Journal of Language and Translation
Volume 9, Number 3, Autumn 2019, (pp.43-55)

The Correlation of Machine Translation Evaluation Metrics with Human Judgement on Persian Language

Marziyeh Taleghani[1], Ehsan Pazouki[2*] and Vahid Ghahraman[3]
[1] MA in Translation Studies, Faculty of Persian Literature and Foreign Languages, South Tehran Branch of Azad University Iran
[2] Assistant Professor of Artificial Intelligence, Faculty of Computer Engineering, Shahid Rajaei Teacher Training University, Tehran, Iran
[3] Assistant Professor of TESOL, Iran Encyclopedia Compiling Foundation, Tehran, Iran

Received: 02 February 2019 Accepted: 23 October 2019

Abstract
Machine Translation Evaluation Metrics (MTEMs) are the central core of Machine Translation (MT) engines as they are developed based on frequent evaluation. Although MTEMs are widespread today, their validity and quality for many languages is still under question. The aim of this research study was to examine the validity and assess the quality of MTEMs from Lexical Similarity set on machine translated Persian texts. This study focused on answering three main questions, which included the extent that Automatic Machine Translation Evaluation Metrics is valid on evaluating translated Persian texts; the probable significant correlation between human evaluation and automatic evaluation metrics in evaluating English to Persian translations; and the best predictor of human judgment. For this purpose, a dataset containing 200 English sentences and their four reference human translations, was translated using four different statistical Machine translation systems. The results of these systems were evaluated by seven automatic MTEMs and three human evaluators. Then the correlations of metrics and human evaluators were calculated using both Spearman and Kendall correlation coefficients. The result of the study confirmed the relatively high correlation of MTEMs with human evaluation on Persian language where GTM proved to be more efficient compared with other metrics.

Keywords: Lexical Similarity; MTEM; Statistical Machine Translation

INTRODUCTION

Machine Translation (MT) is a relatively new field in translation studies. Although it has made a great progress since its creation and different Machine translation systems, like google translate (one of the most known systems ever), have been presented in this field, there is still a long way ahead. Evaluation is the central core of MT systems, they are developed based on that, so in order to be improved these systems need to be evaluated. Human evaluation referred to as subjective evaluation, like in many other fields, is considered as the first approach to MT evaluation. As it is clear from the approach's name, subjective evaluation, the first problem of this approach is that it is subjective which reduces the

*Corresponding Author's Email:
Ehsan.pazouki@sru.ac.ir

reliability of the evaluation, it is also time consuming and expensive. Due to huge corpora containing thousands of pages and millions of words which need to be assessed over and over, using this approach in this field is in fact impossible.

Automatic MT evaluation approach is presented to reduce problems of the subjective evaluation. In this approach it is the machine, not the human, which evaluates the machine translated texts using fixed metrics. Since human is expensive, automatic methods became popular. In order to use this approach there is a need for a set of reference translations of the source text, and also a similarity metric to measure sentence closeness, between the candidate sentence and its set of references (Papineni, Roukos, Ward, & Wei, 2002). There has always been a concern about this approach, whether automatic measures correlate well with human judgement or not.

As Bouamor et.al. state "evaluation of Machine Translation continues to be a challenging research problem. There is an ongoing effort in finding simple and scalable metrics with rich linguistic analysis" (Bouamor, Alshikhabobak, Mohit, & Oflazer, 2014, p.1).

Since creation of the first MTEM, BLEU (Papineni et al., 2002), a wide range of metrics have been proposed and evaluated. Taking a Look at the background of these metrics, it is evident that they are mostly proposed based on European languages' criteria, especially English, and then adapted to other languages, if needed. So their validity for many languages is still under question. Since evaluation is the base of MT training process, as a result of lack of localized MTEM on less focused languages like Persian, MT systems seems inefficient on these languages while they show great results on languages that evaluation metrics are developed based on them.

In order to clarify these issue and improve the performance of MTEMs on less focused languages, they need to be evaluated. These metrics are usually evaluated based on their correlation with human judgments on a set of MT output (Bouamor et al., 2014). In recent years, many works are done, not only focusing on European languages (Callison-Burch, Osborne, & Koehn, 2006) (Agarwal & Lavie, 2008) but also on other languages like Arabic and Chinese (Bouamor et al., 2014) (Dreyer & Marcu, 2012); but to the best of our knowledge, there are not many, if don't say none, works focusing on Persian language in this field.

The aim of this research was to assess the quality of MTEMs on translated texts from English to Persian and to show whether they are valid or not. The rest of this paper is organized as follow: the methodology of the study including data collection and analysis is presented in Methodology. The results of the study are released in Results and the conclusion of the study besides Suggestions for further research are presented in Conclusion.

