COMP4450 Group Assignment3

May 2022

Tong Cai (u6619441)
Huijie Cui (u7261480)
Qiuling Liu (u6845212)

Question 1

Our group is interested in machine translation systems testing. In this area, it mainly focuses on the correctness of the translation. We plan to use the widely developed testing method including BLEU, NIST, METEOR and word error rate in the research. The results of machine translation testing are compared with the human translation and a score is given based on the degree of difference. In this case, our group aims to compare the performance of different machine translators and compare different test methods.

For the first research question, we focus to explore the functionalities of some widely used machine translation systems, in which we compare the testing results for different translation systems with the same data set. For the data set, we also consider the style of contents so we include the small size, larger size, formal language style and colloquial texts. The different sizes of data can help to eliminate the effects of size in testing. The different styles of content help to further compare different machine translation systems for different demands.

For the second research question, due to the most of testing methods, we used in the research are logically similar. We compare the testing method by exploring the algorithms. Based on the findings, we make hypotheses about the advantages, disadvantages and suitable texts for the different testing methods. For each hypothesis, we test it utilizing control variables. For example, for the synonym feature, we replaced the difference in test results before and after the synonym comparison.

Question 2

Natural language processing is an indispensable research direction in the field of artificial intelligence, aiming to solve the problem of how humans and machines communicate in natural language. As machine translation is an important application of natural language processing, today, there are many translation systems based on different algorithms, and the performance of these systems needs to be properly evaluated.

Nowadays, we can use all kinds of machine translators. Although people may use some sentences that are easily translated incorrectly to compare them, it cannot fully measure their translation performance. If we want to compare the accuracy of the output of two translators, we find that if different test sets or different test methods are used, the final comparison results may be different. If we make some improvements to the algorithm of a certain system, in order to prove that these improvements really improve the performance of the system, we need to evaluate the output results before and after these improvements. Using multiple test methods and test sets will greatly improve the reliability of the conclusion, and if the above-mentioned problems occur during this test, it will probably cause confusion for researchers.

In order to analyze these different results, we think that first, we need to understand how different test methods work and compare the characteristics of these test methods. We have chosen four commonly used test methods for comparison, which is also helpful to choose a more effective method according to the size of the test set when scoring machine translation later. For example, a certain method may be more suitable for evaluation at the data set level, but its scores on several sentences are often unreliable, or a certain method may not run successfully on a very large data set, so we need to adjust the size of the data set. In this process, we use multiple translators to translate the same text, which not only increases the credibility of our conclusions but also makes an initial attempt to compare different machine translators. Therefore, we will also compare different machine translators as our second research question.

Question 3

The state of art regarding our research question:

Machine learning has been successful in providing general-purpose natural language translation systems, with many systems able to translate between thousands of pairs of languages effectively in real-time. (Kim Hazelwood, etc， 2018) With a large number of uses, there is a concern about the effectiveness and accuracy of machine translation. The difficulty of testing machine translation lies in the Oracle problem and machine learning techniques. The criteria for translation accuracy are very difficult to define and require consideration not only of the meaning of words but also of grammar, sentence structure, sentence meaning and style of expression. The criteria and aspects of the system that are of concern will have a bias on the testing system. Human evaluations of machine translation are extensive but expensive. (Papineni.K, etc,2002) So people started to develop automatic testing methods.

Initial testing methods such as BLEU were based on manual ratings, but the results of automated human and machine assessments varied considerably. Nowadays, there is a wide variety of testing methods for machine translation. In many papers, we found several BLEU-based tests such as NIST, METEOR, METEOR-PL, etc. which have improved the functionalities. For BLEU, the developers used statistical analysis of the relevance of judgments for translating from four different languages into English to enhance the capabilities of machine translation and optimize the model in the article. BLEU averages judgment errors for individual sentences in the test corpus. The author said: "We believe that BLEU will accelerate the MT R&D cycle by allowing researchers to rapidly home in on effective modelling ideas." (Papineni. K, etc., 2002)

For the metrics of NIST, they are analyzed at the sentence, document, and system-level with results conditioned by various properties of the test data. (Przybocki, M, 2009) This is implemented based on the BLEU. Compared with BLEU, this method adds a synonyms package and focuses on the importance of grams and sentences which makes this method's results closer to human evaluation results. In terms of validation, the authors provide statistics on the results of the human evaluation and the automated evaluation of the test method to demonstrate the consistency of the results of both evaluations. In other words, although the scoring data differs, the evaluation results are consistent.

For METEOR, there are many versions. The researchers have presented Ranking, Adequacy, and Tuning versions of Meteor 1.3. The Ranking and Adequacy versions are shown to have a high correlation with human judgments except in cases of overfitting due to skewed tuning data. (Denkowski, M, 2011) They are still exploring an improved version of the METEOR testing method. The meteor method uses the recall function in order to adjust the penalty for short grams. The solutions of the recall function are infinite, some of them perform better with other side-effects, so there are lots of new versions generated.

In that case, our team thinks that the accuracy of the testing method is increasingly important, that's why our main purpose is to use the testing method and compare the various testing method. For testing, we found some papers using statistics to analyse the methods. So, for the most relevant state of the art papers, our group focused on finding papers compared to the testing methods, especially which can prove the improvements between those testing methods.

Huijie Cui:
Paper name: Interpreting bleu/nist scores: How much improvement do we need to have a better system?

