Acceptability of Google Translate Machine Translation System in Translation from English into Kurdish: A Study on Evaluating Machine Translation Outputs

Fereydoon Rasouli[1], Keivan Seyyedi[1], Soma Soleimanzadeh[2]

[1]Department of Translation, Cihan University-Erbil,
Kurdistan Region, Iraq
[2]Department of Computer Science, Cihan University-Erbil,
Kurdistan Region, Iraq

Abstract—**The development of machine translation (MT) systems and their application in performing translation projects gave a crucial position to the evaluation of these systems' outputs. Recently, the Google Translate MT system added the central accent of the Kurdish language to its language list. The present study is an attempt to evaluate the acceptability of the translated texts produced by the system. Different text typologies have been considered for the study's data. To evaluate the MT outputs, the bilingual evaluation understudy evaluation model has been administered. The findings show that the performance of the understudy MT system in the translation of English into the Sorani accent of Kurdish is affected by some linguistic and technical hindrances, which in general affect the acceptability of translated text.**

***Keywords*—Acceptability of outputs, Automatic evaluation systems, Kurdish language (Sorani).**

I. Introduction

Evaluation of machine translation (MT) outputs is at the center of attention from different perspectives. This can encompass different players in the field, such as operators, designers, and researchers of the systems, according to Arnold et al. (1993), these groups of people are interested in the assessment of MT system outputs from different viewpoints. For a user, the quality and cost of operating MT systems are of the highest importance; however, the theoretical framework and analyzing the sorts of errors to develop systems are crucial questions to which other groups seek to find reliable answers through the assessment of MT systems. The participation of commercials in the debate through different modes of quality assessment, such as using metrics to evaluate the quality of the MT system automatically and the extent of post-editing efforts to measure the usability of an MT system, can be considered in evaluating MT outputs. Dobrinkat (2008) believes that the main purpose of evaluating MT outputs is to compare different MT systems and determine which one performs better in a certain aspect or domain. The author also believes that the system performance can be optimized through modifications based on the findings of an evaluation.

The general viewpoint of the present study is to evaluate the acceptability of translated texts by the Google Translate MT system. The system newly added the Central Kurdish language to its list and the present study tries, for the 1st time, to assess the system's performance based on the principles of automatic evaluation of MT systems, and more specifically, the acceptability of the quality of translated texts into the Central Kurdish language.

II. Literature Review

A. Translation Quality Assessment (TQA)

The translation is a very complicated process that embedded all linguistic, cultural, social, and technological dimensions. A close description of this process and assessing the quality of the final version of a translated text can illustrate the exact complexity and, consequently, the difficulty of assessing the quality of the translation. Different theoretical definitions

Cihan University-Erbil journal of Humanities and Social Sciences (CUEJHSS)
Volume VIII No. 1 (2024) 8 pages
DOI: 10.24086/cuejhss.vol8n1y2024.pp 7-14
Received: 20 May 2023; Accepted: 20 June 2023; Regular research paper: Published: 01 January 2024
*Corresponding author's e-mail: fereydoon.rasouli@cihanuniversity.edu.iq

of the quality of a translated text are open grounds for discussing quality as the source text-the oriented concept of accuracy or fluency that is oriented toward the target text. In general, Gaspari et al. (2015) believed that by the evolution and widespread application of translation technology such as MT and the difficulties and constraints of the process, various definitions of TQA have been presented. Taking into consideration the issue of translation quality, "different sectors in the field pursue different aims and accordingly ask different questions" (Drugan, 2013). A user of the MT system is interested in the TQA to see whether a specific level of quality is determined and evaluated appropriately and whether the results of such an assessment are satisfying. Obtaining a measure to demonstrate the change in quality or improvement from previous versions is the aim of TQA research.

Despite, difficulties and inconsistencies in the forms of evaluating MT system outputs, various approaches proposed by scholars in the field (Graham, 2015). In general, these approaches are classified into the human evaluation and automatic evaluation of MT systems.

Human evaluation, the traditional model of MT evaluation, has been done by bilingual evaluators who understand both source and target languages well and judge the quality of MT output at the sentence level. In this Model, the quality can be assessed based on two perspectives of linguistic correctness and usability of the outputs. According to White (2003), human evaluation is "an intuitive way by which the evaluators try to decide over the goodness of the final outputs" (p.232). To determine the goodness of a translated project, a set of attributes such as fluency, adequacy, and intelligibility are presented; however, fluency and adequacy are the most common ones that are scored out of 5 and illustrated in the following table: (Koehn, 2010).

Although the evaluation of MT outputs by humans is to some extent precise and assesses different aspects of the final projects, conducting such comprehensive work is very time-consuming and, at the same time, very expensive; therefore, doing the evaluation automatically can be a reasonable alternative to doing the task faster and more efficiently than human evaluation.

