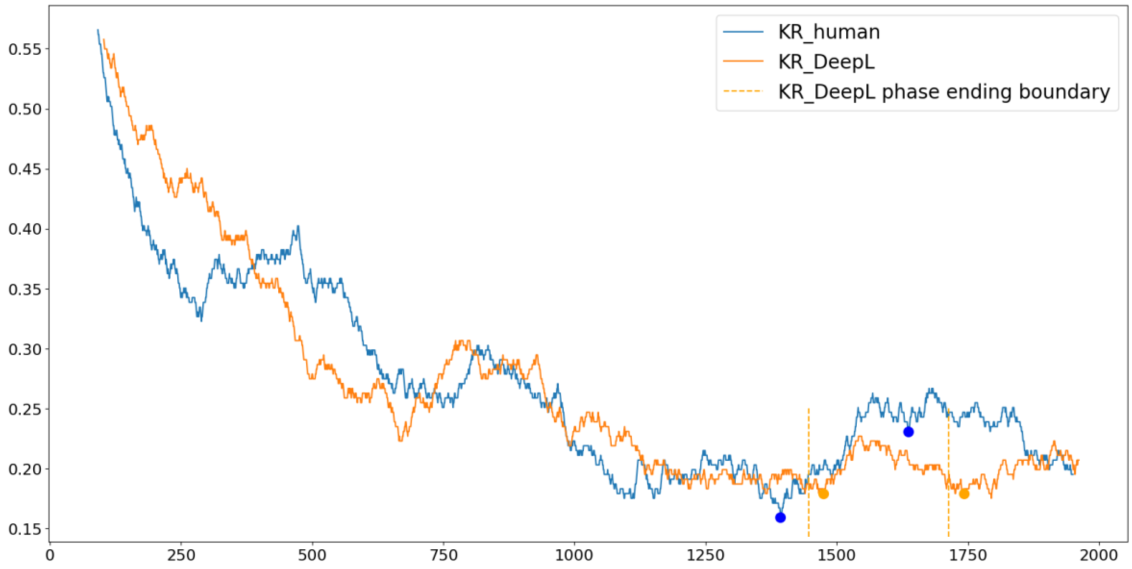

Figure 6. Comparison of KR HT TTR and DeepL TTR

4.2.2 Comparison between the Chinese HT and AI-based MT systems

In this section, Chinese HTs served as the standard benchmark for comparison with MT systems. The analysis procedure was similar to that for Korean translation. The TTR curves for each MT system and the Chinese HT were shown in Figures 7, 8, and 9. The result of the DTW analysis for the three MT systems was shown in Table 2 below. The results were further compared with those obtained for Korean MT translation in the end of this section.

In Figure 7, the Google Translate TTR curve exhibited marked increases starting from the two lowest points, which indicated a clear detection of significant boundaries for the three phases of the story. In addition, it is

© 2024 KASELL All rights reserved

noteworthy that almost the entire Google TTR line is below that of the HT. In Figure 8, the ChatGPT TTR curve also displayed two significant boundaries, with peaks closely resembling those of the HT TTR, representing the last two new episodes. Furthermore, similar to the result of the Korean translation, most parts of the ChatGPT TTR line were lower than that of the HT. In Figure 9, the DeepL TTR curve clearly revealed the lowest point which distinguishes the first phase from the other two and the following two peaks near the end of the story were also present. Still, for almost every part, the DeepL line is below that of the HT.

