applied sciences

MDPI

Article

Automated Scoring of Translations with BERT Models: Chinese and English Language Case Study

Yizhuo Cui [1] and Maocheng Liang [2,*]

[1] School of Humanities and Law, North China University of Technology, Beijing 100144, China; cuiyz@ncut.edu.cn
[2] School of Foreign Languages, Beihang University, Beijing 100191, China
* Correspondence: frankliang0086@163.com

Abstract: With the wide application of artificial intelligence represented by deep learning in natural language-processing tasks, the automated scoring of translations has also advanced and improved. This study aims to determine if the BERT-assist system can reliably assess translation quality and identify high-quality translations for potential recognition. It takes the Han Suyin International Translation Contest as a case study, which is a large-scale and influential translation contest in China, with a history of over 30 years. The experimental results show that the BERT-assist system is a reliable second rater for massive translations in terms of translation quality, as it can effectively sift out high-quality translations with a reliability of $r = 0.9$ or higher. Thus, the automated translation scoring system based on BERT can satisfactorily predict the ranking of translations according to translation quality and sift out high-quality translations potentially shortlisted for prizes.

Keywords: large language model; BERT; automated scoring of translations; large-scale translation contest

check for updates

Citation: Cui, Y.; Liang, M. Automated Scoring of Translations with BERT Models: Chinese and English Language Case Study. *Appl. Sci.* **2024**, *14*, 1925. https://doi.org/10.3390/app14051925

Academic Editor: Marco Palomino

Received: 31 January 2024
Revised: 20 February 2024
Accepted: 22 February 2024
Published: 26 February 2024

1. Introduction

In recent years, large language models, powered by advanced deep learning techniques, have revolutionized the field of natural language processing (NLP) and found extensive applications across various domains. These models, often pre-trained on massive datasets, have demonstrated remarkable capabilities in a wide range of NLP tasks, and they can be fine-tuned to improve performance on various downstream tasks, such as sentence similarity [1–3], translation [4–6], text classification [7–10], question answering [11–13], summarization [14–17], etc. Additionally, the emergence of Big Data provides opportunities to leverage large amounts of data for building high-quality models and improving the performance of NLP systems.

This study explores the possibility of developing a system for the automated scoring of translations for large-scale translation contests based on the large language model, the Bidirectional Encoder Representations from Transformers (BERT). The scoring of translations in large-scale translation contests suffers from challenges related to the huge growth in the number of participants. In the 31st Han Suyin International Translation Contest, the total number of valid Chinese-to-English (C–E) and English-to-Chinese (E–C) translations exceeded 10,000, including 3822 C–E submissions and 6825 E–C submissions. As the assessment of translations is a time-consuming and labor-intensive process, it significantly increases the costs in terms of time, manpower, and money, all of which greatly increase the burden on the organizing committee.

The automated scoring of translations for the large-scale translation contest, Han Suyin International Translation Contest, focuses on selecting high-quality translations for awards, rather than assigning scores to translations like those assigned in an examination. Therefore, a good approach is expected to satisfactorily predict the ranking of translations on an independent basis according to translation quality and sift out high-quality

translations potentially shortlisted for prizes. The selective nature of translation contests makes automated scoring possible. This paper explores the possibility of applying BERT to the automated scoring of translations for the large-scale translation contest with a view to addressing the problems mentioned above. Spearman's rank correlation (r) between the expected ranking (actual ranking) and the predicted ranking (automated assessment conducted by the system), mean absolute error rate of ranking and maximum absolute error rate of ranking were used for model evaluation. Accuracy, recall and F-measure were used for the reliability analysis.

This paper has two main contributions. First, it contributes to the field by providing empirical evidence of the viability of automated translation scoring systems based on BERT models focusing specifically on the evaluation of translations in the context of a well-established translation contest, thereby demonstrating real-world applicability and relevance. Second, it establishes effective methods for processing lengthy texts to conform with BERT's 512-token input constraint.

The remainder of this article is organized as follows. Section 2 shows an overview of previous research related to the topic. Section 3 provides a detailed exposition of the research methodology applied in this study, encompassing the corpus, BERT models, and research procedures. Sections 4 and 5 discuss the research findings in depth and engage in subsequent discussions on BERT in the automated scoring of C–E and E–C translations. The concluding remarks are summarized in Section 6.

2. Related Work

This section provides a comprehensive review of pertinent studies, focusing on a key concept—semantic similarity, and exploring approaches based on similarity for the assessment of translation.

2.1. A Key Concept: Semantic Similarity

Semantic similarity, as an important concept, is introduced in this study, as translation is a kind of restricted writing that is expected to semantically express the same content with the original text in a new language. Namely, if a translated text is semantically similar to the original text, it can be seen as a good translation.

Semantic similarity is a method to compute the semantic distance between two concepts according to a given ontology [18], and it relates to computing the similarity between concepts that are not lexically similar [19]. It is based on the likeness of semantic meaning, regardless of graphical and lexical similarity. It yields high values for pairs of words in a semantic relation (synonyms, hyponyms, free associations, etc.) and low values for all other, unrelated pairs [20]. For example, while the English words "CAR" and "AUTOMOBILE", look different in form, they have a quite high degree of semantic similarity because they share common features at the semantic level. However, the words "LEAD" and "HEAD" have a low degree of semantic similarity even though they look similar and share 75% of the same letters in the same positions. It seems to be a good solution to the problem that the reference translations provided by the experts cannot include all acceptable translations. To be specific, if the reference translation provides "CAR", and the word "AUTOMOBILE" then appears in a translated text, this can be well identified as a kind of expression highly similar to "CAR". By introducing semantic similarity, the assessment of translation quality can be transformed into a problem of how to calculate the degree of overall semantic similarity between a participant's translation and several expert translations. The higher the value of semantic similarity, the higher the quality of the translation. Conversely, if the value of semantic similarity is low, it indicates the lower quality of a translation.

2.2. Similarity-Based Approaches

Similarity-based approaches transform the task of assessment into similarity calculations. The assumption behind this method is prosed by Papineni et al. [21], "The closer a machine translation is to a professional human translation, the better it is". To be specific,

for a machine translation, the degree of similarity between it and one or more reference human translations determines its translation quality. There are a number of measures to calculate the degree of similarity, all of which are expected to achieve the same goal, which is to generate a numerical metric as the result. In terms of assessment measures, there are three main measures: Levenshtein distance-based, n-gram co-occurrence-based, and vector-based.