Different researchers have proposed a variety of models of automatic evaluation of MT outputs, like Meteor, which was initially proposed in 2004 and designed to correlate with a human evaluation of MT outputs (Lavie et al., 2007). Meteor is working on the level of words and computing a score based on a clear word-to-word match between the translated text by MT and the presented reference translation (Rasouli, 2018). In case of the availability of more than one reference translation, the system scores the output against each reference separately and uses the best one as the evaluation score.

Another model for evaluating the performance of MT systems is the word error rate (WER). The model originated from the Levenshtein distance and works based on the minimum number of editing steps at the level of words in the text. (Koehn, 2010) In the WER, editing steps include substitution, omission, and addition; in addition, the WER measures the similarity of word sequences by assessing the lowest number of editing steps needed to turn the MT output into the reference sentence (Dobrinkat, 2008). The most widely used automatic model of evaluation of MT output is the bilingual evaluation understudy (BLEU), which was designed in the IBM labs (Kishore et al., 2001) to acquire a quick and economical method of evaluation of MT outputs. The translation edit rate (TER) proposed by Snover et al. (2006) is based on the edit distance; however, in TER, reordering of the blocks is allowed, and at the same time, it uses additional editing steps for changing the sequences of words. Turian et al. (2003) have presented a model of automatic MT evaluation that applies the notion of maximum matching strings. In this work, BLEU has been selected as the automatic evaluation metric to assess the acceptability of the Kurdish MT system. In the next section, the main framework of BLEU will be discussed in detail, and after that, the concept of acceptability will be elaborated.

B. BLEU

As mentioned, a variety of factors are weighed in the human evaluation of the MT outcome; these factors include adequacy, fluency, and fidelity. Florence Reeder (2001) proposed comprehensive models and techniques for the evaluation of MT; however, as Hovy (1999) elaborated, most MT evaluation approaches conducted by humans are very expensive and at the same time can take months to complete. In general, human evaluation methods have not been warmly welcomed from different perspectives. For instance, from the point of view of MT developers, the factors of time and price are the major problems of human evaluation approaches because they want to observe the effect of ongoing changes on their systems to find out the weak points of the systems. To solve these problems, researchers in the field have proposed several automatic MT evaluation metrics. BLEU metrics is a widely used automatic evaluation system that is similar to human evaluation and has been selected as the evaluation model of the present study.

The main principle in BLEU is comparing a translated sentence by the MT system to one or more human translations, which are called candidate sentences and reference sentences, respectively. The model was introduced to the field by Kishore et al., 2002. The central idea in the BLEU metric is the extent to which a MT outcome is closer to a translated text performed by a professional translator based on the numerical metric (Kishore et al., 2002). In this model the procedure of evaluation is done through two phases: first, by comparing n-grams in both candidate sentences and reference translation/s of the same candidate, then, by counting the number of matched words regardless of word order (ibid).

Table I
Scored Criteria in Human Evaluation MT

Adequacy	Grade	Fluency	Grade
All meaning	5	Flawless	5
Most meaning	4	Good	4
Much meaning	3	Non-native	3
Little meaning	2	Diffluent	2
None	1	Incomprehensible	1

MT: Machine translation

Table II
Evaluation of Sentence 1 by BLEU Metric

Sentence 1	BLEU	Length ratio	Text
Source			The Kurdistan Region is known for having valleys, mountains, forests, as well as small and large rivers.
Reference	100.00	1.00	هەرێمی کوردستان بە هەبوونی دۆڵ، کوێستان، دارستان وهەروەها رووباری گەورە و بچووک ناسراوە.
Candidate	0.32	1.14	هەرێمی‘، ’کوردستان‘ بە هەبوونی دۆڵ و شاخ و دارستان و هەروەها ڕووباری بچووک و گەورە ناسراوە‘،.

Table III
Evaluation of Sentence 2 by BLEU Metric

Sentence 2	BLEU	Length ratio	Text
Source			Tourists can enjoy mountain climbing at Halgurd, Pera Magrun, Korek, Bradost, Sheren, and Saffin mountains.
Reference	100.00	1.00	گەشتیاران دەتوانن چێژ لە شاخەوانی چیاکانی هەڵگورد، پیرە مەگرون ، کۆرەک ، برادۆست، شیرین و سەفین وەربگرن.
Candidate	0.15	1.4	گەشتیاران دەتوانن چێژ لە سەرکەوتن بەسەر شاخ وەربگرن لە هەڵگورد , پێرا شاخەکانی ماگرون و کۆرەک و برادۆست و شێرێن و سەفین