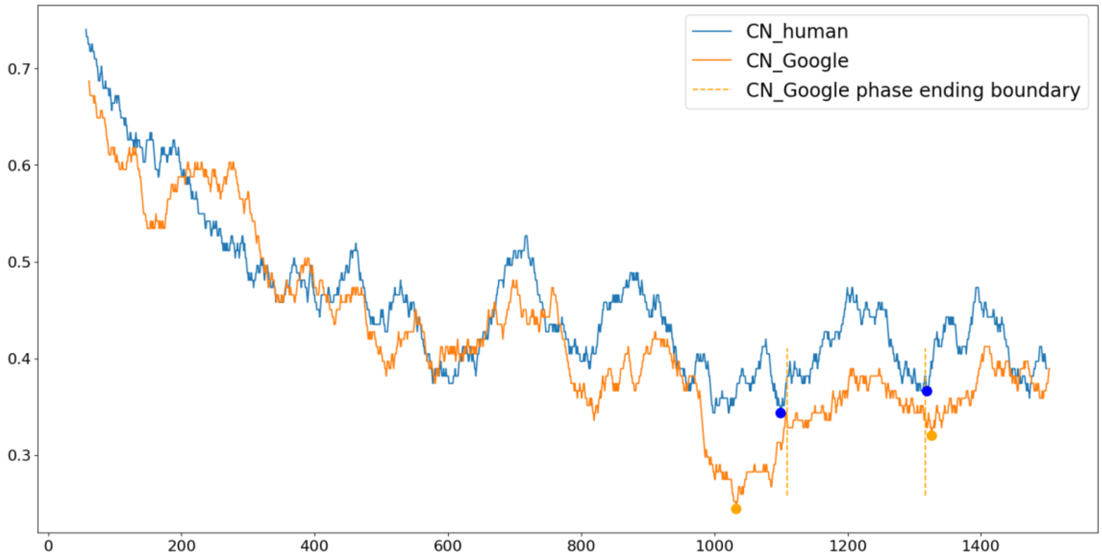

Figure 7. Comparison of Chinese TTR Graph of HT with That of Google Translate

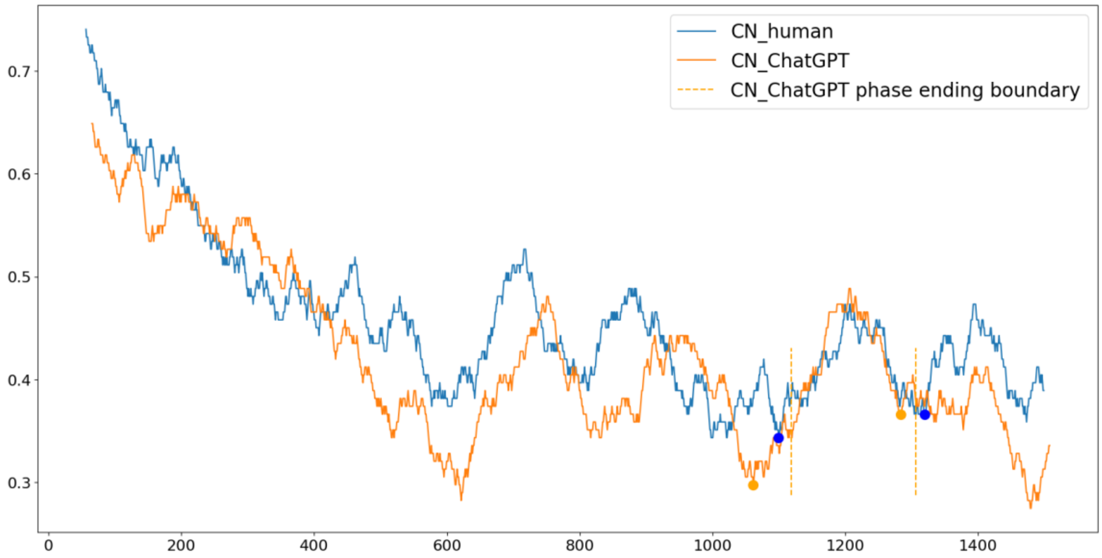

Figure 8. Comparison of Chinese TTR Graph of HT with That of ChatGPT

© 2024 KASELL All rights reserved

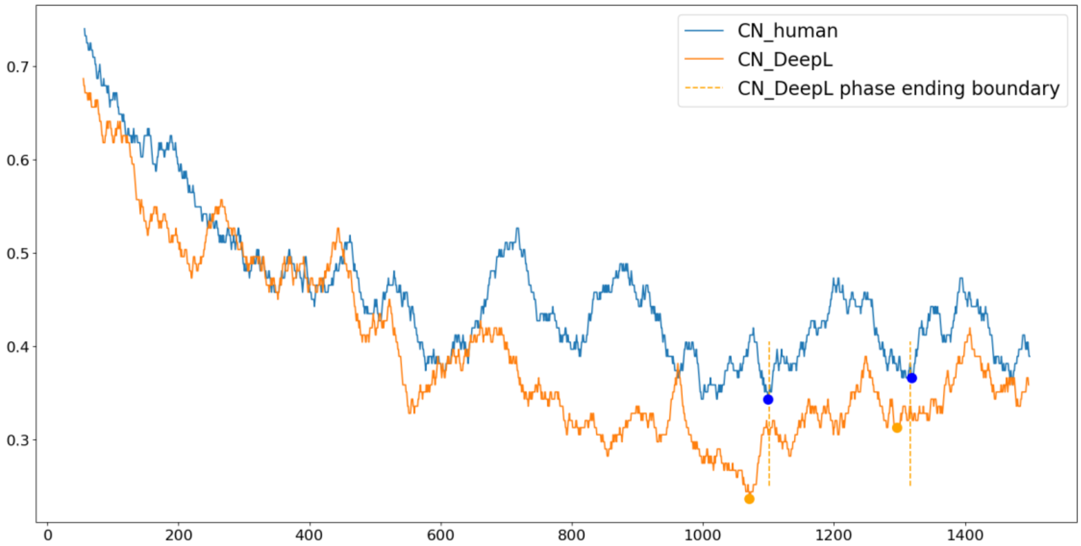

Figure 9. Comparison of Chinese TTR Graph of HT with That of DeepL

The results for each phase and the overall weighted average DTW similarity score between Chinese HT and MT systems were shown in Table 2. ChatGPT obtained the highest average score of 0.621, followed by Google Translate with an average score of 0.621 and DeepL with a score of 0.469. In other words, for translation from English to Chinese, it seems that the quality of ChatGPT ranks the best, followed by Google Translate and DeepL.

Table 2. DTW Similarity between Chinese HT and MT Systems

Chinese	Google Translate	ChatGPT	DeepL
Part 1	0.570	0.584	0.393
Part 2	0.574	0.882	0.405
Part 3	0.887	0.718	0.885
Average (Ranking)	0.621 (2nd)	0.642 (1st)	0.469 (3rd)

We further compared the results of the above three dimensions for Korean and Chinese MT translations shown in Table 3. Text structure is marked from 0 (no trace of story phases) to 3 (three complete phases) to reveal how many structures were detected in the corresponding TTR graph; similarity ranking is recorded from 1(most similar) to 3 (least similar). TTR level is recorded as ‘below’ (the MT TTR curve is generally below the HT TTR line) or ‘above’ (the MT TTR curve is generally above the HT TTR line) to show its relationship to the HT.

Three interesting points emerge from Table 3. First, in terms of text structure, the TTR graphs show that the three machine-translated texts tend to be able to reveal the information flow of the original story, but with different levels of quality. For EN-CN, all three MT systems show a similar level of performance, clearly including the three phases of the story congruent with the original English text. However, for