In this article, firstly, the author makes an in-depth comparison of the characteristics of BLEU and NIST. According to the specific calculation method of BLEU and NIST, it is found that the precision of 3-gram and the precision of 4-gram contribute a high proportion to the final result of BLEU. This is because BLEU generally uses the geometric average of the precisions of n-gram as the final result, while NIST implicitly weights the precision of n-gram with the information contained in all n-grams, so 3-gram and 4-gram with very little information hardly contributed to the final result. As we know, unigram precision represents the precision of words in the output of a translation system, while n-gram precision represents the precision of the order of the output of a translation system. Therefore, the author finds that BLEU pays more attention to the order precision of the output, which sometimes makes the output with a very low word precision get a better result strangely. NIST, on the other hand, pays more attention to the word precision of the output, but hardly pays attention to the order.

Then, the author raises a question: How to determine the accuracy of the score? Because it is very difficult to get enough suites containing source language sentences and multiple reference sentences, the author first uses bootstrapping to get multiple replacement samples and then scores these sample suites. As expected, it is found that the score obtained has a normal distribution. It is easy to get the confidence interval of the score then. By observing the data obtained from seven different translation systems, the author finds that the translation system with higher average scores will also have a higher relative standard deviation. But the author did not analyze this phenomenon.

Finally, the author explores the question that we are most concerned about: how to determine whether there is really a good or bad relationship between the two systems? As in the previous study, the author also uses bootstrapping to get enough suites. This time, the author calculates the difference in scores between different machine translation systems on each suite and finds that these differences also match a normal distribution. If the confidence intervals of the score differences between two translation systems are all located on the side of zero, then these two translation systems are regarded as having good and bad differences. By observing the actual data, the author finally mentions the possible differences between NIST and BLEU in the results of this comparison method: NIST may consider that there are actual differences between the two translation systems, while BLEU thinks there are no differences.

Qiuling Liu :
Paper name: "Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking."

This paper focuses on the automatic assessment of re-speaking. This article proposes a 're-speaking' technique based on automatic speech recognition technology. Re-speaking is the use of automatic speech recognition to reconstruct a scene and generate fully annotated subtitles for the re-recorded audio. Unlike risking the occurrence of incorrect speech recognition due to ambient sounds, re-speaking means instead that a specially trained person repeats the speech from the recorded event in a quiet and controlled environment, (Wołk, K,2016) which eliminates double talk, cocktail party effects and grammatical-semantic problems of paraphrasing.

However, this paper only discusses the issue of quality assessment, with the main comparison being the accuracy of the assessment. One of the most challenging aspects of data accuracy is how to consider semantics, rather than simply comparing words. The article compares machine translation testing methods such as BLEU, NIST, METEOR, METEOR-PL, TER, etc. with several automatic and human-assisted metrics.

The first research part of this article is analysing the testing metrics of each testing method, along with the comparison of similar testing methods including BLEU, NIST, METEOR, METEOR-PL and EBLEU. For other testing methods, the paper analyses those separately. The main differences between those testing methods are synonyms, rare words and determination of cumulative score which consists of penalty score. In order to eliminate those differences, the researcher designed a rewards and penalties system. The synonyms, they design to reward matches of synonyms so that sentences can be rewarded while maintaining the correct meaning. The sentence meaning is set to 'exact match' and simple synonyms are set to synonymy, with different score processing to ensure the integrity of the sentence meaning. When dealing with rare words, extra points are given to matches for rare words. The cumulative score section incorporates the default BLEU score, using logarithms and exponents to reduce the impact of overly heavy penalties on test accuracy.

The NER model is an indicator of word accuracy which is specifically used to measure the accuracy of subtitles. This method will take into account consistency and repeatability.
The source of the database was generated based on the concept of ‘re-speaking'. So, there will be 3 types of data, the original recording, the re-speaking text and the recognition of the re-speaking speech.

The data were collated and processed in the direction of linear regression, presenting a variable on the degree of NER variation and discharging some prominent minor variables at each set of treatments. This allows the data to be discharged as anomalies after multiple treatments, and a comparison of the regression curves for accuracy allows the accuracy of the different testing methods to be compared. Based on the regression curves for the different test methods, the data were processed and the authors generated a NER function - 'NER = 86.55 + 0.254 - BLEU + 0.924 - NIST - 0.221 - EBLEU ' (Wołk, K,2016) to improve the accuracy of the test, combining the advantages of the various testing methods to eliminate potential over-penalisation.

Tong Cai
Paper name: "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments."

This paper focuses on an evaluation metric for machine translation systems called METEOR. As an important automatic metric in machine translation, its importance in improving testing efficiency and reducing testing costs was first introduced by comparing it to traditional human evaluations. However, except for METEOR, there are other automatic metrics like BLEU and NIST have already been used extensively, thus, to better explain the algorithm of METEOR, a comparison with BLEU and NIST was also made.

The authors pointed out that a major difference between METEOR and BLEU is that BLEU takes the brevity penalty as compensation for the lack of recall. A similar issue also exists in NIST because it is developed based on BLEU's algorithm. In addition to this issue, METEOR is expected to solve other drawbacks of BLEU/NIST both in the use of N-gram and explicit word-matching. In this case, METEOR makes an alignment containing a set of stages between machine translation and references, after that, the score for METEOR is calculated as a combination of F-Mean and Penalty, where the calculation of F-Mean takes both Precision and Recall into consideration.