The proposal of distance measure approach is based on Levenshtein distance calculations [22], a string comparison metric that counts the number of edit operations (insertion, deletion, replacement and swap) required to transform one string into another [23]. As a sensitive measure with which distances between strings (in this case transcriptions of word pronunciations) can be calculated [24], it is an efficient and reliable method for computing string alignments [25]. The number of operations required is inversely proportional to the quality of the translation. In 1992, Su et al. [26] firstly employed it to assess machine translation quality with one reference human translation so that the assessment process could be done quickly and automatically. Subsequently, Nießen et al. [27] and Akiba et al. [28] tried to use multiple reference human translations in a dataset. Continuously upgraded by introducing advanced models like bag-of-words [29], the accuracy of scoring has been improved. Various measures for the automated assessment of machine translation quality, such as WER [29], TER [30], CDer [31], etc., are based on Levenshtein distance.

The n-gram co-occurrence statistics method is an innovative assessment method for machine translation. The assessment system, developed on the basis of n-grams co-occurrence statistics, is still widely used in various assessment tasks in the field of NLP. The core of this method is that "a good translation will have a distribution of n-grams similar to other good translations" [32]. To be specific, the segments in the tested translation text are compared with the corresponding aligned segments in the reference translation texts based on the number of their n-grams. The number of all the compared segments finally enter the final score calculation with various weights and factors. The main representative systems are BLEU [21], NIST [33] and METEOR [34].

For the similarity-based approaches mentioned above, the assessment reliability largely depends on the set of reference human translations. If the computer-generated translation is similar to one or more reference translations, it will be considered a good translation. While translations with a high degree of similarity are generally of high quality and have relatively reliable scoring accuracy, translations with low similarity are not necessarily bad translations, as reference translations cannot encompass all acceptable translations. This means that the matching rate of synonyms may be low. A limited number of reference translations will likely reduce the reliability of the assessment. Moreover, the process requires a significant amount of manual intervention, as invited experts must contribute to building the set of reference human translations.

Semantic similarity between two long texts can also be achieved by vectorizing them and then comparing the values of the two vectors generated by large language models. The vectorization of words can technically be achieved by word embedding programs such as Word2Vec [35,36]. The linguistic theoretical basis comes from Distributional Hypothesis [37], that is, words that occur in the same contexts tend to have similar meanings. Based on Word2Vec, Doc2Vec upgraded semantic similarity calculation to a document level [38]. BERT is designed to pre-train deep bidirectional representations from unlabeled text by conditioning on both the left and right context in all layers [39]. This allows it to obtain context-aware embeddings for words and documents as it captures intricate semantic relationships and contextual information. A vector space model can be built and used to calculate semantic similarity after a large amount of corpus training, which can accurately identify the translations semantically similar to the expert translations and is not restricted by the languages and genres of the texts. The overall semantic similarity between two

texts can be derived by computing the cosine value between their document vectors. For instance, the cosine value of text A and text B can be calculated by Equation (1):

$$Similarity(A, B) = cos\theta = \frac{A \times B}{\|A\| \times \|B\|} \quad (1)$$

where $\|A\|$ and $\|B\|$ are the Euclidean norms (or magnitudes) of the vectors A and B, respectively.

This study endeavors to employ the vector-based approach utilizing the large language models, the BERT series, for the automated evaluation of translations in a large-scale translation contest. BERT models offer a unique combination of contextual understanding, bidirectional encoding, and pre-training on large corpora, which collectively enhance the accuracy and effectiveness of semantic similarity-based approaches for translation scoring. This saves time, energy, knowledge, and resources that would otherwise be required to train a language-processing model from scratch. It fills a gap in the field by demonstrating the effectiveness of BERT models in automating the scoring process for large-scale translation contests, which traditionally rely on manual assessment.

3. Materials and Methods

This section introduces some issues related to the methodology employed in this study, including the corpus, BERT models, and research procedures.

3.1. The Corpus

The corpus used in this study consists of participant and expert translations from the 31st Han Suyin International Translation Contest. The contest has a long history and is considered the most prestigious translation contest in China. It has inspired many young people to learn translation and has provided highly trained translators for various industries and has contributed to the development of translation teaching in China.

The data used in this research consist of participant translations and expert translations from the 31st Contest. As shown in Table 1, the sub-corpus used in this study consists of two parts: C–E translations and E–C translations. The C–E set includes 3822 participant translations and 4 expert translations, while the E–C set includes 6825 participant translations and 4 expert translations. The expert translations are provided by senior translation experts with extensive experience in Chinese and English translation.

Table 1. The corpus.

Group	Number of Expert Translations	Number of Valid Translations
C–E Translations	4	3822
E–C Translations	4	6825

3.2. BERT Models

This paper employs different BERT models tailored to the system design within the domain of English-to-Chinese and Chinese-to-English translation.

For C–E translations, BERT models were tested in several variations, including bert-base-uncased, bert-large-uncased, bert-base-cased, and bert-large-cased, which were originally released for cased and uncased input text. The differences between these models primarily lie in their size (number of parameters) and whether they are case-sensitive or case-insensitive during tokenization.

For E–C translation, the bert-base-chinese model was used for E–C translations. The model is pre-trained on a massive amount of Chinese text data, learning to predict masked words within sentences and understand the relationships between words in context.

As the input length of BERT is limited to 512 tokens, it becomes a challenge to apply it to long texts in this research, where the texts to be translated are significantly longer than 512 tokens. Therefore, there is a need to preprocess the input text in a way that it can

be truncated or split into smaller segments of 512 tokens or less, while still preserving its semantic meaning.

There are two common pre-processing strategies that can be used to limit the input text to 512 tokens. The first approach is text truncation, which involves cutting off the text after the first 512 tokens and dropping the rest. However, this can result in the loss of valuable information and the resulting vector may not fully represent the global information of the entire text. The second strategy is to use summarization to condense the translated text into a summary of no more than 512 tokens. This limits the length of the text to what BERT can handle. The summary text captures the main information of the original text, and only the information that is not critical to the text is lost. Therefore, the vector of the summary text can generally be seen as representing the semantic content of the entire text. However, this also fails to capture the global information of the entire text. Thus, we propose to apply the method of segmentation in this study, which involves comparing the mean embeddings of all the paragraphs in a text as a whole, where all the paragraphs are within 512 tokens. After pre-processing, the BERT model, acting as a text encoder, outputs a vector that captures semantic information from the input text.

3.3. Procedures

The NLP tool used in the study is Python. The process for the implementation of this study involves four main steps, namely, synthetic data generation, model evaluation, scoring, and reliability analysis.

3.3.1. Synthetic Data Generation

In the first place, we perform synthetic data generation. Due to insufficient number of translations rated by experts, a "Text-to-Text" synthesis method was utilized to create synthetic texts based on the four expert translations, in order to obtain a sufficient number of translations with known expected ranking. This step was the data preparation for model evaluation, and the texts generated were used for model evaluation to enable an efficient and quick understanding of model quality.