METHODS

Evaluation Metrics

MTEMs of lexical similarity set are divided into different categories including "Lexical Precision", "F-measure" and "Edit Distance" (Giménez & Màrquez, 2010). In this research the aim was to analyze the most common metrics of these categories which are presented in

Dataset

As automatic machine translation evaluation metrics work based on aligned parallel corpora (bilingual corpuses containing sentences in a language and their equivalent translations in another language) and are mostly programmed to get better results in presence of more than one reference (translation text) for a source (original text), researcher attempted to find a bilingual, multi references corpus. The dataset to which researcher got access was a corpus of 200 English sentences from the books "The Kite Runner" by Khaled Hosseini (an average English text) and "A Tale of Two Cities" by Charles Dickens (a relatively difficult English text), 105 sentences from the former and 95 sentences from the later, and its four aligned Persian translation references. The dataset was made by a PHD student of Tehran University computer engineering faculty (Farzi & Faili, 2015). Table 2 and Table 3 presents some

examples of the English source sentences and their aligned Persian references.

Questionnaire

Correlation with human evaluation is the measure to evaluate the validity and quality of MTEMs, therefore human evaluation was used as the touchstone of this study. Since experts were going to rank translated sentences by four different machine translation systems during this research, there was a need for a questionnaire which facilitated the process of ranking.

The questionnaire contained 200 source sentences of the corpus besides four machine translated translations of each sentence. In order to facilitate the process, researcher made online questionnaires using Google Forms till participants be able to rank the translated sentences without being limited to time or place. In addition, since according to Google form rule, answering all questions of a single form before submitting that is necessary, researcher divided them by ten and made 20 forms of 10 sentences till evaluators be able to answer each form in less than 15 minutes.

As it is shown in **Error! Reference source not found.**, on the top of each form, researcher gave the instructions needed to answer them, in Persian, to help evaluators.

Table 1.
List of Evaluation Metrics

METRIC	CATEGORY	FEATURE
BLEU (Papineni et al., 2002)	Lexical Precision	Based on average of matching n-grams between candidate and reference
NIST (Doddington, 2002)	Lexical Precision	Calculate matched n-grams of sentences and attach different weights to them
METEOR (Banerjee & Lavie, 2005)	F-measure	Based on various modules (Exact Match, Stem Match, Synonym Match and POS Tagger)
GTM (Turian, Shen, & Melamed, 2003)	F-measure	Computes precision recall and F-measure in terms of maximum unigram matches.
TER (Snover, Dorr, Schwart, Micciulla, & Makhoul, 2006)	Edit Distance	Computes the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references
WER (Nießen, Och, Leusch, Ney, & Informatik, 2000)	Edit Distance	Based on the Levenshtein distance: the minimum number of substitutions, deletions and insertions that have to be performed to convert the generated text into the reference text
PER (Tillmann, Vogel, Ney, Zubiaga, & Sawaf, 1997)	Edit Distance	Same as WER but compares the words in the two sentences without taking the word order into account

Machine Translation systems

In order to find the ultimate online translators, researcher tried the list of online translators presented in machine translation page of Wikipedia ("Machine translation," 2015). The list consists of fifteen online translators among which Babylon, translation.babylon.com, was filtered so unavailable.

Since the purpose was to translate the source text from English to Persian the translators thatdid not have the possibility of translating into Persian were crossed out and the list ended up in 8 online translators.

As researcher was looking for translators that were able to translate long texts, in the next step she crossed out those that had limitations for the number of the words in a text and got the list of five online translators.

Further checks showed that www.bing.com and www.freetranslations.org use the same datasets and as a result their translations are exactly

the same so researcher chose one of them, www.bing.com, and came to the final list of four translators (http://translate.google.com, http://www.bing.com/translator, http://www.freetranslation.com and http://www.targoman.com).

Table 2.
First example of the aligned source and reference sentences in dataset

Source	He took a deep breath and sipped his tea.
Reference1	او نفس عمیقی کشید و چایش را سر کشید.
Reference2	او نفس عمیقی کشید و چایش را سر کشید .
Reference3	او یک نفس عمیق کشید و چایش را چشید.
Reference4	نفس عمیقی کشید و جرعه‌ای از چایش نوشید.

Table 3.
Second example of the aligned source and reference sentences in dataset

Source	The gentleman had left London.
Reference1	مرد شریف لندن را ترک کرده بود.
Reference2	مرد اصیل لندن را ترک کرده بود.
Reference3	اقا لندن را ترک کرده بود.
Reference4	این اقای محترم لندن را ترک گفته بودند.

Participants

In order to get reliable results it was decided that three evaluators rank machine translated translations of each source sentence. As ranking four machine translated translation of 200 sentences is a laborious and time-consuming task, it was set in a way that each evaluator ranks translations of only 40 source sentences (four Google forms); in other words 5 evaluators were assigned to evaluate the source sentences each time. And as a result of that researcher wanted each sentence to be evaluated three times, 15 human evaluators were needed to conduct this research. 15 participants who were native speakers of Persian with English as their foreign language were chosen as the evaluators. 11 of the participants were MA students of English translation, two of them had MA degree in English literature and two of them had MS degree in other fields. All of the participants had translation experience, seven of which had translation as their profession.