According to the baselines of the model, the greater the number of matched words between the candidate translation and its reference, the better the performance of the MT system. The familiar precision measure in BLEU is the cornerstone of this metric for evaluating MT outputs. To calculate precision, the number of unigrams (candidate translation words) that occur in any reference translation are counted and then divided by the total number of words in the candidate sentence (Kishore et al., 2002). By counting the similarity of n-grams in both reference and candidate sentences, one can simply decide on the accuracy of the translation, but as pointed out, these similarities are text-independent and do not present any credible data regarding fluency and other criteria to evaluate the acceptability of a translated text. To deal with this system, it is advised to match higher ranks of grams for analyzing a sentence (Kishore et al., 2002). The following is a review of how precision scores for *n* = 1 g through *n* = 4 g are calculated as the first step of acquiring a BLEU score:

Precision 1 g = No. matched Candidate 1-g/Total No. of candidate 1-g

Source sentence

The Kurdistan Region is known for having valleys, mountains, forests, as well as small and large rivers.

Reference sentence

هەرێمی کوردستان بە هەبوونی دۆڵ، کوێستان، دارستان و هەروەها رووباری گەورەو بچووک ناسراوە.

Candidate sentence

هەرێمی‘، ’کوردستان‘ بە هەبوونی دۆڵ و شاخ و دارستان و هەروەها’ ڕووباری بچووک و گەورە ناسراوە‘،.

So, precision 1- g = 14/16 0.87

Precision 2-g = No. matched Candidate 1-g/total No. of candidate 1-g

Reference sentence

هەرێمی کوردستان بە هەبوونی دۆڵ، کوێستان، دارستان و هەروەها رووباری گەورە و بچووک ناسراوە.

Candidate sentence

هەرێمی‘، ’کوردستان› بە هەبوونی دۆڵ و شاخ و دارستان و هەروەها’ ڕووباری بچووک و گەورە ناسراوە‘،.

So, precision 2- g = 07/15 0.46

And so on:

Precision 3-g = 5/14 0.35

Precision 4-g = 4/13 0.30

The higher the match *n* = 1, 2, 3, and 4 g between reference and candidate sentences, the more acceptable the output of MT is. This description is the main rationale behind the BLEU metric, so BLEU can be defined as

$$BLEU = BP \times \left(\prod_{n=1}^{4} p_n \right)^{\frac{1}{4}}$$

Based on this definition, the modified n-gram precision between the reference translation (human translation) and the candidate sentence is measured as Pn. There are some cases where the length of the candidate sentence is shorter than its related reference sentence; in this case, the 1-g precision would have been 1/1 = 1, which illustrates a perfect score; however, this is a very misleading score. To deal with these kinds of problems, the brevity penalty (BP) downscales sentences that are shorter than reference through the following formula:

$$BP = \begin{cases} 1 & if\, c > r \\ \exp(1 - \frac{r}{c}) & if\, c \le r \end{cases}$$

According to the formula

- C refers to the number of words in the candidate sentence
- r shows the number of words in the reference sentence.

As mentioned, the BP makes sure that even if the candidate sentence is much longer than the reference sentence, the BP cannot be greater than one.

C. Acceptability

The concept of acceptability has been widely used to talk about the quality of the translated project in the both human translation and automatic translation. The most common definition of acceptability, which is followed by many researchers in the field, was presented by Van Slype (1979). He defined acceptability as "a subjective judgment by which the final user of a translation work thinks the output is acceptable." The writer believes that acceptability should be measured through survey questions (ibid.). The subjective nature of acceptability made it difficult to present a precise and clear definition of it. In defining acceptability, Chomsky (1969) emphasizes the degree to which a text is acceptable. De Beaugrande et al. (1981) look at acceptability from the reader's viewpoint and define it as their attitude and orientation toward the text. Usability, quality, and being satisfied with the final project are characteristics enumerated by Castilho (2016) to define the concept of acceptability. Castilho et al. (2018) stated that in terms of cohesion, coherence, and accuracy, acceptability is one of the main factors in assessing MT outputs that depict the users' attitude toward the final translated text. To measure the acceptability of MT outputs, factors such as usability (cognitive effort and efficiency), satisfaction (through surveys and questionnaires), and quality (grammar, clarity, syntax, etc.) are used (Castilho, 2016). In the current research, a combination of quality, usability, and user satisfaction is used to highlight the acceptability of MT output.

III. Materials and Methods of the Study

The present study is an attempt to evaluate the acceptability of MT output in the translation of English texts into Kurdish (Sorani), which has recently been added to the list of languages on the Google Translate system. The data were gathered from authoritative sources, mainly from government publishers, and translated by the Google Translate system; the same texts were given to three human translators to collect the reference sentences of the study. The translators have been selected based on their experience in the field of translation, and all of them are qualified translators. To evaluate the performance of the MT system, the BLEU metric was administered. According to this model, a translation can be considered acceptable when the n-gram between a candidate and its reference sentence shows a higher rank of matches at the level of words. These matches are position-independent. Overall, the reasons for the administration BLEU score in the present study are as follows:

- It's understandable and easy to calculate
- The similarity of function in evaluating the same texts by humans
- It's language-independent
- It is widely used, and the results can be compared to similar works (Kishore et al., 2002).