Afterwards, this paper conducted a series of experiments to check METEOR's performances on DARPA/TIDES 2003 Arabic-to-English and Chinese-to-English datasets. The first experiment was to compare METEOR to NIST/BLEU by computing the system-level Pearson correlation of each metric score, the listed results were evident to show that considering Recall as one measurement in the metric helps to improve the correlations between metric scores and human assessments, because METEOR obtained higher correlation than the other two metrics which only considered Precision. Besides, to further investigate the significance of METEOR, similarly, an experiment was conducted to compare the Pearson correlations of METEOR scores and other metrics like Precision, Recall and F-mean. From the results, as a combination of Precision and Recall, F-mean obtained a higher correlation with the human assessment, also, by introducing Penalty as well, METEOR score obtained the highest correlation.

As we have described above, at each stage in METEOR implementation, there are different mapping modules to map unigrams based on a different roles. For instance, these modules map unigrams either by introducing stemmers or synonyms. Thus, four experiments were then conducted in the paper to observe their differences in the correlation with human assessments. Those four experiments contain a different number of stages and mapping modules separately, as well as distinct mapping modules. Observing results on both datasets, it was obvious that adding more stages and distinct mapping modules helped improvement in the correlation.

Moreover, considering the noise that existed in human assessment scores, one more experiment was conducted to check whether the normalization of it will take rise to the correlation, and the obtained data was convincing that the METEOR score received a better correlation with human assessment by normalizing it before.

In summary, the paper discussed METEOR as a valid metric to evaluate machine translation from the results of experiments on checking the availability of METEOR's components and different mapping modules.

Question 4

In the mid-break, once determining the abstract research questions to compare both machine translators and test methods, but currently, there are many sorts of test methods been developed, but we could only focus on comparing several of them, so we reviewed some relevant papers about testing machine translators to get some ideas on selecting the test methods.

We first chose BLEU score as a metric because it is a popular and extensively used approach to test machine translators by comparing machine translations and professional translations, also, we found that both NIST and METEOR have made some improvements based on BLEU's algorithm, so we determined to consider NIST and METEOR as well to check whether these improvements effectively help them to better test machine translators. Moreover, we used WER (Word Error rate) as another metric, WER is implemented in a completely different algorithm compared to the other three methods, we chose it to see its performance in testing the same machine translators. As for machine translators, we first selected Google translator and DeepL translator because both are currently popular among users.

After that, in week 7, we began to search for papers about testing different machine translators to get some hints on choosing the appropriate dataset. We noticed that two papers commonly used the datasets from WMT-17 (the second conference of machine translations), so we decided to give the first attempt to solve our research questions by downloading a parallel data called "News Commentary v12" from the official website of WMT-17, due to the large size of the original data, we only picked the text containing 13000 words as our test data to check whether our research questions are valid or not.

In the next week, we implemented the algorithm of BLEU and WER by referring to the codes on Github. Before that, we translated the chosen text from Chinese to English using Google translator and DeepL translator respectively. After testing the two different translations in BLEU, we found that Google Translate obtained a similar score as DeepL translator. However, when testing those translations in WER, our computer automatically terminates the program due to the big size of the text, we only got the WER score of the translation by reducing the text to 800 words.

In the next two weeks, as one of our purposes is to study the different results of one translation system testing on different data sets, in the first, we wanted to choose a new dataset that is different from the original data set in both sizes and language style to prove that there are indeed differences in results. So, we collected a colloquial text containing around 500 words from textbooks for the interpreting profession to get a more comprehensive evaluation result. Also, we observed that most published papers have already tested Google translator, so we chose SouGou translator as well to get more comparisons of different machine translators. Besides, we implemented both NIST and METEOR by importing nltk package. We used two different data sets for evaluation, and then we found that when using BLEU and NIST for evaluation, every translator was scored differently on different data sets. However, when using BLEU for testing, we even found that different test sets led to a difference in the ranking of the performances in these three translators——SouGou translator performed the best on small colloquial text but performed the worst on large WMT-17 text. However, when using Meteor for evaluation, every translator obtained similar scores on different data sets, which really proved that there might be a situation we were worried about.

However, we were not sure whether this kind of situation was caused by the language styles or the sizes of different data sets. To better clarify the reasons, we selected a data set that contains the first 500 words out of all 13,000 words from the WMT-17 text to make it the same size as the colloquial text. After that, we evaluated the performances of different translators on their translations of big-sized WMT-17 text and the small-sized WMT-17 text, and we discovered that most evaluation metrics obtained similar scores for all translators on these two datasets. It was evident to show that the difference we found above is caused by the different language styles of these test datasets. Besides, we also tested the small-sized WMT-17 text in WER to observe how different machine translators behave, as well as find differences between WER and other test methods. Moreover, because BLEU will only be limited to one language style when using only one reference, hence, if without considering the influence of synonyms, there will be a huge fluctuation. However, when considering Meteor, because it uses the WordNet synonym module distinguishably, so the score will be similar. The results also indicated that SouGou translator's output is more colloquial so it still got the best result even though its language style is limited to the colloquial text.

To better prove this point, we used another dataset by substituting synonyms of the WMT-17 text but keeping the data size unchanged. After that, we translated the new dataset containing synonyms with Google translator, DeepL translator and SouGou translator respectively, and then tested the three translations in BLEU, NIST and METEOR.

What worked in the previous steps could be summarized as:

1. **Comparison of Google translator, DeepL translator and SouGou translator:** In previous steps, we adopted four different test datasets in different sizes, language styles and verbalizations to check how different machine translators perform on different datasets.