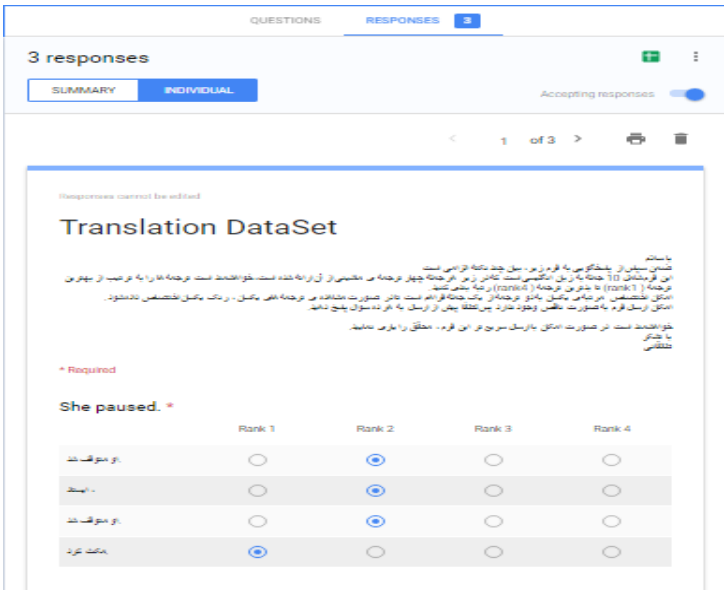

Figure 1. Sample of Google form Participants were asked to rank readability and fluency of the four machine translated translations of each source sentence.

Data collection and analysis

In order to conduct this research, researcher needed to set the data obtained from human evaluators and MTEMs in a way that accomplish the correlation coefficient calculating input's criteria. In order to do so first the corpus including 200 English sentences was translated by four different machine translation systems, Bing, Free translation, Google Translate and Targoman into Persian (the process of choosing these systems is explained in detail in subsection 2.3).

Then the output of these machine translation systems for each sentence was evaluated by three individual human evaluators, who were native speakers of Persian language and had English as their foreign language, translations were ranked from the best (rank 1) to the worst (rank 4) while ties were allowed. One example of these forms is shown in

Then the output of machine translation systems were once again evaluated by seven MTEMs, BLEU, NIST (Doddington, 2002), METEOR (Banerjee & Lavie, 2005), GTM (Turian et al., 2003), TER (Snover et al., 2006), WER (Nießen et al., 2000), and PER (Tillmann et al., 1997), using Asiya[1] tool, an open toolkit aimed at covering the evaluation needs of system and metric developers along the development cycle (Giménez & Màrquez, 2010), examples of these results are presented in

In this step researcher had two sets of data needed as input of Spearman and Kendall correlation coefficients to calculate the correlation between human evaluators and Evaluation Metrics. MATLAB, a multi-paradigm numerical computing environment and fourth-generation programming language developed by MathWorks[2] ("MATLAB," 2017), is the tool which was used in this research for computing these correlation coefficients. The correlation between automatic evaluation and human evaluation at sentence level was obtained following the practice of Agarwal & Lavie (2008).

Table 4.
List of Ultimate Translators

Primary list of online translators	English to Persian translators	Unlimited translators	Final list
http://translate.google.com/	✓	✓	✓
http://translate.reference.com/	✓		
http://translation.babylon.com/	×	×	×
http://transsoftware.info/scripts/webtrans2.dll		✓	
http://turkceingilizce.ingilizceturkce.gen.tr/		✓	
http://www.bing.com/translator	✓	✓	✓
http://www.englishdictionaryonline.org/		✓	
http://www.freetranslation.com/	✓	✓	✓
http://www.freetranslations.org/	✓	✓	✓
http://www.ingilizceceviri.org/	✓		
http://www.reverso.net/		✓	
http://www.spanishenglish.com/		✓	
http://turkce.cevirsozluk.com/	✓		
http://www.systranet.com/translate/		✓	
http://www.targoman.com	✓	✓	✓