The notion of acceptability, which is the pivotal criterion of evaluation in the current research, is a subjective judgment by which the final user of a translation work thinks the output is acceptable. To assimilate BLEU results to human evaluation, accuracy, and fluency, the researchers counted the similarity of grams to four ranks of words, so that the precision scores for 1-ram through $n = 4$ g should be accounted for as the first step.

IV. Results of the Study

Automatic evaluation of MT outputs is proposed to solve the main problems of human evaluation methods. The BLEU metric has been administered to calculate the match grams between reference and candidate sentences. To evaluate the collected data in this paper, a Python program was used, and the result is presented in the following tables:

As can be seen, the BLEU score of the reference sentence is considered 1 an optimal score in the system, the evaluation of candidate sentence 1 shows a score of 0.32, and the length ratio between reference and candidate sentences is 1.14 according to Python. The following details have been acquired by analyzing sentence No. 1.

Sentence 1: {'bleu': 0.3237722713145643,
'precisions': [n1 = 0.625, n2 = 0.4, n3 = 0.2857142857142857, n4 = 0.15384615384615385],
'brevity_ penalty': 1.0,
'length_ ratio': 1.2307692307692308,
'translation_ length': 16,
'reference_ length': 14}

The same procedure has been followed regarding sentence No. 2, and details are presented in the following table:

Sentence 2: {'bleu': 0.15362208233245514,
'precisions': [n1 = 0.47619047619047616, n2 = 0.2, n3 = 0.1052631578947, n4 = 0.055555555555], 'brevity_ penalty': 1.0, 'length_ ratio': 1.4, 'translation_ length': 21, 'reference_ length': 15}

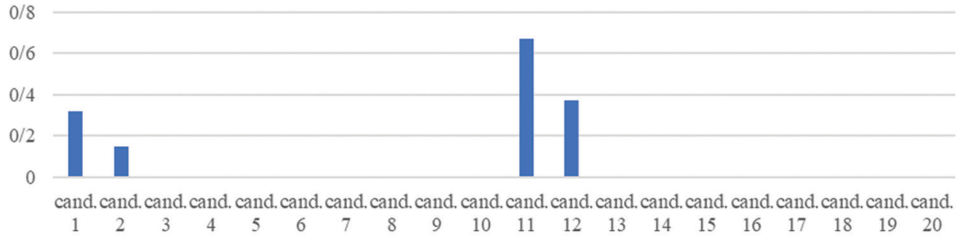
The following bar chart 1 shows the BLEU score of analyzed data by Python; overall, the acquired BLEU score of most of the sentences is equal to zero, and only sentences no. 1, 2, 11, and 12 show a rise toward their related references with scores of 0.32, 0.15, 0.67, and 0.37, respectively. More details are presented in the discussion section, and the analyzed data are available in the appendix of the present study.

Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as secondary units (in parentheses). This applies to papers in data storage**.** For example, write "15 Gb/cm$^{2\,(}$100 Gb/in^2)." An exception is when English units are used as identifiers in trade, such as "3½-in disk drive." Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation.

The SI unit for magnetic field strength H is A/m. However, if you wish to use units of T, either refer to magnetic flux density B or magnetic field strength symbolized as $\mu_0 H$. Use the center dot to separate compound units, e.g., "A·m^2."

V. Discussion

The main concern of this study is the acceptability of the output of the available MT system in the translation of English into Kurdish (Sorani). To have a better understanding of the notion of acceptability, the two factors of accuracy and fluency have been taken as the main scales. Counting precision matches between reference and candidate sentences showed that most sentences enjoy a significant number of matches between pairs of sentences



Bar Chart I: BLEU scores of candidate sentences.

on the level of 1 g, and this can result in the accuracy of translated sentences; in contrast, this trend is reversed by matching sentences on the higher gram levels, which bring the fluency of the MT outputs into question. To express results statistically, collected data have been analyzed by BLEU metric as the model of evaluation in the present study. Based on the definition of BLEU, the closer the MT output is to the reference translation, which is performed by a human translator, the better translation has been provided by the MT system. The findings presented the reality that the majority of translated sentences by the Google Translate MT system have not received a significant score to be called acceptable. Throughout all 20 sentences analyzed by the Python program, only four sentences partially showed compatibility with their related reference sentences. In general, the findings lead to this result: since the analyzed MT output has a score that is lower than the BLEU expectation, the translated sentences in the present study cannot be considered acceptable. Some factors are influential in this regard; as it turned out, BLEU cannot be administered to analyze Kurdish language texts due to the lack of a unified spelling system among speakers of the language. When it comes to the acceptability of the translated sentences, most of them have a relatively similar meaning to the reference sentences; however, the statistics show a considerable disparity in this reality. In addition, comparing a candidate sentence to more reference sentences can lead to better results and can be applied as a strategy to solve the above-mentioned problem.