EN-KR, the text translated by Google Translate retains three story phases, whereas the text translated by ChatGPT and DeepL has only two less distinctive story phases, fewer than the original text. Second, based on the “similarity” comparison, the three MT systems have completely different behavior with respect to EN-KR and EN-CN. For EN-KR, DeepL performs best, followed by Google Translate and ChatGPT. In contrast, when it comes to EN-CN, DeepL behaves worst, and instead ChatGPT behaves best, Google Translate still ranks at Second. Third, for both EN-KR and EN-CN, the TTR level of the output generated by the three MT systems is below that of HT.

© 2024 KASELL All rights reserved

Table 3. Discrepancies of MT Systems between EN-CN and EN-KR

Google Translate	Text structure	Similarity ranking	TTR level
KR	3	2nd	below
CN	3	2nd	below
ChatGPT	Text structure	Similarity ranking	TTR level
KR	2	3rd	below
CN	3	1st	below
DeepL	Text structure	Similarity ranking	TTR level
KR	2	1st	below
CN	3	3rd	below

The results have some implications for some hypotheses in the context of MT research. First, the lower level of MT TTR indicates that the lexical diversity of MT output is lower than HT, which is consistent with the hypothesis proposed by Vanmassenhove et al. (2019) that the process of MT causes a general loss in terms of lexical diversity and richness compared to human-generated text. This could be related to the disappearance (or 'non-appearance') of rare words due to the inherent nature of MT systems. Furthermore, combining our findings of lower lexical diversity and reduced segments compared to the original English text, it seems reasonable to argue that information density is likely to be affected in the machine translated text. This result provides some evidence for the hypothesis proposed by Rubino et al. (2016) that "simplification and explication may affect the average information density measured in translated texts compared to comparable originally written texts in the same language".