   From the results, we observed that these three translators obtained different scores in each testing metric on all test datasets, which verifies that one of our research questions to compare machine translators is valid. Also, these results are valuable for us to further analyze the reasons to cause distinct performance in testing each translator on different test datasets. In this case, we can better compare these translators based on a set of comprehensive results.

2. **Comparison of WER and the other three test methods (BLEU, NIST and METEOR):** When considering the large-sized WMT-17 text containing 13000 words as the test dataset, BLEU, NIST and METEOR immediately returned the evaluation scores, but WER is not valid to test the same dataset, it only returned a result when reducing the size of the test data to around 800 words. However, when testing the small-sized WMT-17 test dataset, all four test methods immediately returned a result.

   Even though there is a limitation of the size of the dataset in testing machine translations for WER, the relevant steps help us to solve our research question to compare different test methods, because the results indicate that WER is not suitable to test small datasets, while BLEU, NIST and WER are effective to test datasets regardless of the size of datasets.

3. **Comparison of BLEU and the other two test methods (NIST and METEOR):** From the evaluation results of three translators on the first three test datasets (big-sized WMT-17 text, small-sized WMT-17 text and colloquial text), we cannot clearly see the obvious differences between BLEU, NIST and METEOR in testing machine translations.

   However, when testing the final dataset which substitutes synonyms of the big-sized WMT-17 text, from the results, we noticed that compared to the other two test methods, BLEU is not always effective to test this dataset due to the impact of synonyms. These findings work in helping us compare these three test methods.

What doesn’t work in the previous steps could be summarized as:

1. **Test machine translation with a larger size (around 13000 words) in WER:**
   When we tried to test all machine translators with WER on the large-sized dataset, it always took a very long time until the system automatically interrupted the program, so we failed to get a direct WER score for further comparison and analysis of both test methods and machine translators on this test dataset.

2. **Comparison of NIST and METEOR:**
   As we did in the previous steps, despite the differences in their implementation algorithm, we still wanted to seek a more intuitive comparison of different test methods by observing the obtained scores on each of them.

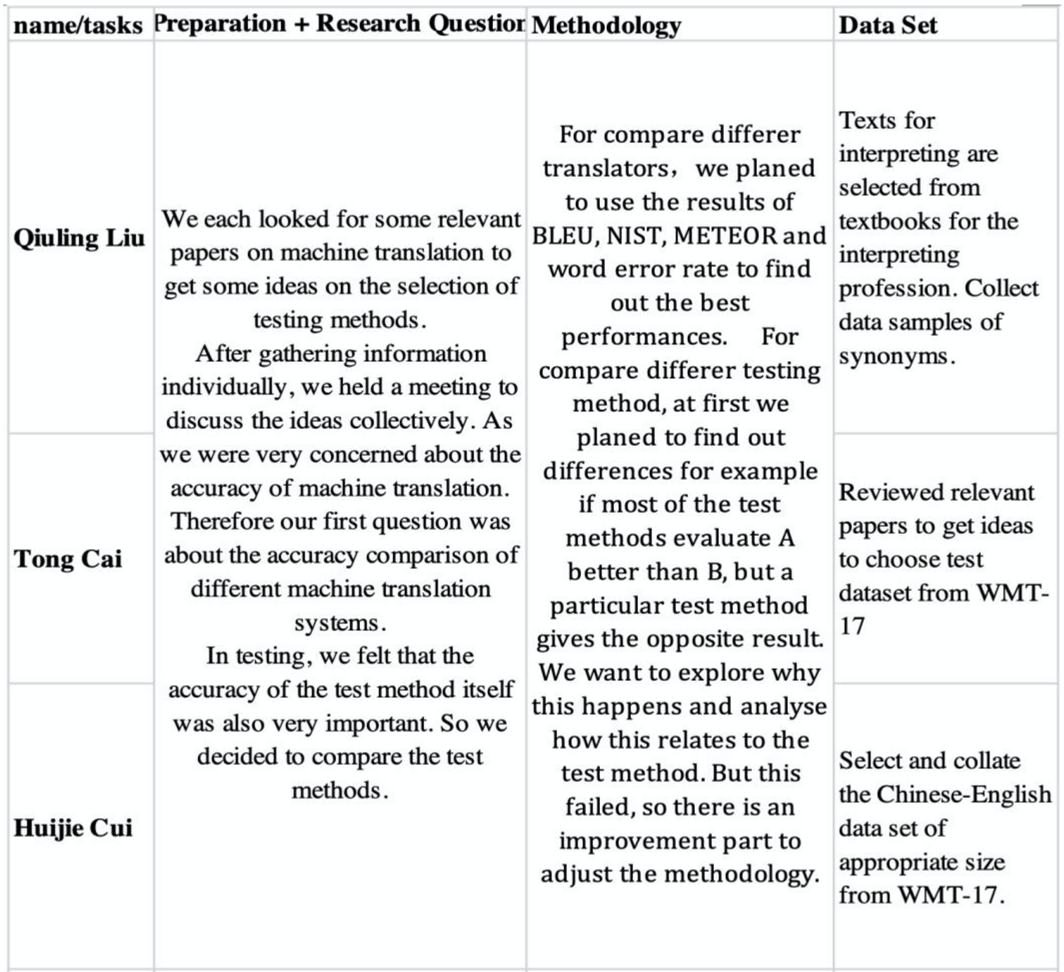
   We have already tested several datasets in different sizes and different language styles, and we also introduced synonyms in one dataset to better compare BLEU, NIST and METEOR. However, from the results, we could only observe the difference between BLEU and the other two methods rather than the major differences between NIST and METEOR. Then those steps didn't work in making the comparison in all test methods.