VI. Conclusion

The acceptability of MT output in translating texts from English into Kurdish is the essential concern of the present study. To reach a result about the mentioned concern, twenty sentences have been analyzed by the BLEU model of evaluation. To do so, a Python program has been used, and the findings showed no reliable data regarding the acceptability of translated text in general. However, closer attention to the results makes it clear that, due to some linguistic problems, the BLEU metric cannot be a useful metric to decide on the acceptability of Kurdish-translated text conducted either automatically or manually. However, looking at the BLEU system from another perspective and giving more value to the n-grams in the lower ranks can lead to more valuable results.

References

Arnold, D, Sadler, L., & Humphreys, R.L. (1993). Evaluation: An assessment. *Machine Translation*, 8(1-2), 1-24.

Castilho, S (2016). *Measuring Acceptability of Machine-Translated Enterprise Content*. Ph.D. Thesis, Dublin City University

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). *Approaches to Human and Machine Translation Quality Assessment. Translation Quality Assessment: From Principles to Practice*. Berlin: Springer. p.9-38.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

De Beaugrande, R.A., & Dressler, W.U. (1981). *Introduction to Text Linguistics*. Vol. 1. London: Longman.

Dobrinkat, M. (2008). *Domain Adaptation in Statistical Machine Translation Systems via User Feedback (Doctoral Dissertation, Master's thesis, Helsinki University of Technology).*

Drugan, J. (2013). *Quality in Professional Translation: Assessment and Improvement*. London: Bloomsbury.

Florence Reeder. (2001). Additional MT-Eval References. Technical Report, International Standards for Language Engineering, Evaluation Working Group. Available from: https://isscowww.unige.ch/projects/isle/taxonomy2

Gaspari, F., Almaghout, H., & Doherty, S. (2015). A survey of machine translation competencies: Insights for translation technology educators and practitioners. *Perspectives Studies in Translatology*, 23(3), 333-358.

Graham, Y. (2015). Improving the Evaluation of Machine Translation Quality Estimation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Vol. 1. Long Papers. pp.1804-181.

Hovy, E.H. (1999). Toward finely differentiated evaluation metrics for machine translation. In: *Proceedings of the Eagles Workshop on Standards and Evaluation*, Pisa, Italy.

Kishore, P., Salim, R., Todd, W., John, H., & Florence, R. (2002). Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish results. In: *Proceedings of Human Language Technology 2002*, San Diego, CA.

Koehn, P (2010). *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In: *Proceedings of the Workshop on Statistical Machine Translation*, Prague. pp.228-231.

Rasouli, F. (2018). Assessment of machine translation output: A comparative study between human and automatic models. *Cihan University-Erbil Scientific Journal*, 2, 119-141.

Rasouli, F. (2022). The impact of developing short-term memory on the interpretation performance of students. *Cihan University-Erbil Journal of Humanities and Social Sciences*, 6(1), 64-68.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). 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*. pp.223-231.

Turian, J.P., Shen, L., & Melamed, I.D. (2003). Evaluation of machine translation and its evaluation. In: *Proceedings of MT Summit IX*, New Orleans. pp.386-393.

Van Slype, G. (1979). *Critical Study of Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-1914*2. Bureau Marcel van Dijk, Bruxelles.

White, J.S. (2003). How to evaluate machine translation. *Benjamins Translation Library*, 35, 211-244.

Appendix I

Analyzed results of Candidate sentences via Python

Sentence 1: {‘bleu’: 0.3237722713145643, ‘precisions’: [0.625, 0.4, 0.2857142857142857, 0.15384615384615385], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.2307692307692308, ‘translation_length’: 16, ‘reference_length’: 13}

Sentence 2: {‘bleu’: 0.15362208233245514, ‘precisions’: [0.47619047619047616, 0.2, 0.10526315789473684, 0.05555555555555555], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.4, ‘translation_length’: 21, ‘reference_length’: 15}

Sentence 3: {‘bleu’: 0.0, ‘precisions’: [0.0, 0.0, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.1428571428571428, ‘translation_length’: 8, ‘reference_length’: 7}

Sentence 4: {‘bleu’: 0.0, ‘precisions’: [0.25, 0.0, 0.0, 0.0], ‘brevity_penalty’: 0.7788007830714049, ‘length_ratio’: 0.8, ‘translation_length’: 4, ‘reference_length’: 5}