Based on the comprehensive TTR analysis in section 4.2, it seems reasonable to claim that the TTR graphs are sensitive to the quality discrepancies of different MT systems with respect to both Korean and translations. However, we are cautious in concluding that this TTR-based method is valid for measuring the discrepancy in translation quality of different AI-based MT systems. In the following section, the discrepancies in translation quality are measured based on automatic evaluation metrics.

4.3 Results of Automatic Evaluation on MT Systems

In this section, we analyze scores from three AI-based MT systems using BLEU, METEOR, and ROUGE metrics, as shown in Table 4 and Table 5.

Table 4 showed that from an overall perspective, DeepL performed best in the translation of the short story Eveline from English to Korean, followed by Google Translate and ChatGPT 3.5. In contrast, Table 5 revealed that DeepL gained the worst scores in the translation from English to Chinese, and meanwhile ChatGPT 3.5 performed best, followed by Google Translate. This result suggests that the translation quality of a particular MT system seems related to certain language pairs (EN-KR / EN-CN).

Table 4. Evaluation Scores of Korean Outputs

Korean	Google Translate	ChatGPT 3.5	DeepL
BLEU	0.2167 (3rd)	0.2213 (2nd)	0.2402 (1st)
METEOR	0.3395 (2nd)	0.3176 (3rd)	0.3549 (1st)
ROUGE	0.4347 (2nd)	0.3965 (3rd)	0.4438 (1st)
Ranking	2nd	3rd	1st

© 2024 KASELL All rights reserved

Table 5. Evaluation Scores of Chinese Outputs

Chinese	Google Translate	ChatGPT 3.5	DeepL
BLEU	0.1873 (1st)	0.1795 (2nd)	0.1790 (3rd)
METEOR	0.3523 (2nd)	0.3606 (1st)	0.3444 (3rd)
ROUGE	0.4971 (2nd)	0.5156 (1st)	0.4767 (3rd)
Ranking	2nd	1st	3rd

The results of the three automatic evaluation metrics align with the TTR analysis. Since the three metrics evaluate the quality of MT from different dimensions, the consistent result provides reliable evidence for the validity of the separately counted TTR method in demonstrating the quality discrepancies of different MT systems.

5. Discussion

The exploration of the adapted TTR approach, a traditional corpus linguistic technique, flexibly applied to the domain of MT has innovative meaning in the AI date. And the underlying idea of the information flow revealed by lexical features in discourse has been used for various purposes such as analyzing the style and competence of the authors in literature, checking or providing evidence for the literal critics. The present study moves a step forward by shedding light on the potential of this linguistic knowledge being employed to the domain of MT. The prevalent evaluation metrics for MT mainly focused on the accuracy of the outputs, which essentially match n-grams or relevant of n-grams of the reference and the candidate. The present study extends the scope of evaluation by comparing the coherence of the translations on a discourse spectrum rather than limited to sentence level. Our analytical framework, based on span-based TTR calculations, mitigates the need for strict sentence matching across the English, Korean, and Chinese texts. The selection of spans allows us to assess translation quality and lexical diversity without the constraints imposed by direct sentence alignment. This methodology acknowledges and accommodates the natural variability in sentence structure and length across languages, focusing instead on overarching lexical and structural trends observable within contiguous text segments. Additionally, since the TTR graph visualizes the semantic structures and reveals of the style and competence of the original authors, it implies the method might contributes to the analysis of style and competence alignments between the translated texts and source texts. The future study on 'machine translationese' will not only on lexical and information features but also on style and even cultural properties. To be further, the present study also sheds light on the plausibility of investigation on the style and competence (Youmans 1990) of narrative authors of other languages. One of the findings is that the TTR graphs are sensitive to different languages. Chinese and Korean reveals different overall TTR landscapes, with Chinese graph showing overall higher levels and shorter lines. It implies that the future study, on the one hand, can investigate stories of different length to compare the styles and competence of different authors within a particular language. On the other hand, linguistic features of narratives regarding different languages can be investigated.