1.http://www.lsi.upc.edu/~nlp/Asiya
2. https://www.mathworks.com

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

ref.sgm per.report

```
SET          DOC          SEG METRIC          1_bing.txt     2_free.txt    3_google.txt 4_targoman.txt
=============================================================================================================
UNKNOWN_SET UNKNOWN_DOC 1    -PER          -0.65000000    -0.50000000    -0.57142857    -0.55000000
UNKNOWN_SET UNKNOWN_DOC 2    -PER          -0.57142857    -0.88888889    -0.51851852    -0.52380952
UNKNOWN_SET UNKNOWN_DOC 3    -PER          -0.50000000    -0.66666667    -0.40909091    -0.21428571
UNKNOWN_SET UNKNOWN_DOC 4    -PER          -0.55555556    -0.57142857    -0.77777778    -0.55555556
UNKNOWN_SET UNKNOWN_DOC 5    -PER          -0.22222222    -0.62500000    -0.22222222    -0.11111111
UNKNOWN_SET UNKNOWN_DOC 6    -PER          -0.57142857    -0.66666667    -0.52380952    -0.52380952
UNKNOWN_SET UNKNOWN_DOC 7    -PER          -0.76190476    -0.76190476    -0.80952381    -0.38095238
UNKNOWN_SET UNKNOWN_DOC 8    -PER          -0.61538462    -0.84210526    -0.76923077    -0.72222222
UNKNOWN_SET UNKNOWN_DOC 9    -PER          -0.56521739    -0.62500000    -0.45833333    -0.30000000
UNKNOWN_SET UNKNOWN_DOC 10   -PER          -0.66666667    -0.53333333    -0.60000000    -0.46153846
UNKNOWN_SET UNKNOWN_DOC 11   -PER          -0.50000000    -0.71428571    -0.64285714    -0.28571429
UNKNOWN_SET UNKNOWN_DOC 12   -PER          -0.69230769    -0.69230769    -0.61538462    -0.60000000
UNKNOWN_SET UNKNOWN_DOC 13   -PER          -0.66666667    -1.00000000    -0.66666667    -0.44444444
UNKNOWN_SET UNKNOWN_DOC 14   -PER          -0.53333333    -0.80000000    -0.66666667    -0.33333333
UNKNOWN_SET UNKNOWN_DOC 15   -PER          -0.86666667    -0.68750000    -0.93333333    -0.33333333
UNKNOWN_SET UNKNOWN_DOC 16   -PER          -0.83333333    -0.85714286    -0.62500000    -0.62500000
UNKNOWN_SET UNKNOWN_DOC 17   -PER          -0.65217391    -0.73913043    -0.65217391    -0.52941176
```

Figure 2. Examples of the per metric evaluation results

As it is shown in Figure 2 the results of the MTEMs extracted by Asiya tool are scores allocated to each translation and not ranks. These scores are transformed in to ranks during the correlation calculating process by the correlation coefficients in order to make comparison of the two sets of evaluation results possible. For this purpose, for each sentence the correlation coefficients set the evaluation results of human evaluators in a four three matrix and also the evaluation results of the MTEMs in a four seven matrix. Then the correlation of these two matrix for each sentence is calculated, the result which is a three seven matrix shows the concordance between each human translator and each Evaluation Metric. Each matrix includes numbers between (-1) to (1), where the bigger the number is, the concordance between the evaluators is more so the correlation is higher.

Then the average concordance of three evaluators with each metric is calculated and presented in a one seven matrix these are the final results for each sentence. Figure 3 presents results extracted from these matrix for all 200 sentences calculated by both Spearman and Kendall correlation coefficients.

	WER		TER		PER		METEOR		GTM		BLEU		NIST	
Sentense	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall
1	0.383187033	0.38838	-0.03153	0.06086	0.094591	0.161364	0.33694	0.28308	0.516228	0.526513	0.516228	0.526513	0.516228	0.526513
2	0.916227766	0.85985	0.916228	0.85985	0.610819	0.515908	0.916228	0.859846	0.916228	0.859846	0.916228	0.859846	0.916228	0.859846
3	0.849561099	0.74873	0.849561	0.74873	0.877485	0.849241	0.877485	0.849241	0.849561	0.748735	0.849561	0.748735	0.877485	0.849241
4	0.28685294	0.23299	0.210819	0.18257	0.286853	0.232991	0.238799	0.157135	-0.18306	-0.13807	0.305466	0.268246	0.383064	0.360293
5	0.808137624	0.66944	0.808138	0.66944	0.611111	0.533333	0.808138	0.669439	0.808138	0.669439	0.851852	0.733333	0.851852	0.733333
6	0.172075922	0.06086	0.32193	0.18838	0.62344	0.498482	0.744152	0.627019	0.782894	0.748735	0.782894	0.748735	0.744152	0.627019

Figure 3. The correlation between MTEMs and three human evaluators for all sentences calculated by Spearman and Kendall coefficients

At last the average Spearman and Kendall correlation coefficient for all 200 sentences was calculated and the final result was obtained which are presented in Results.

RESULTS

In order to check the validity of human evaluator's answers the inter-annotator agreement was calculated using both Spearman and Kendall correlation coefficients. Figure 4 illustrates this agreement.

According to Figure 4, the inter annotator correlation between the experts in this study was more than 0.40 where the first and second experts showed the highest correlation, more than 0.45 while the second and third experts had the lowest correlation.

Researcher also calculated the correlation between each evaluation metric and each expert for all sentences based on two correlation coefficients. Obtained results are presented in Figure 5 and Figure 6.

Taking a look at Figure 5 and Figure 6, MTEMs have a relatively high correlation with experts one and two, more than 0.43, while their correlation with expert three is far lower, around 0.20.

At last the mean of each Spearman and Kendall correlation coefficients for all evaluators was calculated, and the final result was obtained which are presented in Figure 7.

