JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics

Somayyeh Karami[*]
PhD Candidate in Translation, Department of Foreign Languages, University of Isfahan, Isfahan, Iran
Email: s.karami@fgn.ui.ac.ir

Dariush Nejadansari[**]
(corresponding author)
Assistant Professor in Teaching English Language, Department of English Language and Literature, University of Isfahan, Isfahan, Iran
Email: d.nejadansari@fgn.ui.ac.ir

Akbar Hesabi[***]
Assistant Professor in General Linguistics, Department of English Language and Literature, University of Isfahan, Isfahan, Iran
Email: a.hesabi@fgn.ui.ac.ir

ABSTRACT

Considering the costly nature of translation quality assessment in terms of time, money and energy, it seems logical to benefit from the modern technologies that are introduced in the field of machine translation (MT). Automated Translation Quality Evaluation Understudy Metrics (ATQEUMs) are one of these technologies that have revealed a promising application in assessing the MT output quality. This study, however, attempts to examine the reliability of the scores provided by the lexical ATQEUMs to human translated texts (i.e. the ones provided by 51 senior students of translator training programs in Iran) using 1, 2, …, 5 reference translations successively and separately. To this end, an empirical applied study is conducted following a quantitative approach to assess the reliability of the lexical ATQEUMs' scores in comparison to the expert scorers' scores. The higher the correlation between the sets of scores (in different stages of using 1, 2, …, 5 reference translations), the higher the reliability is interpreted to be. The results of the Pearson correlation coefficient analysis revealed that using 5 reference translations had led to the highest correlations in 37.80% of cases, which is more than the number for any other situation considered (i.e. using 4 reference translations (3.65%), 3 reference translations (10.97%), 2 reference translations (31.70%), and 1 reference translation (15.85%)). However, using 2 reference translations achieved the second position in having the highest correlations which contradicted the hypothesis that more reference translations would

ARTICLE INFO
Article history:
Received:
5th, September, 2020
Accepted:
6th, October, 2020
Available online:
Autumn 2020

Keywords:
Translation quality assessment, Lexical automated translation quality evaluation understudy metrics, Automated scoring, Reliability, Reference

Karami, Somayyeh, Nejadansari, Dariush, Hesabi, Akbar (2020). Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics. *Journal of Foreign Language Research*, 10 (3), 618-629.
DOI: 10.22059/jflr.2020.309025.751

[*] Somayyeh Karami is a PhD. candidate in Translation at the University of Isfahan. She is interested in all areas related to technology in translation and its quality evaluation.

[**] Dariush Nejadansari is an associate professor in Teaching English as a Foreign Language in the Department of English Language and Literature at the University of Isfahan.

[***] Akbar Hesabi is an associate professor in Linguistics in the Department of English Language and Literature at the University of Isfahan.

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 3, Autumn 2020, Page 618 to 629

1. Introduction

As an integral part of the education, evaluation refers to the process of gathering information about the learner to decide upon his/her progress (Cheng, Rogers, & Hu in Beikmohammadi, Alavi, & Kaivanpanah, 2020). However, it is indeed an under-researched and under-discussed area in the field of translation studies. Bowker, (2001) summarizes the reason for this relative neglect as "the problems associated with the evaluation of translated texts are of cosmic proportions" which make it an arduous and challenging task. "The main difficulty surrounding translation evaluation is its subjective nature: the notion of quality has very fuzzy and shifting boundaries" (Bowker, 2001, p. 347). Thus, a translation may be deemed appropriate, acceptable, or totally unacceptable in various circumstances even when being evaluated by the same evaluator (Vanderschelden, 2000 in Olohan, 2004), since evaluators conventionally rely on their unverified intuition, anecdotal evidence or small samples, dictionaries, printed parallel texts, subject field experts which are not always highly conducive to an objective evaluation (Bowker, 2001). However, as House (1998) puts it, translation evaluation needs to develop beyond subjective, one- sided or dogmatic judgments using large-scale empirical studies to posit intersubjectively verifiable evaluation criteria, i.e. it needs to move towards being a more objective task (Bowker, 2001). Looking from another angle, Màrquez (2013) truly argues that the TQA task is too costly in terms of time, energy, and money. Human expert evaluators are required to spend an undeniably long time and great energy to assess and score the translations whose result is, almost always, not appreciated or even accepted as being reliable.