3. **Comparison of WER and the other three test methods (BLEU, NIST and METEOR) in other aspects except for the size of test data:**
   From the previous steps, we only found that a major difference between WER and the remaining three test methods is that there is a size limitation for the test dataset in WER. However, when testing the small-sized WMT-17 text on all four methods, although we already knew that a lower WER score points to a better translation performance, while a higher BLEU, NIST and METEOR score points to a better performance instead, we still failed to observe a direct difference between WER and the other three test methods.

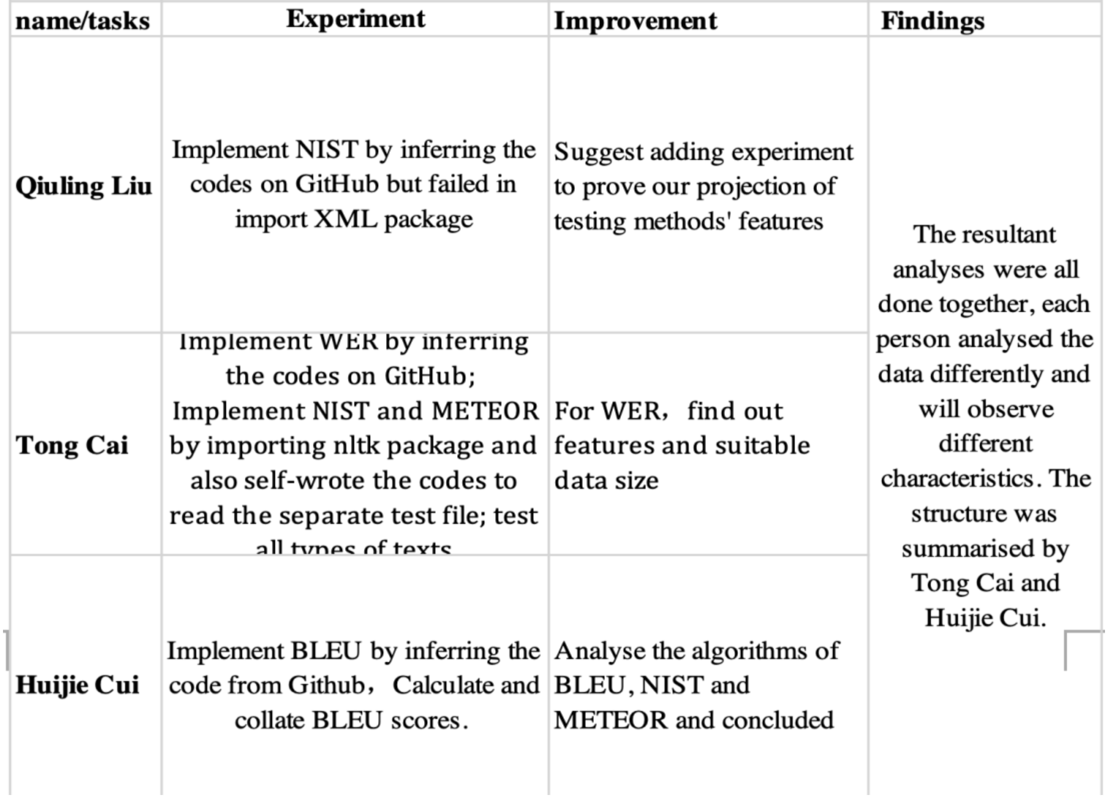
Question 5

Most of our group's tasks are done together in discussion. As most of our work and difficulties lie in coming up with suitable ways to compare different approaches to machine translation testing. Our group felt that adequate communication and collective ideas would enrich our research. For example, when thinking about how to compare test methods, our original expectations did not match the test results making it difficult to progress with the experiment. We were able to change our minds thanks to timely and active communication. In terms of new experimental designs, it was also because we brainstormed together as a group that we were able to come up with the right approach.

On tasks that can be done by one person, we distribute them fairly. For example, in the collection of data, we consider that one person can collect general data. Another person can prepare the spoken database. In terms of test methods, we also initially planned for one person to be responsible for the generation of results for one test method. But we found a very well-established package that could test BLEU, NIST and METEOR at the same time and we changed the allocation to be done by one person.



name/tasks	Preparation + Research Questior	Methodology	Data Set
Qiuling Liu	We each looked for some relevant papers on machine translation to get some ideas on the selection of testing methods. After gathering information individually, we held a meeting to discuss the ideas collectively. As we were very concerned about the accuracy of machine translation. Therefore our first question was about the accuracy comparison of different machine translation systems. In testing, we felt that the accuracy of the test method itself was also very important. So we decided to compare the test methods.	For compare differer translators， we planed to use the results of BLEU, NIST, METEOR and word error rate to find out the best performances. For compare differer testing method, at first we planed to find out differences for example if most of the test methods evaluate A better than B, but a particular test method gives the opposite result. We want to explore why this happens and analyse how this relates to the test method. But this failed, so there is an improvement part to adjust the methodology.	Texts for interpreting are selected from textbooks for the interpreting profession. Collect data samples of synonyms.
Tong Cai			Reviewed relevant papers to get ideas to choose test dataset from WMT-17
Huijie Cui			Select and collate the Chinese-English data set of appropriate size from WMT-17.