MT has contributed to L2 learning from various aspects including L2 writing and translation practice. The post editing strategies and translation errors have been involved in the studies on application of MT to teaching. The novelty quantitative method provides new possible translation errors to be considered when doing post-editing (Shin and Chon 2023). Specifically, the error analysis (Costa et al. 2015, Ferris 2011, Lee and Briggs 2021) on discourse coherence of the translated text detected using the proposed TTR method enables the consideration of

© 2024 KASELL All rights reserved

translation errors to be extended beyond words, phrases and sentences to discourse, or rather episodes matching in translation for narratives.

However, we recognize the limitations of our current methodology, particularly the challenges of quantitatively comparing MT and HT using a span-based rather than commonly-recognized sentence-based approach. However, the advancement of the AI-based MT systems requires the development of new method to evaluate the nuanced discrepancies of MT outputs from the source language. Our methodology is designed to explore lexical diversity and information flow across broader text segments. If, in the further study, this approach is improved such as by combining qualitative error analysis in translation, it will deliberately capture more nuanced aspects of translation quality that might be overlooked in a simpler sentence-by-sentence analysis. Additionally, the present study focuses on a single translation direction of only two target languages (from English to Korean and Chinese). A study includes backward translation directions among comparable datasets, along with the qualitative analysis will offer deeper insights into the variances across different MT systems and between HT and MT. Such investigations will further enrich the theory on systemic machine translationese, providing stronger evidence to support discussions on this topic.

6. Conclusion

Three conclusions can be drawn from the quantitative analysis, and some implications for the future study are revealed. First, the adapted TTR computed separately within a moving span has the power to detect the significant boundaries of a translated text in both Korean and Chinese. However, this is now limited to the narrative genre, and the story is well edited using professional narrative techniques. Further research on different genres or other narrative stories will shed more light on the validity of this quantitative measure. In addition, for both Korean and Chinese, using a suitable span, the TTR graphs are highly valid in showing the discrepancies in translation quality among the three different MT systems. This claim is supported not only by TTR analysis per se, but also by the comprehensive evaluation metrics, which show exactly the same ranking of the three different MT systems in terms of comparison within either EN-KR or EN-CN. Finally, the different performance of the three MT systems with respect to Korean and Chinese shows that the three AI-based MT systems are sensitive to different language pairs.

This study not only offers insights relevant to the existing framework of MT theory, as discussed in Section 4.2.2, but also enriches second language learning by introducing new evaluative perspectives at the discourse level for MTs (detailed in section 5). Regarding future research, the current findings, derived from a small, self-compiled multilingual corpus focusing on translations from English to Korean and Chinese, suggest that further studies utilizing larger corpora encompassing a broader range of languages and genres could yield valuable insights. Furthermore, combining this quantitative method with qualitative analysis of crucial segments—specifically, those between the lowest points and the areas identified by literary critics as boundaries—could offer profound insights into the stylistic and cultural characteristics of machine-translated outputs or human translations.

© 2024 KASELL All rights reserved

References

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis and E. Tognini-Bonelli, eds., *Text and Technology*, 233-250. Amsterdam: John Benjamins.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C. Lin and C. Voss, eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65-72. Michigan: Association for Computational Linguistics.

Bentivogli, L., A. Bisazza, M. Cettolo and M. Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In J. Su, K. Duh and X. Carreras, eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 257-267. Austin: Association for Computational Linguistics.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brglez, M. and Š. Vintar. 2022. Lexical diversity in statistical and neural machine translation. *Information* 13(2), 93-107.

Broeder, P., G. Extra and R. V. Hout. 1993. Richness and variety in the developing lexicon. In C. Perdue, ed., *Adult Language Acquisition: Cross-linguistic Perspectives. Vol. I: Field Methods*, 145-163. Cambridge: Cambridge University Press.