However, there is the potential of Automated Translation Quality Evaluation Understudy Metrics (ATQEUMs) which can remarkably attenuate the time and energy that is to be spent in the quality assessment process of such complicated skills as translation (Weigle, 2011). The researchers of the field of automated scoring have, in most cases, simply compared the automated scores of machine translations to the scores provided by human scorers to the same translated texts using just one reference translation, and have influentially concluded that scores provided by these ATQEUMs are at least as reliable as the scores given by human scorers (e.g. Papineni et al. 2002; Doddington, 2002; Banerjee & Lavie, 2005). In contrast, this research is an attempt to take advantage of this development in the field of machine translation evaluation to score the human translated texts. However, considering the creative nature of human translators' mind in such processes as problem solving and decision making of the translation act, it does not seem logical to base the whole assessment process on just one reference translation as the benchmark to decide upon. Therefore, this study tries to discover the difference that is created by increasing the number of reference translations to five, one by one, to encompass all (or at least most) of the possible equivalents that are deemed as appropriate in a particular context.

2. Literature Review

Automated Translation Quality Evaluation Understudy Metrics

The automatic metrics can be distinguished based on their scope, i.e. that particular dimension of texts that they focus on. The present research study has concentrated on those metrics which limit their scope to lexical dimension (or n-gram based metrics) that actually constitute the dominant approach to automatic TQA today. Apart from the technique(s) that they apply to

calculate the score, these metrics all follow the same rule: They reward lexical similarities (n-gram matchings) among the candidate translations and a set of reference translations (i.e. human quality translations or gold translations). In the following, the metrics that have been used in the course of the present study are briefly introduced based on the technique(s) they apply to calculate the score that defines the quality of candidate translations.

Edit Distance Measures

The edit distance measures (Levenshtein, 1966), provide an estimation of translation quality based on the minimum number of changes that are required to convert the candidate translations into a reference translation. These changes include all the necessary additions, deletions and substitutions of a word by another one in translation evaluation. Such implementations in the field of MT evaluation include:

- WER (Niessen et al., 2000);
- PER (Tillmann et al., 1997);
- TER (Snover et al., 2006) with four variants, including:

1. -TER default (i.e., with stemming and synonymy lookup);
2. -TERbase (i.e., without stemming, synonymy lookup, nor paraphrase support);
3. -TERp with stemming, synonymy lookup and paraphrase support;
4. -TERpA i.e., -TERp tuned towards adequacy.

Precision Oriented Measures

In the field of translation, precision refers to the ratio between acceptable n- grams in the candidate translation (i.e. the ones which are also used in reference translations) to the total number of n-grams in the corresponding candidate translation segment. The implementations of these measures in the field of MT evaluation include:

- BLEU (Papineni et al., 2002);
- NIST (Doddington, 2002);
- Pl (Lexical Precision) (Gonzàlez & Giménez 2014).

Recall Oriented Measures

In the field of translation, recall computes the proportion of acceptable n- grams in the candidate translation to the number of n-grams of the reference translation. In other words, it computes the proportion of n-grams in the reference translations covered by the candidate translation. Such implementations in the field of MT evaluation include:

- Rouge (Lin & Och, 2004) with Eight variants, including:

1. $ROUGE_n$ (for several n-gram lengths, up to 4);
2. $ROUGE_L$ (longest common subsequence);
3. $ROUGE_S{}^*$ (skip bigrams with no max-gap-length);
4. $ROUGE_{SU}$* (skip bigrams with no max-gap-length, including unigrams);
5. $ROUGE_W$* (weighted longest common subsequence with weighting factor w = 1.2).

- Rl (Lexical Recall) (Gonzàlez & Giménez, 2014).