As it is shown in Figure 7, although there is a shade of difference between the two correlation coefficients results, they ranked the metrics in the same order. Generally speaking,

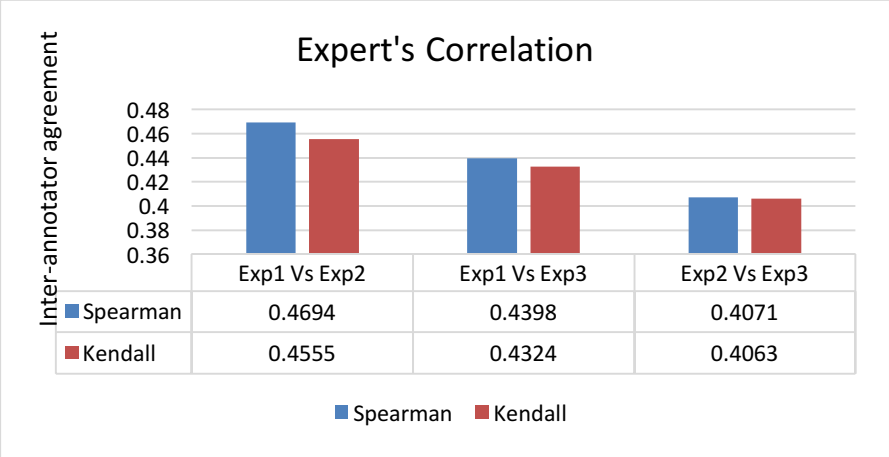

Figure 4. Inter-annotator agreement

MTEMs showed acceptable correlation with human judgements, more than 0.37 based on Spearman and more than 0.33 based on Kendall correlation coefficient. The GTM metric from F-measure category has the highest correlation with human evaluators on Persian language based on both Spearman and Kendall correlation coefficients. Then there are BLEU and NIST both from the Lexical Precision category in the second and third place. Next there are, respectively, WER from Edit Distance category and METEOR from F-measure category in next places. And finally the weakest results are obtained from TER and PER from Edit Distance category.

For addressing the validity questions of this study we must again take a look at Figure 7, as mentioned before correlation is a number between -1 and 1. According to Figure 7, the correlation of these examined metrics at worst is 0.3393 which means these metrics at worst have

more than 33% positive correlation with human evaluation. In other words, although the metrics still has a long way a head they can be considered valid when it comes to assess Machine translated Persian texts.

Figure 8 illustrates the percentage of the distribution of correlation for each metric. As it is shown in these diagrams the distribution of correlation of more than 0.6 (in other words +60%) for each metric is more than 45% and reach 52% for GTM metric which is an acceptable result.

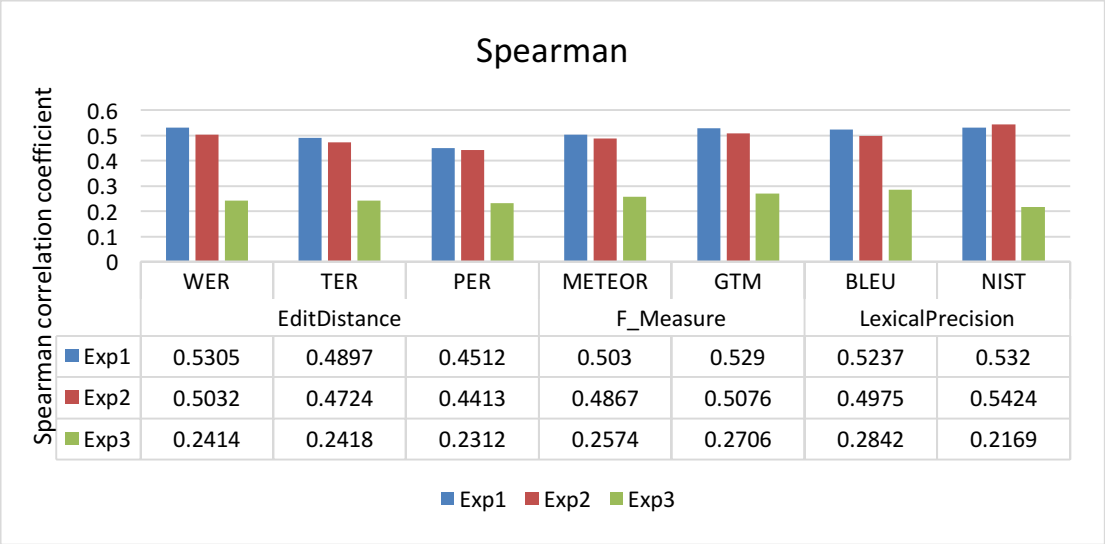

Figure 5. Correlation between each evaluation metric and each expert for all sentences based on Spearman correlation coefficient