Sentence 5: {‘bleu’: 0.0, ‘precisions’: [0.4444444444444444, 0.125, 0.0, 0.0], ‘brevity_penalty’: 0.8007374029168082, ‘length_ratio’: 0.8181818181818182, ‘translation_length’: 9, ‘reference_length’: 11}

Sentence 6: {‘bleu’: 0.0, ‘precisions’: [0.6666666666666666, 0.2727272727272727, 0.1, 0.0], ‘brevity_penalty’: 0.9200444146293233, ‘length_ratio’: 0.9230769230769231, ‘translation_length’: 12, ‘reference_length’: 13}

Sentence 7: {‘bleu’: 0.0, ‘precisions’: [0.21428571428571427, 0.0, 0.0, 0.0], ‘brevity_penalty’: 0.8668778997501817, ‘length_ratio’: 0.875, ‘translation_length’: 14, ‘reference_length’: 16}

Sentence 8: {‘bleu’: 0.0, ‘precisions’: [0.6, 0.2857142857142857, 0.07692307692307693, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.3636363636363635, ‘translation_length’: 15, ‘reference_length’: 11}

Sentence 9: {‘bleu’: 0.0, ‘precisions’: [0.5, 0.17647058823529413, 0.0625, 0.0], ‘brevity_penalty’: 0.9459594689067654, ‘length_ratio’: 0.9473684210526315, ‘translation_length’: 18, ‘reference_length’: 19}

Sentence 10: {‘bleu’: 0.0, ‘precisions’: [0.36363636363636365, 0.2, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.2222222222222223, ‘translation_length’: 11, ‘reference_length’: 9}

Sentence 11: {‘bleu’: 0.6703420896351792, ‘precisions’: [0.8461538461538461, 0.75, 0.6363636363636364, 0.5], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.0833333333333333, ‘translation_length’: 13, ‘reference_length’: 12}

Sentence 12: {‘bleu’: 0.3777331186826423, ‘precisions’: [0.7272727272727273, 0.5, 0.4444444444444444, 0.375], ‘brevity_penalty’: 0.7613003866968737, ‘length_ratio’: 0.7857142857142857, ‘translation_length’: 11, ‘reference_length’: 14}

Sentence 13: {‘bleu’: 0.0, ‘precisions’: [0.4, 0.25, 0.0, 0.0], ‘brevity_penalty’: 0.8187307530779819, ‘length_ratio’: 0.8333333333333334, ‘translation_length’: 5, ‘reference_length’: 6}

Sentence 14: {‘bleu’: 0.0, ‘precisions’: [0.25, 0.0, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.0, ‘translation_length’: 8, ‘reference_length’: 8}

Sentence 15: {‘bleu’: 0.0, ‘precisions’: [0.125, 0.0, 0.0, 0.0], ‘brevity_penalty’: 0.8824969025845955, ‘length_ratio’: 0.8888888888888888, ‘translation_length’: 8, ‘reference_length’: 9}

Sentence 16: {‘bleu’: 0.0, ‘precisions’: [0.2727272727272727, 0.0, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.0, ‘translation_length’: 11, ‘reference_length’: 11}

Sentence 17: {‘bleu’: 0.0, ‘precisions’: [0.0, 0.0, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.0, ‘translation_length’: 4, ‘reference_length’: 4}

Sentence 18: {‘bleu’: 0.0, ‘precisions’: [0.625, 0.2857142857142857, 0.0, 0.0], ‘brevity_penalty’: 1.0, ‘length_ratio’: 1.1428571428571428, ‘translation_length’: 8, ‘reference_length’: 7}

Sentence 19: {‘bleu’: 0.0, ‘precisions’: [0.3333333333333333, 0.2, 0.0, 0.0], ‘brevity_penalty’: 0.7165313105737893, ‘length_ratio’: 0.75, ‘translation_length’: 6, ‘reference_length’: 8}

Sentence 20: {‘bleu’: 0.0, ‘precisions’: [0.46153846153846156, 0.08333333333333333, 0.0, 0.0], ‘brevity_penalty’: 0.5404329964865341, ‘length_ratio’: 0.6190476190476191, ‘translation_length’: 13, ‘reference_length’: 21}