As shown in Table 2, 100 synthetic texts could be generated for each expert translation. The first synthetic text contained the first 1% of the expert translation, was assigned a score of 1 and ranked 100th. The second synthetic text contained the first 2% of the expert translation, which scored 2 and ranked 99th. This process was repeated, generating four sets of synthetic texts with four known expected ranking lists based on the four expert translations.

Table 2. Synthetic texts generated by an expert translation.

Synthetic Texts	Score	Rank
1% Expert translation text	1	100
2% Expert translation text	2	99
3% Expert translation text	3	98
...	...	...
99% Expert translation text	99	2
100% Expert translation text	100	1

3.3.2. Model Evaluation

After generating enough translations with known expected rankings, a model evaluation is necessary to ensure that the ranks generated by the system are reliable. This evaluation process includes analysis of the results of model evaluation indicators.

- Indicators for model evaluation

Given that the system built in this study is based on "picking out the good translations and leaving the bad ones", the following three indicators were used to evaluate the models: (1) Spearman's rank correlation between the expected ranking (actual ranking) and the

predicted ranking (automated assessment done by the system); (2) mean absolute error rate of ranking; and (3) maximum absolute error rate of ranking.

1. Spearman's Rank Correlation Coefficient

According to Weigle [40], the correlation coefficient is a standard measure for inter-rater reliability between Rater 1 and Rater 2, which can be computed using the popular Equation (2):

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

where r represents Spearman's rank correlation coefficient, d_i represents difference between the two ranks of each observation, n represents number of observations.

In this study, the expected ranking is considered to have been done by Rater 1, while the predicted ranking was considered to have been done by Rater 2. As ranking is the focus, Spearman's rank correlation was used for the analysis since it effectively tests the association between two ranked variables. The Spearman's rank correlation coefficient, r, ranges from -1 to +1. A r of +1 indicates a perfect association of ranks, a r of zero indicates no association between ranks, and a r of -1 indicates a perfect negative association of ranks. The closer r is to zero, the weaker the association between the ranked variables. Thus, if r is close to +1, it indicates a strong positive correlation between the expected ranking and the predicted ranking, implying that the system is a highly reliable rater. Conversely, if r is low, the system is deemed to be of low reliability.

Then, the four sets of synthetic texts are assessed by the designed system, and four predicted ranking lists can be obtained. The overall reliability of the system is shown in Equation (3):

$$r_{final} = \sum_{i=1}^{4} r_i \quad (3)$$

where r_{final} is the average of the four Spearman's rank correlation coefficients (r) of the four predicted ranking lists generated by the designed system compared with their expected ranking lists.

2. Absolute error rate of ranking

The mean absolute error rate of ranking and the maximum absolute error rate of ranking are also included in the reliability analysis as they can reflect the prediction error rate of the automated assessment system in two dimensions. Let $D_{1,j}$ denotes the synthetic text generated with the first j% of "Expert Translation 1". The absolute ranking error, denoted by $|E_{1,j}|$, is defined as the absolute value of the difference between the expected ranking and the predicted ranking of $D_{1,j}$. The absolute ranking error rate, denoted by $P_{1,j}$, is calculated as $\frac{E_{1,j}}{100}$. For example, suppose $D_{1,1}$ is expected to rank 100th, but is ranked 98th by the automated assessment system. Then $|E_{1,1}|$ is equal to $|100 - 98| = 2$, and $P_{1,1}$ is equal to $\frac{|2|}{|100|} = 0.02$. The mean absolute error rate of ranking can be expressed as follows in Equation (4):

$$Mean(P) = \sum_{j=1}^{100}\sum_{i=1}^{4} P_{i,j}/4 \times 100 \quad (4)$$

where $P_{i,j}$ represents the absolute ranking error rate of synthetic text j generated on the basis of Expert Translation i. Mean (P) represents the average absolute ranking error rate of all synthetic texts generated in this study.

The mean absolute error rate of ranking is used in this study instead of the mean error rate of ranking. This is because the mean error rate can have both positive and negative error rate values, which may sometimes cancel each other out when performing algebraic summation. Consequently, the mean error rate may not provide an accurate indication of the final error rate. The mean absolute error rate, on the other hand, accurately reflects the

magnitude of the prediction error rate and overcomes the shortcomings of the mean error rate to a certain extent.

The maximum absolute ranking error rate can be calculated using the following Equation (5):

$$Max(P) = Max\begin{cases} P_{1,1}, P_{1,2}, \ldots, P_{1,j}, \ldots, P_{1,99}, P_{1,100} \\ P_{2,1}, P_{2,2}, \ldots, P_{2,j}, \ldots, P_{2,99}, P_{2,100} \\ P_{3,1}, P_{3,2}, \ldots, P_{3,j}, \ldots, P_{3,99}, P_{3,100} \\ P_{4,1}, P_{4,2}, \ldots, P_{4,j}, \ldots, P_{4,99}, P_{4,100} \end{cases} \quad (5)$$

where $P_{1,j}$, $P_{2,j}$, $P_{3,j}$ and $P_{4,j}$ are the absolute ranking error rates of the synthetic texts with the first j% of Expert Translation 1, Expert Translation 2, Expert Translation 3, and Expert Translation 4, respectively.

3.3.3. Scoring

After the model evaluation, the translation scoring was processed, which involved the use of the designed automated scoring systems to score the participant translations and rank them according to their semantic similarity compared with the expert translations. Each system achieved this by comparing each participant translation to all the expert translations and generating a translation quality ranking list. During this process, words, phrases, and sentences with the same or different expressions but that were semantically similar to any of the expert translations were effectively identified. In this study, all the expert translations were considered to have full marks, and any participant translation that was very similar to any of them was considered to be a good translation. The two sets of translations used in this study were the translations provided by the experts and the translations submitted by the participants, as shown below:

(1) expert translations = [$expert_1$, $expert_2$, $expert_3$, $expert_4$]
(2) participant translations = [$participant_1$, $participant_2$, …, $participant_n$].

First, a text-cleaning process was carried out using Python programming language with a custom-written code.

Next, the segmentation method was used to obtain the mean embeddings of all the paragraphs in a text as a whole. The final representation of the participant translations and expert translations can be obtained by calculating the mean embeddings of all its paragraphs, as shown in Equation (6):

$$e_{text} = \frac{\sum_{i=1}^{N} e_i}{N} \quad (6)$$

where e_{text} refers to the mean embeddings of a translated text with N paragraphs.