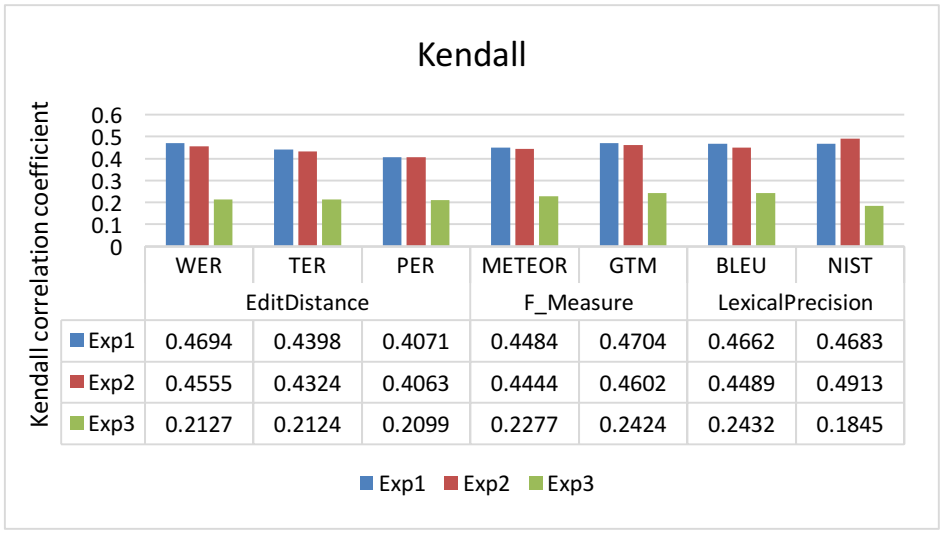

Figure 6. Correlation between each evaluation metric and each expert for all sentences based on Kendall correlation coefficient

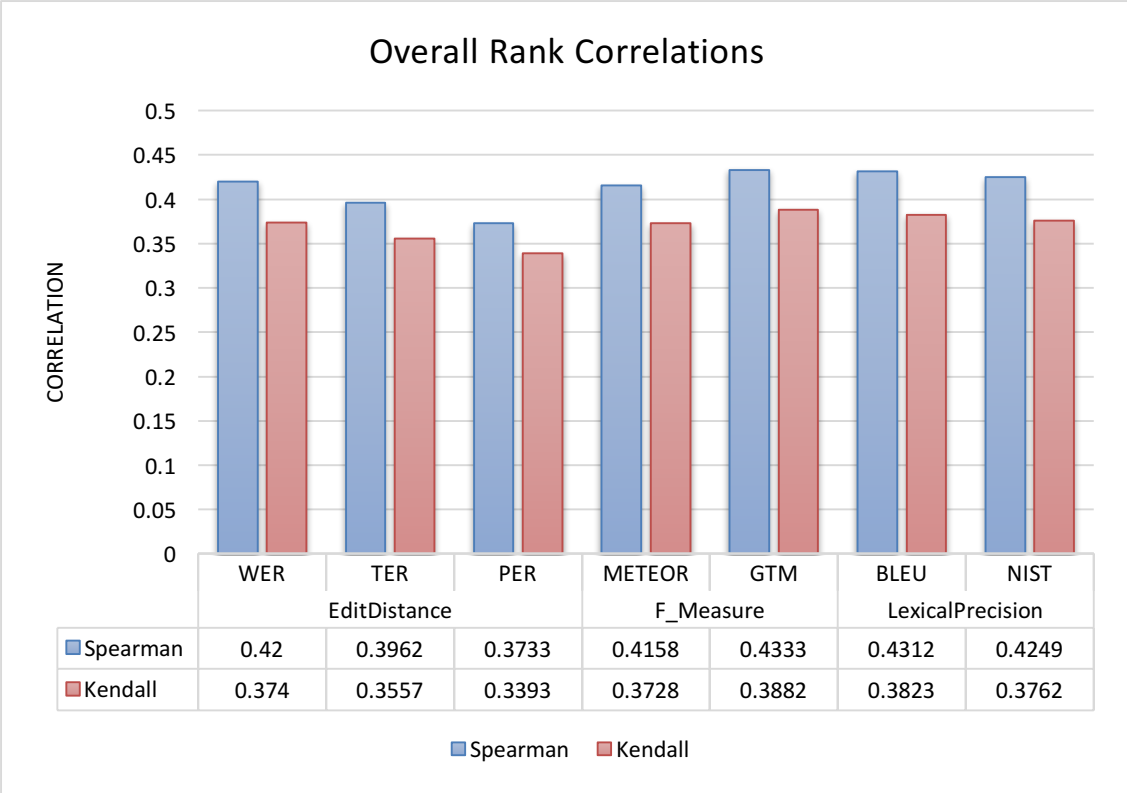

Figure 7. The mean of Spearman and Kendall correlation coefficients for all evaluators

The results of turian et al. (Turian et al., 2003) work to some extent confirm the present study's result when they state that machine translation can be evaluated using well-known evaluation measures. They add that in particular, on the data they have used, the F-measure (here refers to GTM) proved significantly more reliable than the BLEU and NIST measures.

Yanli Sun is another researcher that in his work, "Mining the Correlation between Human and Automatic Evaluation at Sentence Level" (Sun, 2010), works on the correlation of automatic evaluation with human evaluation comparing Chinese translations of different MT using Spearman correlation coefficient (ρ).

As Sun reported in his work GTM correlates better with human evaluation than BLEU and TER at sentence level in Chinese output evaluation, his findings are presented in

Table 5. He also refers to similar findings which have been reported by Cahill (2009) in German evaluation which compared 6 metrics including the three metrics used in Sun's paper.

Kalyani and his friends (Kalyani, Kumud, Singh, & Kumar, 2014) also conducted a similar research on the Quality of MT Systems for Hindi to English Translation using some of MTEMs and compared them with human evaluation. The result of their work presents METEOR and GTM as metrics with the most correlation with human.