F-Measure Oriented Measures

As the last subcategory of the lexical similarity metrics, the F-measure metrics combine lexical precision and recall to calculate the score that defines the quality of candidate translations. The implementations of these metrics in the field of MT evaluation include:

- GTMe (Melamed et al., 2003);
- METEOR (Banerjee & Lavie, 2005) with four variants, including:

1. $METEOR_{ex}$ (only exact matching);
2. $METEOR_{st}$ (plus stem matching);
3. $METEOR_{sy}$ (plus synonym matching);
4. $METEOR_{pa}$ (plus paraphrase matching).

- Fl (Lexical Fl) (Gonzàlez & Giménez, 2014);
- Ol (Lexical Overlap) (Gonzàlez & Giménez, 2014).

Theoretical Framework

The main methodological framework that governs the theoretical foundation of the present study is a set of basic levels of expertise based on the traditional craft guild terminology adapted from Hoffman, 1998 (in Hoffman, et al., 2014, p. 24). It refers to the features, skills, and knowledge that differentiate experts from novices and less experienced people (Ericsson, Hoffman, Kozbelt, & Williams, 2018, p. 4). One of the research approaches to expertise is to study experts in comparison to novices (Chi, 2006). This relative approach presupposes that expertise is one of the various levels of proficiency that novices can achieve.

Table 1. Basic Expertise Categories Based on the Traditional Craft Guild Terminology (Adapted from Hoffman, 1998 in Hoffman, et al., 2014)

Naïve	One who is ignorant of a domain.
Novice	Literally, someone who is new—a probationary member. There has been some ("minimal") exposure to the domain.
Initiate	Literally, someone who has been through an initiation ceremony—a novice who has begun introductory instruction.
Apprentice	Literally, one who is learning—a student undergoing a program of instruction beyond the introductory level. Traditionally, the apprentice is immersed in the domain by living with and assisting someone at a higher level. The length of an apprenticeship depends on the domain, ranging from one to 12 years in the craft guilds.
Journeyman	Literally, a person who can perform a day's labor unsupervised, although working under orders. An experienced and reliable worker or one who has achieved a level of competence. It is possible to remain at this level for life.
Expert	The distinguished or brilliant journeyman, highly regarded by peers, whose judgments are uncommonly accurate and reliable, whose performance shows consummate skill and economy of effort, and who can deal effectively with certain types of rare or "tough" cases. In addition, an expert is one who has special skills or knowledge derived from extensive experience with subdomains.
Master	Traditionally, a master is any journeyman or expert who is also qualified to teach those at a lower level. A master is one of an elite group of experts whose judgments set the regulations, standards or ideals. In addition, a master can be that expert who is regarded by the other experts as being "the" expert, or the "real" expert, especially with regard to subdomain knowledge.

Borrowing Hoffmann's (1998) tentative classification of the various stages of expertise, Kiraly has made an attempt to operationalize it in the domain of translation studies as 'the students who enter the programs of study are usually novices, who are initiated into the domain through introductory courses, and continue to learn as apprentices and are expected to reach the stage of journeyman by the time they graduate. However, true expertise can develop only after many years of real-world experience after graduation' (Kiraly, 2000, p. 59).

These various levels of expertise can be evaluated by measures such as academic qualifications (e.g. graduate students vs. undergraduate students), seniority or experience in performing the related representative tasks, or consensus among peers. However, it can also be evaluated at a more fine-grained level, in terms of domain specific knowledge or performance tests (Chi, 2006). The present study tries to investigate the difference that the addition of more reference translations creates in the reliable designation of experts in comparison to less knowledgeable participants. In other words, the researchers are to find if using more reference translations while applying the lexical ATQEUMs as the benchmark to decide upon the accuracy and/or fluency of the translated texts results in more reliable scores when compared with the gold

scores provided by human scorers.

3. Method

The research that was conducted in the course of this study was of empirical and applied nature, "which is research on practical problems, research that has an application in life" (Saldanha and O'Brien, 2014). In fact, it was an attempt to establish the intended or unintended effects of the initiative — i.e. using up to five reference translations in the application of lexical ATQEUMs to the evaluation of human translation.