For each participant translation, the system compares it with each expert translation and selects the maximum cosine similarity measure out of the four comparisons as the final similarity score. This approach allows for a more comprehensive assessment of the participant translations, taking into account potential variations in the expert translations, as shown in Equation (7):

$$sim_1 = max\begin{cases} cos\_sim(e_{participant_1}, e_{expert_1}) \\ cos\_sim(e_{participant_1}, e_{expert_2}) \\ cos\_sim(e_{participant_1}, e_{expert_3}) \\ cos\_sim(e_{participant_1}, e_{expert_4}) \end{cases}, \ldots, sim_n = max\begin{cases} cos\_sim(e_{participant_n}, e_{expert_1}) \\ cos\_sim(e_{participant_n}, e_{expert_2}) \\ cos\_sim(e_{participant_n}, e_{expert_3}) \\ cos\_sim(e_{participant_n}, e_{expert_4}) \end{cases} \quad (7)$$

where *cos_sim* refers to the calculation of the cosine of the participant translation and the expert translation using Equation (1).

After obtaining the final similarity measures, a ranking list was generated based on these similarity measures as shown in Equation (8):

$$rank\ (sim_1,\ sim_2, \ldots, sim_n) \tag{8}$$

Finally, the top translations proceeded to the human raters.

Drawing from the BERT system's output, human raters can proceed with a second-round assessment of translations identified by the system as of relatively high quality. Human raters can provide additional context or insights that may not be captured by automated analysis alone, leading to more comprehensive and reliable evaluations for the final rewards. The integration of machine and manual scoring not only decreases evaluation time significantly but also conserves resources.

3.3.4. Measures of Reliability

The final step is the reliability analysis. Reliability entails a comprehensive comparison between scores assigned by the automated scoring systems and human raters based on accuracy, recall, and F0.5-Score. Given that automated scoring systems are designed to emulate human scoring, it is systematically evaluated against human scoring metrics. This evaluation serves as a benchmark for assessing the alignment and effectiveness of system scoring in replicating human scoring. A set of randomly selected translations, including 138 C–E participant translations and 262 E–C participant translations, which have been manually assessed, was used for the reliability analysis.

- Accuracy

Accuracy is a metric used to measure how well a model performs across all categories. It is calculated by dividing the number of correct predictions by the total number of predictions. Mathematically, it can be expressed as shown in Equation (9):

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}} \tag{9}$$

where $True_{positive}$ are the number of positive samples correctly identified by the system, $True_{negative}$ are the number of negative samples correctly rejected by the system, $False_{positive}$ are the number of negative samples incorrectly identified by the system, and $False_{negative}$ are the number of positive samples that incorrectly rejected by the system.

- Recall

Recall is a metric used to measure the ability of a system to correctly identify all positive samples. It is calculated as the ratio of the number of true positives to the sum of true positives and false negatives, as shown in Equation (10):

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}} \tag{10}$$

where $True_{positive}$ are the number of positive samples correctly identified by the system, and $False_{negative}$ are the number of positive samples that incorrectly rejected by the system.

- F-measure

F-measure is a metric that combines both precision and recall into a single measure. It is often used to assess the overall performance of a classification system. The $F0.5$-Score is a variant of the F-measure that places more emphasis on precision than recall, as shown in Equation (11):

$$F0.5 = (1 + 0.5^2)\frac{Precision \times Recall}{(0.5^2 \times Precision) + Recall} \tag{11}$$

Precision is calculated as the ratio of the number of positive samples correctly classified to the total number of samples classified as positive (correct or incorrect). Precision

measures the accuracy of the system in classifying samples as positive. The higher the value of precision, the fewer false positives are predicted. It can be calculated as shown in Equation (12):

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}} \quad (12)$$

where *Precision* and *Recall* are as defined previously, the *F*0.5-Score ranges between 0 and 1, with higher values indicating better performance. A score of 1 indicates perfect precision and recall, while a score of 0 indicates no correct predictions.

4. Results

In this section, the evaluation results of the automated scoring systems for C–E translations based on four English BERT models: bert-base-uncased, bert-base-cased, bert-large-uncased, and bert-large-cased, and the automated scoring systems for E–C translations based on the Chinese BERT model: bert-base-chinese, are reported.

4.1. C–E Translations

4.1.1. Results of Model Selection Indicators

Table 3 demonstrates the overall results of model selection indicators for the BERT-based system (C–E). In Table 3, the values of *r* are all greater than 0.99, with a corresponding $p < 2.2 \times 10^{-16}$, indicating a strong correlation between the expected ranking and the predicted ranking. For all the BERT models, the mean absolute error rates of ranking are all below 0.0229, and the maximum absolute error rates of ranking do not exceed 0.1400. These results suggest that the average and maximum fluctuations are under 2.29% and 14%, respectively, indicating that high-quality translations are not being filtered out.

Table 3. Results of model selection indicators for BERT-based systems (C–E).

Measures	Spearman's Rank Correlation Coefficient (*r*)	Significance (*p*)	Mean Absolute Error Rate of Ranking	Maximum Absolute Error Rate of Ranking
bert-base-uncased	0.996504650	$< 2.2 \times 10^{-16}$	0.0137	0.1100
bert-base-cased	0.998805875	$< 2.2 \times 10^{-16}$	0.0077	0.0700
bert-large-uncased	0.991977225	$< 2.2 \times 10^{-16}$	0.0229	0.1400
bert-large-cased	0.999039925	$< 2.2 \times 10^{-16}$	0.0066	0.0600

To be specific, there are four English BERT models: base and large variations, for both cased and uncased input text. The BERT (large cased)-based system outperforms the other models, with a correlation coefficient (*r*) of 0.999039925 and a corresponding $p < 2.2 \times 10^{-16}$. The mean absolute error rate of ranking is 0.0066, and the maximum absolute error rate of ranking is 0.0600. The BERT (large uncased)-based system performs relatively poorly in comparison to the other three systems, but still exhibits a correlation coefficient of over 0.99 with a corresponding $p < 2.2 \times 10^{-16}$. The mean absolute error rate of ranking is 0.0229, and the maximum absolute error rate of ranking is 0.1400. These results suggest that the BERT-based system is reliable enough to predict the ranking of C–E translations.

Additionally, the absolute error rate curves in Figures 1–4 reveal that the error rates are remarkably low.

Figure 1 illustrates the absolute error rate of ranking of synthetic texts generated based on the four expert translations in the BERT-based system (C–E, bert-base-uncased), with four sets of curves in (a–d). Overall, the figure shows a low absolute error rate, with a mean absolute error rate of ranking of 0.0137. Specifically, among the 400 synthetic texts, 49.5% of them have predicted rankings that match their expected rankings, resulting in an absolute error rate of 0. For 46% of the texts, the absolute error rates fall in the range of 0.01 to 0.05. Only 4.5% of the texts have absolute error rates of more than 0.05, with a maximum

of 0.11. These results indicate that the errors of the BERT (base uncased)-based system do not significantly impact the selection of the top-ranked C–E translations.