The results of our research somehow disagree Callison-Burch and his co- researchers' (Callison-Burch, Fordyce, Koehn, Monz, & Schroeder, 2007) as they state that METEOR excels other metrics where in next steps are BLEU, GTM, TER and WER respectively. Different target language can be considered as one reason of this disagreement. They worked on translated English texts from different European languages like French and German.

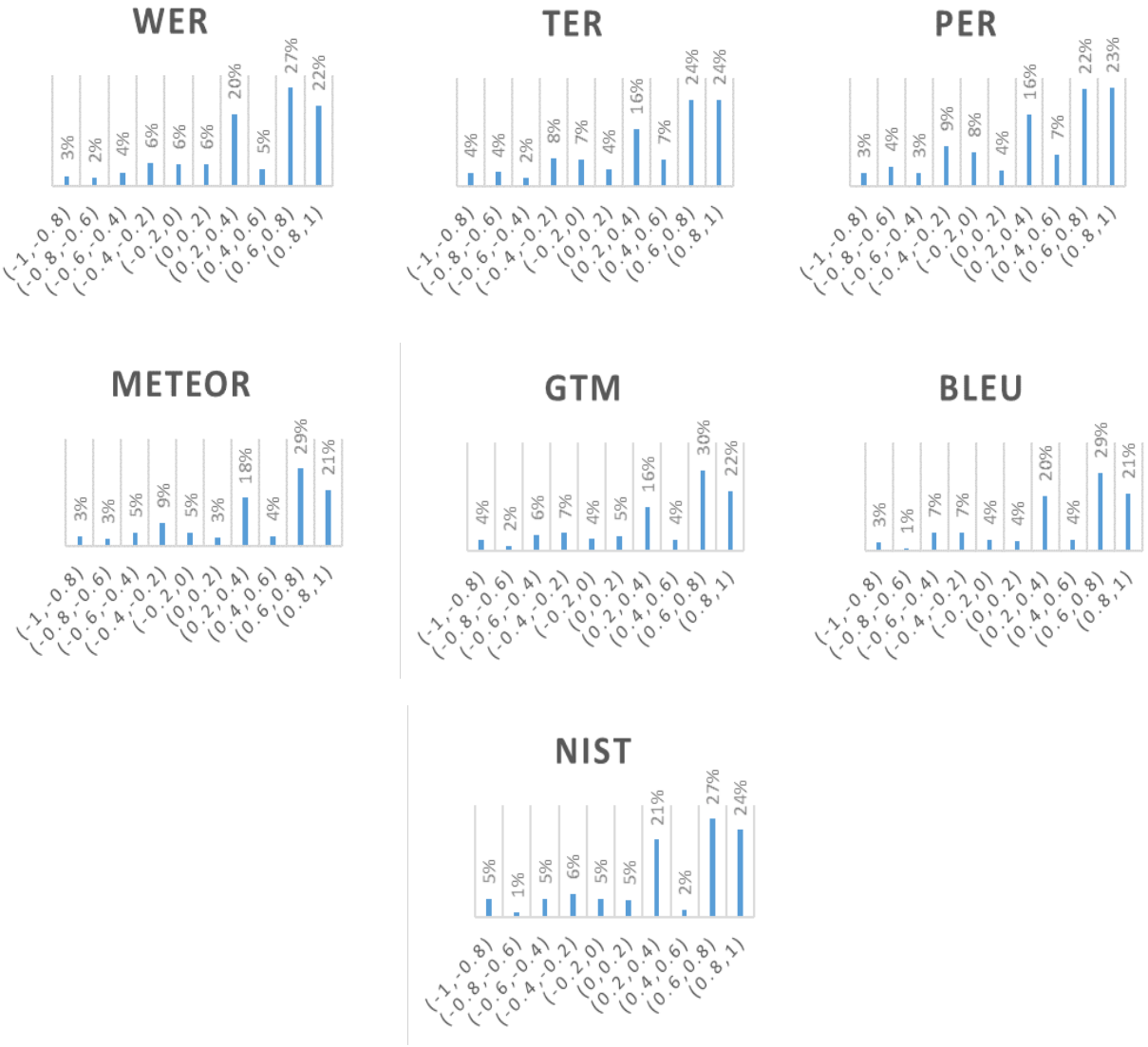

Figure 8. Distribution of correlation for 7 MTEMs. The x-axis shows the metrics' correlation coefficient with human judgement while the y-axis presents the percentage of distribution of correlation coefficients.

Table 5.
Spearman Correlation between Automatic and Human Evaluation as reported in sun's work

	L1	L2	L3	L4	Average
GTM	0.32	0.50	0.14	0.26	0.30
TER	0.33	0.48	0.12	0.24	0.29
BLEU	0.34	0.44	0.13	0.26	0.29

CONCLUSION

Our most important finding is that, even though automatic MTEMs are still far from being able to replace human judgment, they have made a great progress even when utilizing them on less focused languages like Persian. Comparing the result of the study with the previous ones shows that the range of correlation among less focused languages is the same (Sun, 2010) but it is far from the result gathered on researches on more focused languages (Callison-Burch et al., 2007) which shows the long way ahead. It can be concluded that machine translated Persian texts can be evaluated using well-known evaluation measures in future. So they are valid on Persian but they need to be developed based on this language features. In particular, on general Persian texts, GTM proved more reliable than BLEU and NIST measures. As a result when it comes to evaluate machine translated Persian texts, GTM metric could be considered as the best choice.