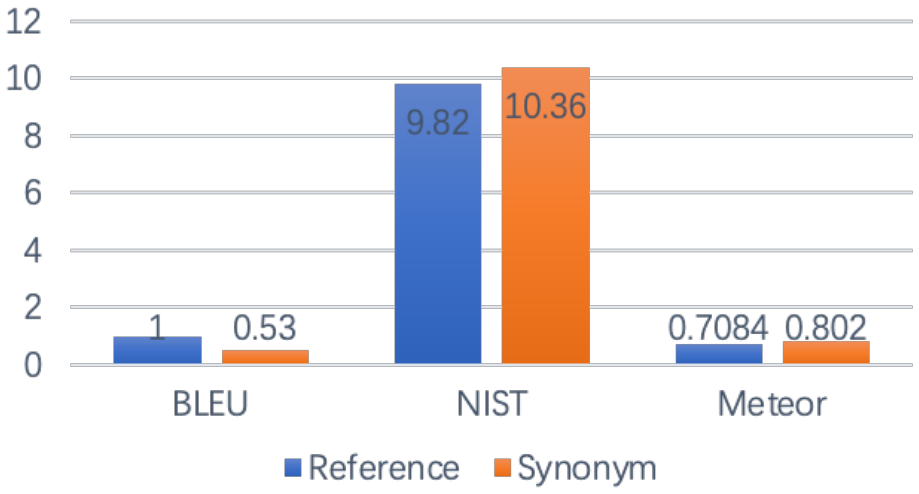


name/tasks	Experiment	Improvement	Findings
Qiuling Liu	Implement NIST by inferring the codes on GitHub but failed in import XML package	Suggest adding experiment to prove our projection of testing methods' features	The resultant analyses were all done together, each person analysed the data differently and will observe different characteristics. The structure was summarised by Tong Cai and Huijie Cui.
Tong Cai	Implement WER by inferring the codes on GitHub; Implement NIST and METEOR by importing nltk package and also self-wrote the codes to read the separate test file; test all types of texts	For WER， find out features and suitable data size	
Huijie Cui	Implement BLEU by inferring the code from Github， Calculate and collate BLEU scores.	Analyse the algorithms of BLEU, NIST and METEOR and concluded	

Question 6

For the first research question, compare different machine translation testing methods. We started by analyzing the metrics of BLEU, NIST and METEOR. For BLEU, we found out that it is convenient, fast and easy to calculate, with results that are closer to human ratings because its logic is followed the human rating logic. But it is a lack of consideration of sentence meaning and synonyms. BLEU needs a package of different language styles' references to support or its accuracy will be affected by language styles. For NIST, it introduces synonyms and increases the weight of some keywords that appear less frequently, which is the rare words. For METEOR, it increases the accuracy and recall of the whole corpus so that it can help to check the completeness of sentences and also introduce the synonyms. But it will over-punish the short grams situation.

For synonyms, we make a projection that the NIST and METEOR can perform better than the BLEU. We use the reference text as a reference, replacing some words with synonyms as a candidate. As you can see from the results in the graph, BLEU is the only one that results in a decrease, the other two will have increased. We, therefore, believe that BLEU is most affected by synonyms and that a large number of references are needed when testing texts, otherwise it has a significant impact on the accuracy of the results.



For word error rate (WER), we failed to test 10000 words data set with this method. Because this method will compare data sentence by sentence which is inefficient, the testing function crashed several times. We found out a small size data set is more suitable for WER.

For the second research question, compare different machine translation systems. We used three types of data sets which are big-sized WMT-17 text, small-sized WMT-17 text and colloquial text. For the translation systems, we chose the Google translator, DeepL translator (British style) and SouGou translator. In order to analyse the results, the translator has better performance if it obtains a higher score in BLEU, NIST and METEOR. The translator obtains a better performance if its WER percentage is smaller.

Figure1 is the results from the WER, we can observe that Google translator obtains the smallest WER for both colloquial and formal texts, while DeepL translator obtains the highest WER.