Castilho, S., N. Resende and R. Mitkov. 2019. What influences the features of post-editese? A preliminary study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 19-27. Shoumen, Bulgaria: Incoma Ltd.

Chafe, W. 1987. Cognitive constraints on information flow. In R. Tomlin, ed., *Coherence and Grounding in Discourse*, 21-51. Amsterdam: John Benjamins.

Chafe, W. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

Costa, Â ., W. Ling, T. Luís, R. Correia and L. Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation* 29(2), 127-161.

Engber, C. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of L2 Writing* 4(2), 138-155.

Ferris, D. 2011. *Treatment of Error in Second Language Student Writing*. Michigan: University of Michigan Press.

Francis, W. N. and H. Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Hart, C. 1969. Eveline. In C. Hart, ed., *James Joyce's 'Dubliners'*, 48-52. London: Faber and Faber.

Ilisei, I.-N. 2012. *A Machine Learning Approach to the Identification of Translation Language: An Inquiry into Translationese Learning Models.* Doctoral Dissertation, University of Wolverhampton, England, UK.

Kruger, H. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target. International Journal of Translation Studies* 24(2), 355-388.

Laufer, B. and P. Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16(3), 307-322.

Lee, S. M. and N. Briggs. 2021. Effects of using machine translation to mediate the revision process of Korean university students' academic writing, *ReCALL*, 33(1), 18–33.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Workshop Text Summarization Branches Out*, 74-81. Barcelona: Association for Computational Linguistics.

Malvern, D., B. Richards, N. Chipere and P. Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Basingstoke: Palgrave Macmillan.

© 2024 KASELL All rights reserved

McCarthy, P. M. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Doctoral dissertation, The University of Memphis, Memphis, TN, USA.

O'Halloran, K. 2007. The subconscious in James Joyce's 'Eveline': A corpus stylistic analysis that chews on the 'Fish hook'. *Language and Literature* 16(3), 227-244.

O'Loughlin, K. 1995. Lexical density in candidate output on direct and semi-indirect versions of an oral proficiency test. *Language Testing* 12(2), 217-237.

Papineni, K., S. Roukos, T. Ward and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak and D. Lin, eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318. Association for Computational Linguistics.

Roberts, N., D. Liang, G. Neubig and Z. C. Lipton. 2020. Decoding and diversity in machine translation. *arXiv:2011.13477 [cs.CL]*.

Rubino, R., E. Lapshinova-Koltunski and J. Genabith. 2016. Information density and quality estimation features as translationese Indicators for human translation classification. In K. Knight, A. Nenkova and O. Rambow, eds., *Proceedings of NAACL-HLT 2016*, 960-970. Association for Computational Linguistics.

Sakoe, H. and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26(1), 43-49.

Shin, D. and Y. V. Chon. 2023. Second language learners' post-editing strategies for machine translation errors. *Language Learning & Technology* 27(1), 1-25.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Massachusetts: Blackwell.

Teich, E. 2003. *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Toral, A. 2019. Post-editese: An exacerbated translationese. In M. Forcada, A. Way, B. Haddow and R. Sennrich, eds., *Proceedings of Machine Translation Summit XVII: Research Track*, 273-281. Dublin: European Association for Machine Translation.

Tuldava, J. 1998. *Probleme und Methoden der Quantitativ-systermischen Lexikologie*. [Translated from Russian original (1987)]. Verlag: Wissenschaftlicher Verlag Trier.

Vanmassenhove, E., D. Shterionov and A. Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In M. Forcada, A. Way, B. Haddow and R. Sennrich, eds., *Proceedings of Machine Translation Summit XVII: Research Track*, 222-232. Dublin: European Association for Machine Translation.

Vanmassenhove, E., D. Shterionov and M. Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In P. Merlo, J. Tiedemann and R. Tsarfaty, eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2203-2213. Association for Computational Linguistics.

Vermeer, A. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing* 17(1), 65-84.

Youmans, G. 1991. A new tool for discourse analysis: The vocabulary management profile. *Language* 67(4), 763-789.

Yule, G. U. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Examples in: English
Applicable Languages: English
Applicable Level: All

© 2024 KASELL All rights reserved