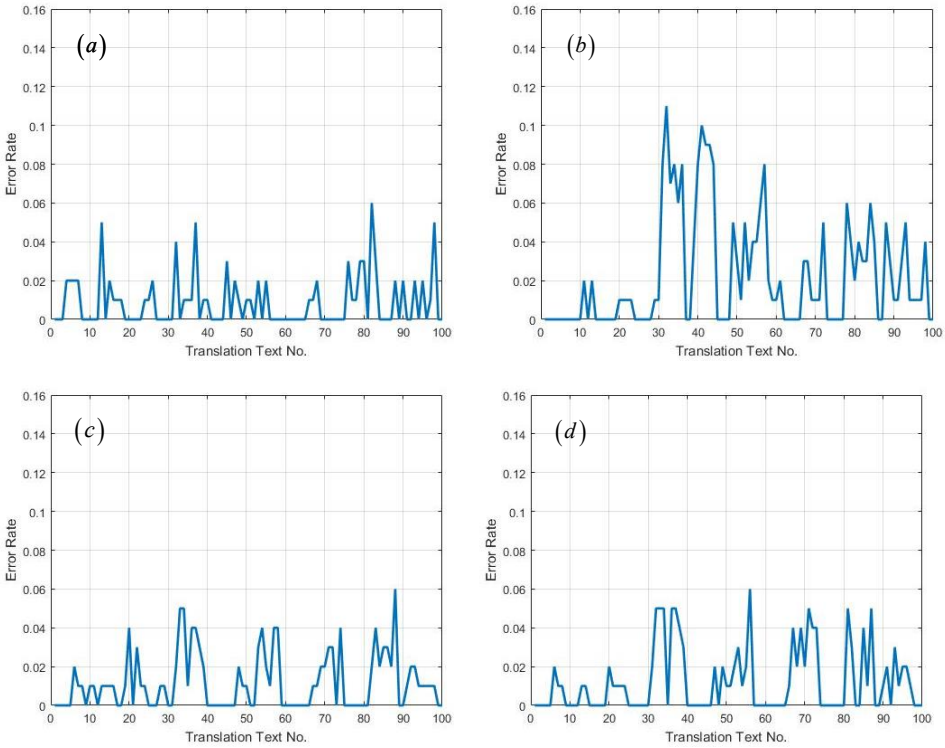

Figure 1. Absolute error rate of ranking of BERT-based system (C–E, bert-base-uncased).

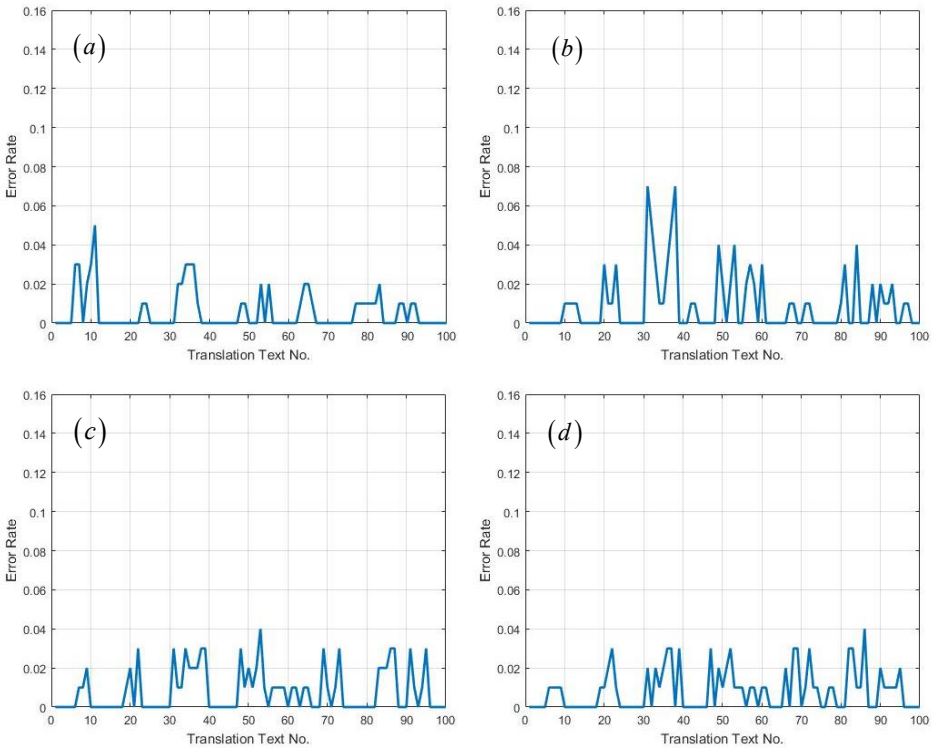

Figure 2. Absolute error rate of ranking of BERT-based system (C–E, bert-base-cased).

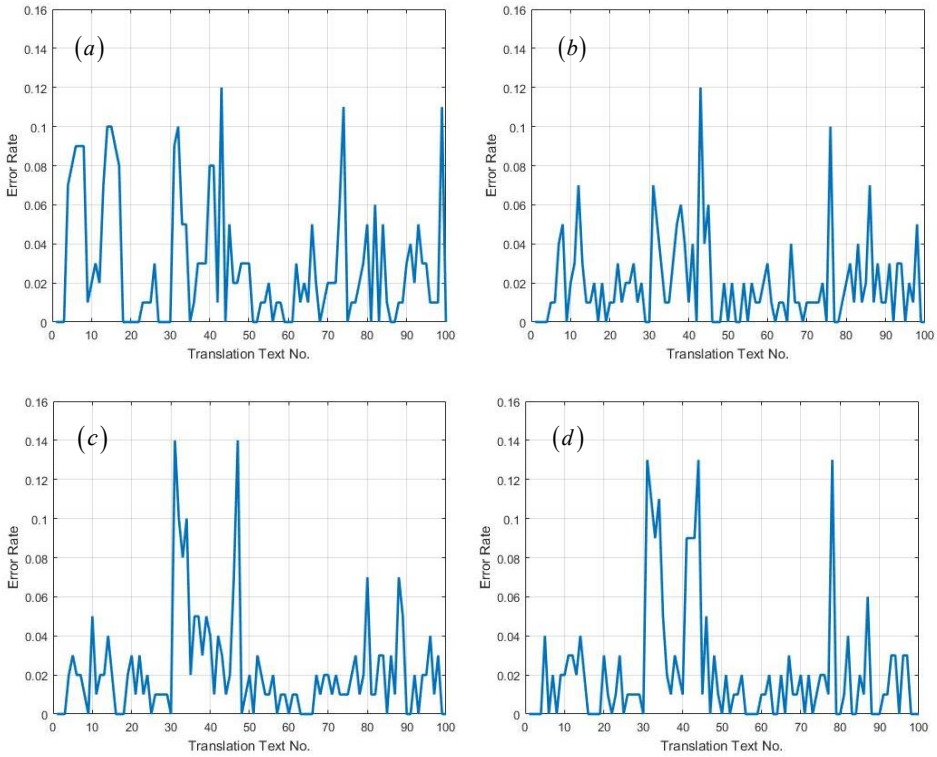

Figure 3. Absolute error rate of ranking of BERT-based system (C–E, bert-large-uncased).