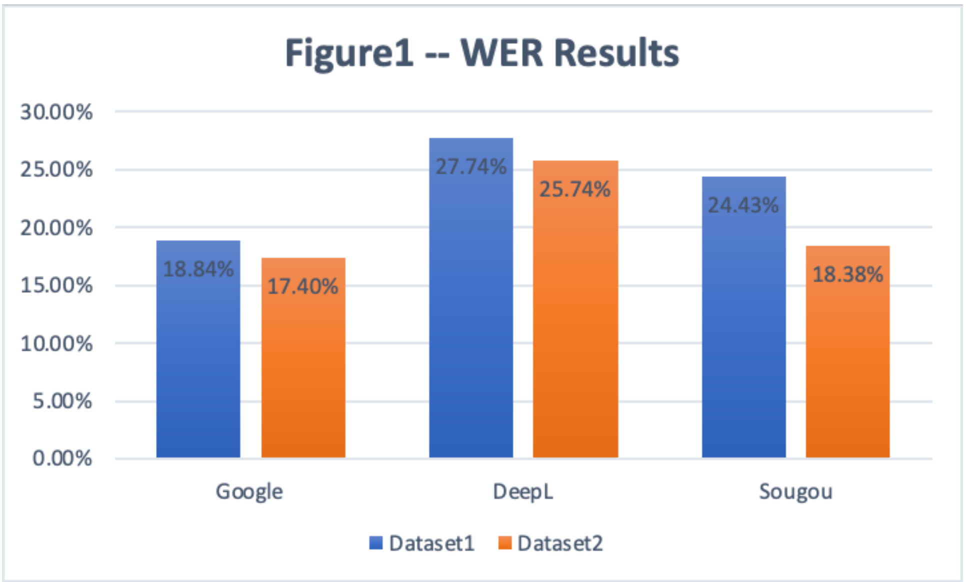
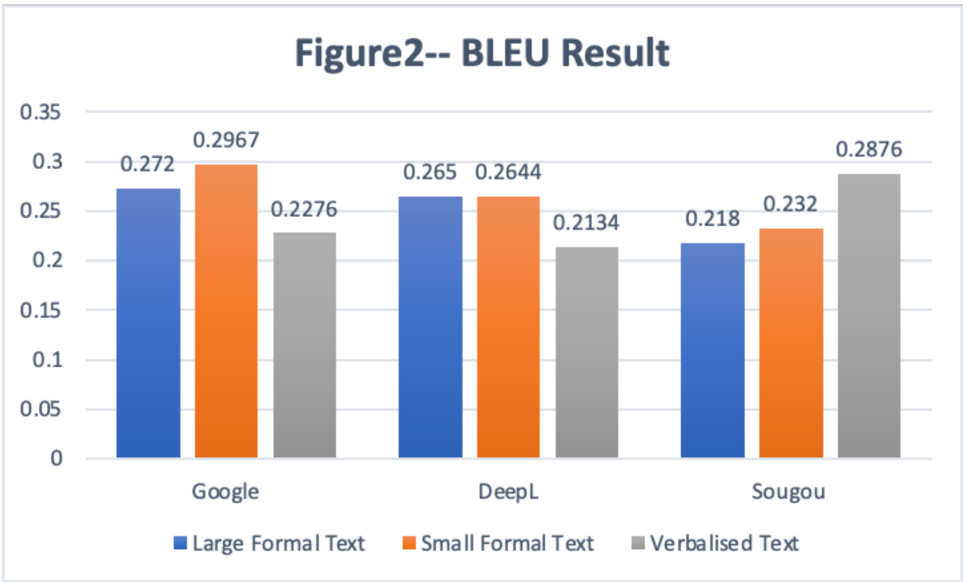
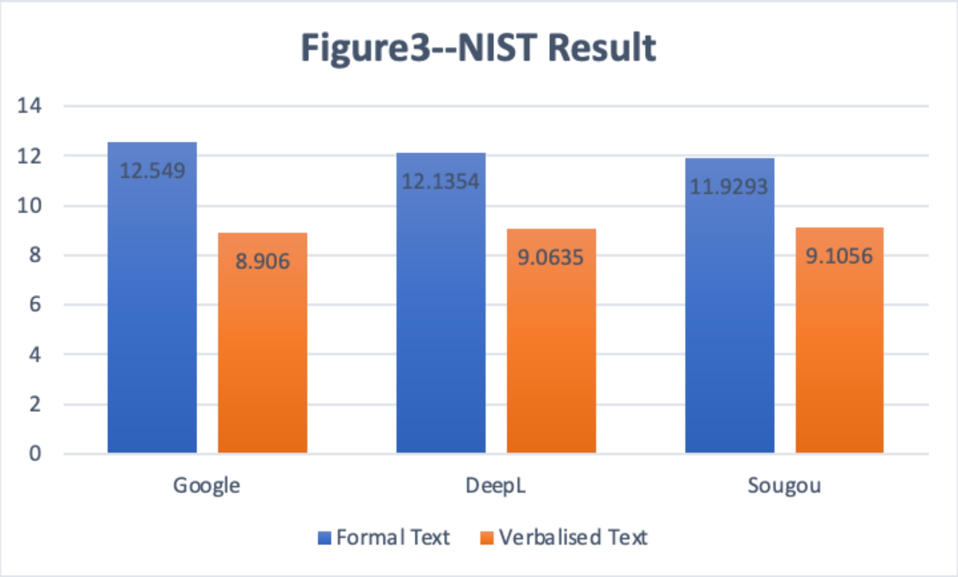
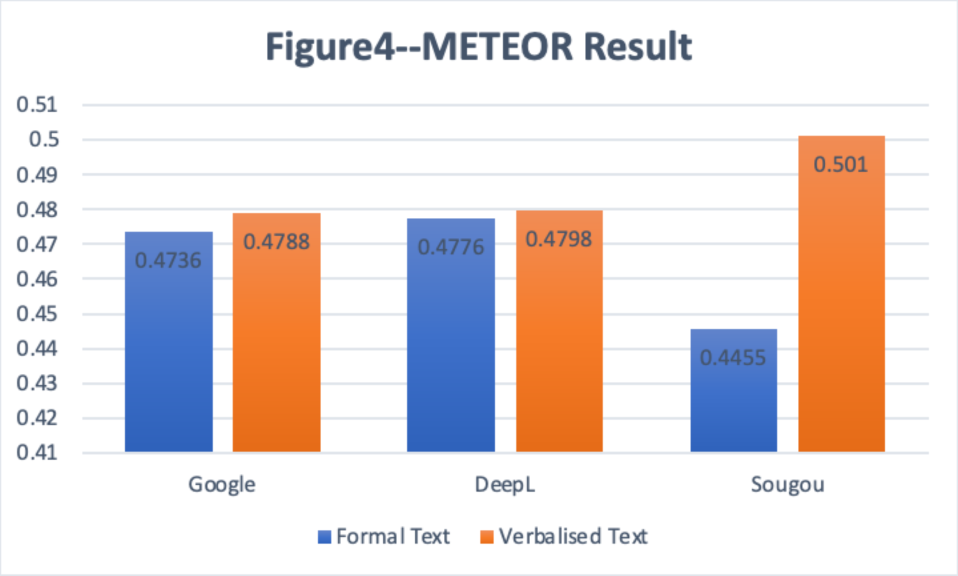


Figure2 is from the BLEU, figure3 is from the NIST, the figure4 is from the METEOR. We can observe that the the SouGou get highest scores in colloquial texts in all three testing methods and get the worst performance scores in formal texts. And in BLEU, the Google performs better than the DeepL with formal text. In NIST and METEOR, DeepL and Google perform equally same.