Appendix II
Related Table of the Data of the Study

Sentence 1	BLEU	Length ratio	Text
Source			The Kurdistan Region is known for having valleys, mountains, forests, as well as small and large rivers.
Reference	100.00	1.00	هەرێمی کوردستان بە هەبوونی دۆڵ، کوێستان، دارستان و هەروەها رووباری گەورە و بچووک ناسراوە.
Candidate	0.32	1.14	‘هەرێمی‘، ’کوردستان‘ بە هەبوونی دۆڵ و شاخ و دارستان و هەروەها ڕووباری بچووک و گەورە ناسراوە‘.
Sentence 2	BLEU	Length ratio	Text
Source			Tourists can enjoy mountain climbing at Halgurd, Pera Magrun, Korek, Bradost, Sheren, and Saffin mountains.
Reference	100.00	1.00	گەشتیاران دەتوانن چێژ لە شاخەوانی چیاکانی هەڵگورد، پیرە مەگرون ، کۆرەک ، برادۆست، شیرین و سەفین وەربگرن.
Candidate	0.15	1.4	گەشتیاران دەتوانن چێژ لە سەرکەوتن بەسەر شاخ وەربگرن لە هەڵگورد , پێرا شاخەکانی ماگرون و کۆرەک و برادۆست و شێرێن و سەفین
Sentence 3	BLEU	Length ratio	Text
Source			A state-of-the-art cable car has recently opened on Korek Mountain
Reference	100.00	1.00	بەم دواییانە تەلەکابینێیەکی سەردەمانە لەچیایی کۆرەک کردراوەتەوە .
Candidate	0.0	1.14	لەم دواییانەدا تەلەفریکێکی پێشکەوتوو لەسەر شاخی کۆڕەک کرایەوە
Sentence 4	BLEU	Length ratio	Text
Source			Takes visitors up to its peak
Reference	100.00	1.00	گەشتیاران دەباتە سەر لووتکەی چیا .
Candidate	0.0	0.8	سەردانکەران دەباتە لوتکەی خۆی
Sentence 5	BLEU	Length ratio	Text
Source			To enjoy wonderful views of the surrounding areas
Reference	100.00	1.00	بۆ ئەوەی چێژ وەربگرن لە دیمەنە سەرنجراکێشەکانی ئەم ناوچەیە و دەورووبەری
Candidate	0.0	0.81	بۆ ئەوەی چێژ لە دیمەنە نایابەکانی ناوچەکانی دەوروبەری وەربگرن.
Sentence 6	BLEU	Length ratio	Text
Source			Visitors can also take tours through the region’s valleys, rivers, and caves.
Reference	100.00	1.00	هەروەها گەشتیاران دەتوانن بە ناو دۆل و رووبار و ئەشکەوتەکانی ناوچەکەدا گەشت بکەن.
Candidate	0.0	0.92	هەروەها سەردانکەران دەتوانن گەشت بکەن بەناو دۆڵ و ڕووبار و ئەشکەوتەکانی ناوچەکەدا.
Sentence 7	BLEU	Length ratio	Text
Source			The Guardian ranked Kurdistan as its top international destination for adventure tourism in 2015.
Reference	100.00	1.00	رۆژنامەی گاردیەنی بریتانیایی ساڵی ٢٠١٥ لە ریزبەندی سالانەیدا کوردستانی وەک یەکەم ولات بۆ گەشتیاری سەرکێشی دەستنیشان کرد.
Candidate	0.0	0.87	ڕۆژنامەی گاردیانی بەریتانی لە ساڵی ٢٠١٥دا کوردستانی لە پلەی یەکەمی گەشتیاریی سەرگەرمیدا ڕیزبەندی کرد.
Sentence 8	BLEU	Length ratio	Text
Source			Religious tourists can visit temples, mosques, and churches in all governorates in the Kurdistan Region.
Reference	100.00	1.00	گەشتیارانی ئایینی لەسەرجەم پارێزگاکانی کوردستان دەتوانن سەردانی پەرستگا،مزگەوت و کڵێساکان بکەن.
Candidate	0.0	1.36	گەشتیارانی ئایینی دەتوانن سەردانی پەرستگا و مزگەوت و کڵێساکان بکەن لە سەرجەم پارێزگاکانی هەرێمی کوردستان.
Sentence 9	BLEU	Length ratio	Text
Source			The remnants of many traditional and locally made weapons can be seen in Kurdistan’s museums, including the Red Terror Museum in Sulaimani
Reference	100.00	1.00	پاشماوەی زۆریەک لە چەکە کۆنەکان کە لە ناوخۆی کوردستان درووستکراون لە مۆزەخانەکانی کوردستان وەک مۆزەخانەی تیرۆری سوور لەسلێمانی دەبینرێت.
Candidate	0.0	0.94	پاشماوەی زۆرێک لە چەکی نەریتی و دروستکراوی ناوخۆیی لە مۆزەخانەکانی کوردستاندا دەبینرێت، لەوانە مۆزەخانەی تیرۆری سوور لە سلێمانی .
Sentence 10	BLEU	Length ratio	Text
Source			The region’s best-known arts and crafts are carpets and other textiles.
Reference	100.00	1.00	فەرش و قوماش بەرچاوترین هونەر و پیشەسازییەی ئەم ناوچەیە.
Candidate	0.0	1.2	ناسراوترین هونەر و پیشەسازییەکانی ناوچەکە بریتین لە فەرش و قوماشی دیکە
Sentence 11	BLEU	Length ratio	Text
Source			These handicrafts can be seen and admired at cultural museums in the cities of Erbil, Sulaimani, Duhok, and Kalar.
Reference	100.00	1.00	ئەم کارە دەستیانە لە مۆزەخانە رەوشەنبیرییەکانی هەولێر، سلێمانی، دهۆک، و کەلار دەبینرێن.
Candidate	0.67	1.08	ئەم کارە دەستیانە لە مۆزەخانە ڕۆشنبیرییەکانی شارەکانی هەولێر، سلێمانی ، دهۆک، و کەلار دەبینرێن
Sentence 12	BLEU	Length ratio	Text
Source			There are more than 3,500 archaeological sites in the Kurdistan Region.
Reference	100.00	1.00	لەهەرێمی کوردستان زیاتر لە سێ هەزار و پێنج سەد شوێنی گرینگی گەشتیاری بوونیان هەیە.
Candidate	0.37	0.78	لە هەرێمی کوردستان زیاتر لە سێ هەزار و 500 شوێنی شوێنەواری هەیە