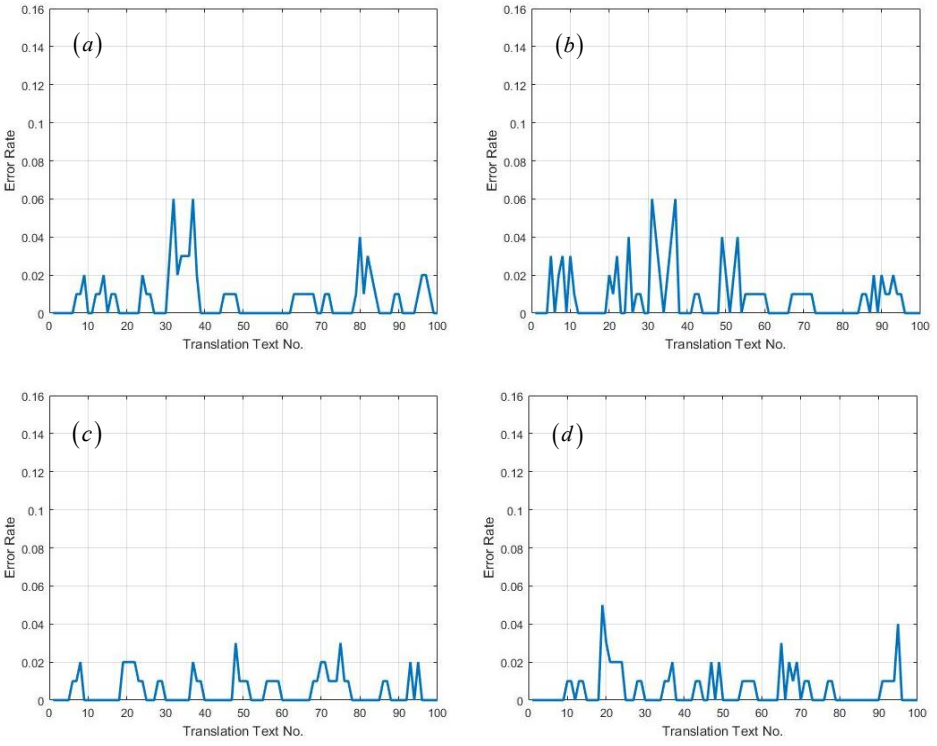

Figure 4. Absolute error rate of ranking of BERT-based system (C–E, bert-large-cased).

Figure 2 depicts the absolute error rate of ranking of synthetic texts generated based on the four expert translations in the BERT-based system (C–E, bert-base-cased), with four sets of curves in (a–d). Overall, the figure indicates a low absolute error rate, with a mean absolute error rate of ranking of 0.0077. Specifically, among the 400 synthetic texts, 59.25% of them have predicted rankings that match their expected rankings, resulting in an

absolute error rate of 0. For 40.25% of the texts, the absolute error rates fall in the range of 0.01 to 0.05. Only 0.5% of the texts have absolute error rates of more than 0.05, with a maximum of 0.07. These results suggest that the errors of the BERT (base cased)-based system do not have a significant impact on the selection of the top-ranked C–E translations.

Figure 3 illustrates the absolute error rate of ranking of synthetic texts generated using the BERT-based system (C–E, bert-large-uncased), with four sets of curves in (a–d). The results show a low absolute error rate, with a mean absolute error rate of ranking of 0.0229. Specifically, among the 400 synthetic texts, 28.25% of them had predicted rankings that matched their expected rankings with an absolute error rate of 0. For 60.75% of the texts, the absolute error rates were in the range of 0.01 to 0.05. Only 11% of the texts had absolute error rates greater than 0.05, with a maximum of 0.14, which does not seem to have a significant impact on the selection of the top-ranked C–E translations.

Figure 4 shows the absolute error rate of ranking of synthetic texts generated based on the four expert translations in the BERT-based system (C–E, bert-large-cased), with four sets of curves in (a–d). Overall, it presents a low absolute error rate, with the mean absolute error rate of ranking of 0.0066. More specifically, among the 400 synthetic texts, 60.25% of the predicted rankings are the same as their expected rankings, while 38.75% have absolute error rates in the range of 0.01 to 0.05. Only 1% of the texts have absolute error rates of more than 0.05, with a maximum of 0.06, which does not seem to have a significant impact on the selection of the top-ranked C–E translations.

4.1.2. Reliability Report

The results shown in Table 4 indicate that the BERT-based systems (C–E) appear to have a high degree of agreement with human raters. Specifically, the accuracy ranges from 0.56 to 0.66, the recall ranges from 0.63 to 0.72, and F0.5-Score ranges from 0.63 to 0.72, which are all encouraging. Interestingly, the accuracy, recall, and F0.5-Score of the four kinds of models do not vary greatly from each other. The BERT (base uncased)-based system and BERT (large cased)-based system both have a relatively better agreement with human raters in comparison with BERT (base cased)-based system and BERT (large uncased)-based, with an accuracy of 0.6522, and a recall and F0.5-Score of 0.7108.

Table 4. Consistency between BERT-based system and human raters (C–E).

Model	Accuracy	Recall	*F*0.5-Score
bert-base-uncased	0.6522	0.7108	0.7108
bert-base-cased	0.5652	0.6385	0.6385
bert-large-uncased	0.6377	0.6887	0.6887
bert-large-cased	0.6522	0.7108	0.7108

After conducting the reliability analysis, it can be concluded that the BERT-based systems (C–E) all have high reliability, in which the BERT-based system (C–E, bert-base-uncased) and BERT-based system (C–E, bert-large-cased) perform the best.

4.2. E–C Translations

4.2.1. Results of Model Selection Indicators

Table 5 shows the overall indicator results of model selection for BERT-based system (E–C) is good. In Table 5, the value of r is 0.999157000 with a corresponding $p < 2.2 \times 10^{-16}$, which indicates a strong correlation between the expected ranking and the predicted ranking. The value of the mean absolute error rate of ranking is 0.0051, and the value of the maximum absolute error rate of ranking is 0.0700, which indicates that the average fluctuation is under 0.51%, and that the maximum fluctuation is no more than 7%; therefore, it can be considered that the high-quality translations are not filtered out.

Table 5. Results of model selection indicators for BERT-based system (E–C).