We found out most of the testing results can prove that SouGou translation is more colloquial, and Google translation and DeepL translation are more official.







Question 7

For comparing the translation systems, our aim is almost completed. We found out SouGou translation fits the colloquial content, but Google translation and DeepL translation are more official. For the next step, we also want to explore whether the language style, for example, the bureaucratization, will also influence the performance. Also, in our previous research, we only use the Chinese - English translation content, the performances of other languages are also essential to evaluate the capabilities of translation systems.

Furthermore, our group considers how to use the testing results to improve the functionalities of each translator. In response to this, we have thought about a few approaches that might work. We plan to introduce the 'transRepair' method which can automatically test and repair. The "transRepair" result may just fix the consistency, but the accuracy of the translation remains the same. To check the capability, we plan to do tests using both the original and the repaired data, and the difference in the results can demonstrate whether the correctness has improved. Another idea is to use the results of the word error rate to judge whether the main issues caused by the error rate are over-translation or under-translation. Based on the specific issue, adjusting the algorithm or using the targeted data to train the translation model may help the translators improve the accuracy.

Our main aim of this research question is thronging the comparison of these widely used machine translation testing methods trying to explore suitable content which can improve the accuracy of testing. And we only focus on the BLEU, NIST, and METEOR. Our original plan was to find out the features of each method's throng algorithms and then use the targeted data set to prove. For example, by consideration of synonyms we use a data replaced part of words with synonyms, and then use the original and replaced data to test. But the results are quite the same, we cannot identify which one is better at synonyms. We found some papers use linear regression and normal distribution to analyse the differences. The paper "Wołk, K., & Koržinek, D. (2016). Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. arXiv preprint arXiv:1601.02789." has shown that the scores of the same testing method obey the linear regression. And the author of "Interpreting bleu/nist scores: How much improvement do we need to have a better system?." calculated the differences in scores for different machine translation systems on each data set and found that these differences were consistent with a normal distribution. For the next step, we plan to use a large sample size trying to do the linear regression to eliminate the abnormal data. For example, if the data is far from the linear regression line we can identify it as abnormal data. And using the normal distribution to identify the confidence intervals of the testing scores. For example, if two methods' confidence intervals are in the same location of zero then we can identify as they have differences.

Question 8

When we read some related papers, we find that sometimes authors only rely on one score or use one reference to evaluate their results. From our research, the result of such an evaluation method is uncertain, because if synonyms are not considered in this evaluation method, it will treat synonyms and wrong words in the same way. The resulting increase in scores is probably related to the language styles of the output results from different translation systems. This also makes us realize that the design of each step in the research needs full investigation in advance and reasonable design. Any mistake may lead to unreliable results.

When comparing the performances of different translators, we didn't design a specific comparison method in advance. We only briefly found some features of the output results of translators, and we didn't propose a method to further test these features. This does not contribute much to the purpose of comparing whether the translator's performance is improved. I think this is because we didn't make sufficient preparations and detailed research plans before we started the first experiment. In general, we started our research with one attempt and planned the future research through the results of attempts. In hindsight, if we can read more papers before starting the research, to learn the strategies of successful researchers and specify a more detailed plan and the purpose of each experiment, we may get more creative and credible results.

In the process of research, we find that if we only study the test methods of the translation system, most of the related research methods are based on calculating the overlap rate between words in the output result and words in reference. There is an obvious improvement relationship among these methods, so the results of comparison among them are mostly that the improved test method is better than that before improvement, and most of these results can be intuitively analyzed from the calculation process of the method. In hindsight, we might as well compare the test method and system performance of dialogue generation. Dialogue generation, like machine translation, belongs to the process of machine generation of natural language and plays an increasingly important role in practical applications. Its related testing methods include not only the machine learning testing methods mentioned in our research, but also some word embedding evaluation methods, which will greatly broaden our research scope and perhaps get more results.

References (APA style)

Zhang, Y., Vogel, S., & Waibel, A. (2004, May). Interpreting bleu/nist scores: How much improvement do we need to have a better system?. In *LREC*.

Banerjee, S., & Lavie, A. (2005, June). 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 (pp. 65-72).

Wołk, K., & Koržinek, D. (2016). Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. arXiv preprint arXiv:1601.02789.

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, and Xiaodong Wang. 2018. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. In 24th International Symposium on High-Performance Computer Architecture (HPCA 2018), February 24-28, Vienna, Austria.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Przybocki, M., Peterson, K., Bronsart, S., & Sanders, G. (2009). The NIST 2008 Metrics for machine translation challenge—overview, methodology, metrics, and results. Machine Translation, 23(2), 71-103.

Denkowski, M., & Lavie, A. (2011, July). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the sixth workshop on statistical machine translation (pp. 85-91).