Since the MTEMs have found wide application on evaluating translated Persian texts in recent years, especially in academic researches (pilevar & Faili, 2010), (Ansari, Sadreddini, Tabebordbar, & WALLACE, 2014), there seemed to be a need for a study that concern their validity and priority on Persian language.

The result of the study also indicate that Iranian researchers must focus on developing localized metrics based on Persian language features and structure with higher performance on Persian language. Development of Such a metric can cause and facilitate developments of new brand Translation engines focusing on Persian languages and help flourishing previous ones, like Targoman. Although Bleu is the first metric developed, the results of this study shows that it is better that Iranian researches focus on GTM structure for developing a localized metric as it shows better performance on evaluating Persian.

There are more MTEMs and other ways to ascertain the reliability of these metrics. More data and more rigorous analysis is necessary to conform the results of this study. Therefore, a number of suggestion for eager researchers are put forward next:

• This research is a scalable work therefore, developing the study on a wider dataset with a wider number of evaluators in order to conform the reliability of the study can be a fruitful area for research.

• What works on one corpus might not work on another, so the future studies could take place on different corpora from different kind of texts.

• There are numerous MTEMs from different sets and categories that working on all of them in one study is impossible so there is a need for further works on other metrics.

References

Agarwal, A., & Lavie, A. (2008). METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output (pp. 115–118). Presented at the Third Workshop on Statistical Machine Translation, Columbus.

Ansari, E., Sadreddini, M. H., Tabebordbar, A., & WALLACE, R. (2014). Extracting Persian-English parallel sentences from document level aligned comparable corpus using bi-directional translation. *Advances in Computer Science: An International Journal*, *3*(5), 59–65.

Banerjee, S., & Lavie, A. (2005). 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). Michigan.

Bouamor, H., Alshikhabobak, H., Mohit, B., & Oflazer, K. (2014). A human judgment corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 207–213). Doha, Qatar.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136–158). Stroudsburg, PA, USA: Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *In Proceedings of EACL-2006*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence Statistics. In *HLT '02 Proceedings of the second international conference on Human Language Technology Research* (pp. 138–145). Morgan Kaufmann Publishers Inc.

Dreyer, M., & Marcu, D. (2012). HyTER: Meaning-Equivalent Semantics for Translation Evaluation (pp. 162–171). Presented at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Farzi, S., & Faili, H. (2015). A swarm-inspired re-ranker system for statistical machine translation. *Computer Speech & Language*, *29*(1), 45–62.

Giménez, J., & Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, *94*, 77–86.

Kalyani, A., Kumud, H., Singh, S. P., & Kumar, A. (2014). Assessing the Quality of MT Systems for Hindi to English Translation. *International Journal of Computer Applications*, *89*(15), 41–45.

Machine translation. (2015, September 6). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Machine_translation&oldid=679685135

MATLAB. (2017, January 17). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=MATLAB&oldid=760467403

Nießen, S., Och, F. J., Leusch, G., Ney, H., & Informatik, L. F. (2000). A Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*.

Papineni, K., Roukos, S., Ward, T., & Wei, J. Z. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association forComputational Linguistics (ACL)* (pp. 311–318). Philadelphia.

pilevar, M. T., & Faili, H. (2010). Persian SMT: A first attempt to English-Persian statistical machine translation. In *JADT 2010: 10th international conference on statistical analysis of textual data*.

Snover, M., Dorr, B., Schwart, R., Micciulla, L., & Makhoul, M. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *In Proceedings of Association for Machine Translation in the Americas* (pp. 223–231).

Sun, Y. (2010). Mining the Correlation between Human and Automatic Evaluation at Sentence Level. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. In *Proceedings of European Conference on Speech Communication and Technology*. Rhodes, Greece.

Turian, J., Shen, L., & Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. In *In Proceedings of MT Summit IX* (pp. 386–393).

Biodata

Ms Marziyeh Taleghani received her master's degree in the Faculty of Persian Literature and Foreign Languages, South Tehran Branch of Azad University, Iran. Her current research interests include machine translation.
Email: taleghani.marziyeh89@gmail.com

Dr Ehsan Pazouki received his PhD in Artificial Intelligence from Amirkabir university of Technology. He is an assistant professor of Shahid Rajaei Teacher Training University. His current research interests include Big Data with special focus on wide area surveillance and machine translation.
Email: ehsan.pazouki@sru.ac.ir

Dr Vahid Ghahraman is assistant professor of applied linguistics / TESL at Iran Encyclopedia Compiling Foundation, Tehran, Iran. His areas of interest include teaching EFL language skills cross-cultural pragmatics, and cultural translation.
Email: ghahraman@iecf.ir