Measures	Spearman's Rank Correlation Coefficient (r)	Significance (p)	Mean Absolute Error Rate of Ranking	Maximum Absolute Error Rate of Ranking
bert-base-chinese	0.999156900	$< 2.2 \times 10^{-16}$	0.0051	0.0700

Figure 5 illustrates the absolute error rate of ranking of synthetic texts generated based on the four expert translations in the BERT-based system (E–C, bert-base-chinese), with four sets of curves in (a–d). Overall, the figure shows a low absolute error rate, with a mean absolute error rate of ranking of 0.0051. Specifically, out of the 400 synthetic texts, 73.75% of them have predicted rankings that match their expected rankings, with an absolute error rate of 0. For 25.50% of the texts, the absolute error rates range from 0.01 to 0.05, while only 0.75% of the texts have absolute error rates higher than 0.05, with a maximum of 0.07. These error rates seem unlikely to significantly impact the selection of the top-ranked E–C translations.

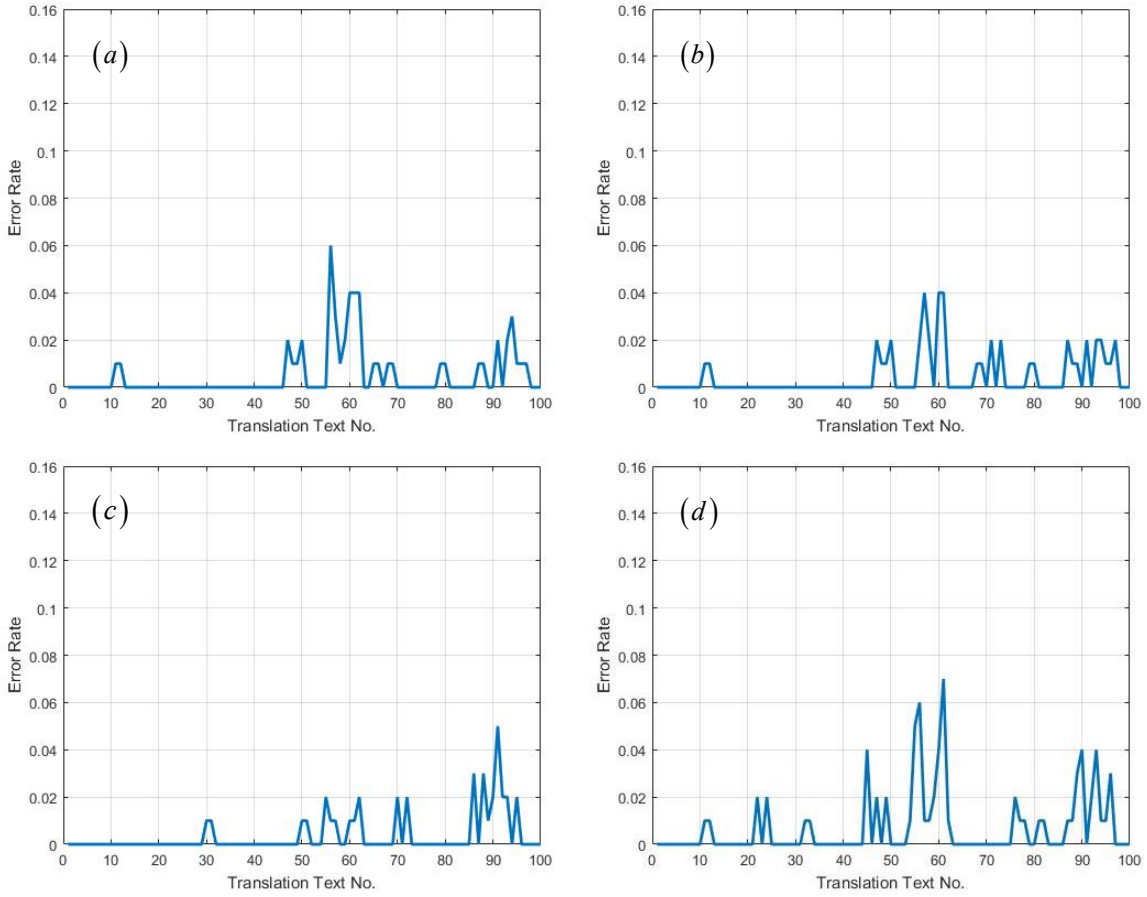

Figure 5. Absolute error rate of ranking of BERT-based system (E–C, bert-base-chinese).

4.2.2. Reliability Report

The BERT-based system (E–C) exhibits a high degree of consistency with human raters. Table 6 presents the system's results, which demonstrate good accuracy at nearly 0.75, with recall and F0.5-scores of nearly 0.85.

Table 6. Consistency between BERT-based systems (E–C) and human raters.

Text Pre-Processing	Accuracy	Recall	F0.5-Score
bert-base-chinese	0.7480	0.8421	0.8421

After conducting the reliability analysis, it can be concluded that the BERT-based system (E–C, bert-base-chinese) has high reliability and can be put into operation.

5. Machine Translation Detection

With the development of NLP, there is concern as to whether machine translations will be submitted directly to the contest and whether the system will regard such translations as high-quality translations. Thus, a test was conducted to see if the machine-translated texts could be recognized as good translations by the system. Four widely used machine translation services, Baidu translation (fanyi.baidu.com, accessed on 18 November 2022), DeepL translation (deepl.com, accessed on 18 November 2022), Google translation (translate.google.com, accessed on 18 November 2022), and Sogou translation (fanyi.sogou.com. accessed on 18 November 2022), were selected for this test.

For C–E systems, a total of 142 translations, including the four machine translations, were entered into the BERT-based systems for scoring, and the average score was 83.09, 79.15, 78.96, and 78.97, with a maximum score of 100 and a minimum score of 0. The scores for the machine translations are shown in Tables 7–10. Although all the machine translations receive a certain number of points, indicating that they are acceptable, none of them receive a very high score, indicating that there is still a gap between the quality of the machine translations and the quality of the human translations.

Table 7. Machine translations scored by BERT-based system (C–E, bert-base-uncased).

Machine Translation	Score
Baidu translation	80.25
DeepL translation	83.28
Google translation	81.52
Sogou translation	82.91

Table 8. Machine translations scored by BERT-based system (C–E, bert-base-cased).

Machine Translation	Score
Baidu translation	53.02
DeepL translation	67.53
Google translation	59.33
Sogou translation	56.13

Table 9. Machine translations scored by BERT-based system (C–E, bert-large-uncased).

Machine Translation	Score
Baidu translation	80.50
DeepL translation	84.34
Google translation	74.82
Sogou translation	77.71

Table 10. Machine translations scored by BERT-based system (C–E, bert-large-cased).