(*Contd*...)

Appendix II
(*Continued*)

Sentence 1	BLEU	Length ratio	Text
Sentence 13	BLEU	Length ratio	Text
Source			Some of them are significant in terms of tourism.
Reference	100.00	1.00	هەندێکیان لە رووی گەشتیاری گرینگیان هەیە.
Candidate	0.0	0.83	هەندێکیان لە ڕووی گەشتیارییەوە گرنگن.
Sentence 14	BLEU	Length ratio	Text
Source			Employment openings for the month totaled 10.72 million,
Reference	100.00	1.00	کۆی ١٠.٧٢ ملیۆن هەلی کار لەم مانگەدا بەردەست بووە.
Candidate	0.0	1.0	کۆی گشتی ژمارەی دامەزراندن بۆ مانگەکە 10.72 ملیۆن بووە،
Sentence 15	BLEU	Length ratio	Text
Source			well above the FactSet estimate of 9.85 million,
Reference	100.00	1.00	زیاتر لە ڕێژەی خەملێندراوی ٩.٨٥ هەلی کار لەلایان فاکتسێتە.
Candidate	0.0	0.88	زۆر زیاترە لە خەمڵاندنی فاکتسێت بۆ 9.85 ملیۆن کەس.
Sentence 16	BLEU	Length ratio	Text
Source			according to September’s Job Openings and Labor Turnover Survey.
Reference	100.00	1.00	بە پێی توێژینەوەی بەردەستبوون و ئالوگۆری هەلی کارلە مانگی ئەیلول دا
Candidate	0.0	1.00	بەپێی ڕاپرسیی کردنەوەی هەلی کار و گۆڕانی کار لە مانگی ئەیلولدا،
Sentence 17	BLEU	Length ratio	Text
Source			The data indicates
Reference	100.00	1.00	زانیارییەکان ئەمە دەخەنە روو.
Candidate	0.0	1.0	داتاکان ئاماژە بەوە دەکەن
Sentence 18	BLEU	Length ratio	Text
Source			there are 1.9 job openings for every available worker.
Reference	100.00	1.00	کە بۆ هەر کرێکارێک ١.٩ هەلی کار بەردەستە .
Candidate	0.0	1.14	کە 1.9 هەلی کار بۆ هەر کرێکارێکی بەردەست هەیە.
Sentence 19	BLEU	Length ratio	Text
Source			The ISM Manufacturing Index posted a 50.2 reading,
Reference	100.00	1.00	پێوەرەکانی بەرهەمهێنانی ئای ئێس ئێم نیشاندەری ٥٠.٢ دەرفەتی کارە .
Candidate	0.0	0.75	پێوەرەکانی بەرهەمهێنانی ISM خوێندنەوەی 50.2ی تۆمارکردووە،
Sentence 20	BLEU	Length ratio	Text
Source			slightly better than the Dow Jones estimate of 50.0 but 0.9 percentage points lower than September.
Reference	100.00	1.00	تا ڕادەیەک لە مەزەندەکانی داونجۆنز کە ٥٠.٠ پێشان دەدن بەرزترە بەڵام بە بەراوەرد لەگەڵ مانگی ئەیلول ئەم ڕێژەیە ٠.٩ لە سەد کەمترە .
Candidate	0.0	0.61	کەمێک باشترە لە خەمڵاندنی داو جۆنز بۆ 50.0 بەڵام 0.9خاڵ کەمترە لە مانگی ئەیلول.