Machine Translation	Score
Baidu translation	49.15
DeepL translation	61.66
Google translation	53.86
Sogou translation	49.34

For the E–C system, a total of 266 translations, including the four machine translations, were entered into the system for scoring, and the average score was 86.13, with a maximum

score of 100 and a minimum score of 0. The scores for the machine translations are shown in Table 11. It can be seen that, although all the machine translations receive a certain number of points, indicating that they are acceptable, none of them receive a very high score, indicating that there is still a gap between the quality of the machine translations and the quality of the human translations.

Table 11. Machine translations scored by BERT-based system (E–C, bert-base-chinese).

Machine Translation	Score
Baidu translation	77.64
DeepL translation	72.29
Google translation	61.90
Sogou translation	47.63

6. Conclusions

The experimental results show that the BERT-assist system serves as a reliable and efficient second-rater for the large-scale translation contest in terms of translation quality, as it can effectively sift out high-quality translations with a reliability of $r = 0.9$ or higher. Moreover, it demonstrates the capability to assess 10,000 translations within a mere 3 h timeframe.

The consistency between the system-generated scores and the human-assigned scores is satisfactory, with a maximum accuracy 0.65+ (C–E) and 0.74+ (E–C); a maximum recall of 0.71+ (C–E) and 0.84+ (E–C); and a maximum F0.5-Score of 0.71+ (C–E) and 0.84+ (E–C). The use of segmentation provides possibilities to handle the BERT input constraint.

While this preliminary case study demonstrates the effectiveness of a BERT-based system in predicting translation rankings based on translation quality with a high degree of confidence and sifting out high-quality translations that can potentially be awarded, further validation is necessary to ensure its generalizability. Additional studies, including evaluations in diverse, large-scale translation contests, are essential. Furthermore, considering the potential variations in translation requirements across different genres, such as expository, descriptive, and narrative genres, it may be prudent to develop different systems tailored to each genre, with each utilizing specific, large language models.

Author Contributions: Conceptualization, Y.C. and M.L.; methodology, Y.C. and M.L.; software, Y.C. and M.L.; validation, Y.C.; formal analysis, Y.C.; investigation, Y.C.; resources, M.L.; data curation, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, M.L.; visualization, Y.C.; supervision, M.L.; project administration, M.L.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundamental Research Funds by North China University of Technology, grant number 2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We are grateful for the data provided by the TAC program.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Severyn, A.; Moschitti, A. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 373–382.
2. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q. V XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.

3. Devlin, B.; Liu, R. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
5. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
6. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
7. Kant, N.; Puri, R.; Yakovenko, N.; Catanzaro, B. Practical Text Classification with Large Pre-Trained Language Models. *arXiv* **2018**, arXiv:1812.01207.
8. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676.
9. Chae, Y.; Davidson, T. *Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning*; Open Science Foundation: Charlottesville, WV, USA, 2023. [CrossRef]
10. Ahn, S. Experimental Study of Morphological Analyzers for Topic Categorization in News Articles. *Appl. Sci.* **2023**, *13*, 10572. [CrossRef]
11. Robinson, J.; Rytting, C.M.; Wingate, D. Leveraging Large Language Models for Multiple Choice Question Answering. *arXiv* **2022**, arXiv:2210.12353.
12. Guo, Q.; Cao, S.; Yi, Z. A Medical Question Answering System Using Large Language Models and Knowledge Graphs. *Int. J. Intell. Syst.* **2022**, *37*, 8548–8564. [CrossRef]
13. Kang, B.; Kim, Y.; Shin, Y. An Efficient Document Retrieval for Korean Open-Domain Question Answering Based on ColBERT. *Appl. Sci.* **2023**, *13*, 13177. [CrossRef]
14. Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; Hashimoto, T.B. Benchmarking Large Language Models for News Summarization. *arXiv* **2023**, arXiv:2301.13848. [CrossRef]
15. Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.-B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Collins, W.; Ahuja, N.; et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *arXiv* **2023**, arXiv:2309.07430.
16. Keswani, G.; Bisen, W.; Padwad, H.; Wankhedkar, Y.; Pandey, S.; Soni, A. Abstractive Long Text Summarization Using Large Language Models. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 160–168.
17. Hasan, T.; Bhattacharjee, A.; Islam, M.S.; Samin, K.; Li, Y.-F.; Kang, Y.-B.; Rahman, M.S.; Shahriyar, R. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *arXiv* **2021**, arXiv:2106.13822.
18. Slimani, T. Description and Evaluation of Semantic Similarity Measures Approaches. *arXiv* **2013**, arXiv:1310.8059. [CrossRef]
19. Petrakis, E.G.M.; Varelas, G. X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. *J. Digit. Inf. Manag.* **2006**, *4*, 233–237.
20. Panchenko, A. RUSSE: The First Workshop on Russian Semantic Similarity. *arXiv* **2015**, arXiv:1803.05820.
21. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
22. Levenshtein, V.I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
23. Kruskal, J.B. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Rev.* **1983**, *25*, 201–237. [CrossRef]
24. Heeringa, W.J. Measuring Dialect Pronunciation Differences Using Levenshtein Distance. Ph.D. Thesis, University of Groningen, Groningen, The Nederlands, 2004.
25. Fiscus, J.G.; Ajot, J.; Radde, N.; Laprun, C. Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 22–28 May 2006; pp. 803–808.
26. Su, K.-Y.; Wu, M.-W.; Chang, J.-S. A New Quantitative Quality Measure for Machine Translation Systems. In Proceedings of the COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992.
27. Nießen, S.; Och, F.J.; Leusch, G.; Ney, H. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May 2000.
28. Akiba, Y.; Imamura, K.; Sumita, E. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In Proceedings of the Machine Translation Summit VIII, Santiago de Compostela, Spain, 18–22 September 2001.
29. Leusch, G.; Ueffing, N.; Ney, H. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In Proceedings of the Machine Translation Summit IX, New Orleans, LA, USA, 23–27 September 2003.
30. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
31. Leusch, G.; Ueffing, N.; Ney, H. Efficient MT Evaluation Using Block Movements. In Proceedings of the EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics), Trento, Italy, 6 April 2006; pp. 241–248.
32. Culy, C.; Riehemann, S.Z. The Limits of N-Gram Translation Evaluation Metrics. In Proceedings of the Machine Translation Summit IX: Papers, New Orleans, LA, USA, 23–27 September 2003.

33. Doddington, G. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, USA, 24–27 March 2002; pp. 138–145.
34. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
35. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
36. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
37. Harris, Z.S. Distributional Structure. *WORD* **2015**, *7956*, 146–162. [CrossRef]
38. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the The 31st International Conference on Machine Learning (ICML 2014), Beijing, China, 21–26 June 2014; Volume 32, pp. 1188–1196.
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
40. Weigle, S.C. Effects of training on raters of ESL compositions. *Lang. Test.* **1994**, *11*, 197–223. [CrossRef]