Participants

The participants of the present study consisted of various groups, including:

1. Four expert certified translators with at least 10 years of professional experience in the field who were asked to translate the sample texts into English which have been used as reference translations.
2. Five expert translation scorers with at least 10 years of extensive experience in both the profession itself and teaching and scoring translations.
3. Fifty-one undergraduate learners of translator training program in the Iranian universities as the potential applicants of the Test with no important determinant criteria who were required to translate the sample texts into English which have been used as candidate translations.

Materials and Instruments

The design of the present research demanded the application of an open toolkit for automatic translation (meta-) evaluation, namely Asiya (Giménez & Màrquez, 2010), and a translation test comprised of two parts (Text A and Text B):

The Translation Test

The sample test of the Certified Translator Accreditation Test of the Judiciary of the Islamic Republic of Iran held in 2006, to be translated from Persian into English, is chosen as the translation test text. It constituted of 674 words, including three Articles from the Civil Code of the Islamic Republic of Iran (i.e. Articles 976, 977, and 978 as Text A) and six Articles from the Criminal Code of the Islamic Republic of Iran (i.e. Articles 626, 627, 628, 629, 630 and 631 as Text B).

Asiya: An Automated Metric Repository

Being publicly available at http://asiya.lsi.upc.edu, Asiya is an open toolkit that provides an interface to a collection of not only reference-based but also Quality Estimation metrics (Gonzàlez & Giménez, 2014). It includes a resourceful collection of reference-based metrics based on different similarity measures operating at various textual dimensions, and provides schemes for metric combination and mechanisms to determine optimal metric sets (Gonzàlez & Giménez, 2014). Asiya has originally been designed for the (meta-) evaluation of the quality of MT output; however, the researchers have tried to extend the scope of the efficiency of the online toolkit to cover the human translations as well.

Procedure

The researchers first asked 51 English Translation senior undergraduates to translate the designated sample texts into English. Since, the scoring process of the translated texts was the pivot around which this whole research study rounds, there was no other considerations with regard to other conditions of taking the test. Similarly, the same texts were simultaneously given to four expert translators in order to provide the reference translations (since one authorized sample translation of the texts is publicly available).

Having collected the candidate translations, the researchers asked 5 expert human raters to score the candidate translations (from 20) taking

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 3, Autumn 2020, Page 618 to 629

a correct sample translation provided by the researchers as the reference based on which they were supposed to decide upon the quality of the translations as this is the same approach that is applied in the lexical ATQEUMs method.

As the next step of the study, the researchers had to score the same candidate translations using the lexical ATQEUMs and up to 5 reference translations available in order to collect the second set of data required to answer the research question.

Data Analysis

Following a quantitative approach, the relationships between lexical ATQEUMs' and human scores on sample Texts A and B using 1, 2, …, 5 reference translations are calculated in terms of Pearson correlation coefficient accompanied by 95% confidence intervals using bootstrap resampling. The different sets of lexical ATQEUMs (i.e. edit distance, precision oriented, recall oriented and F-measure oriented measures) are to be examined separately for both Texts A and B respectively using 1, 2, …, 5 reference translations to find the difference, if any, that is made in the reliability level of the automated scoring method.

4. Results and Discussion

Results

Edit Distance Lexical ATQEUMs

The relationship between edit distance lexical ATQEUMs and expert human scores for Texts A and B is represented in Tables 2 and 3 respectively in terms of Pearson correlation coefficient accompanied by 95% confidence intervals using bootstrap resampling. The researchers had hypothesized that adding more reference translations would lead to more objective scores and higher reliability when compared with the scores given by expert human scorers. However, the results obtained with regard to Text A contradicted this hypothesis, i.e. in 66.66% of the cases, use of just 2 reference translations [-WER, -PER, -TER_{base}, and -TER_PA], and in 33.33% of the cases use of just 1 reference translation [-TER, and -TER_P], have led to the highest correlations. Interestingly, using 4 [-TER, -TER_P, and -TER_PA] and 5 [-WER, -PER, -TER_{base}] reference translations have each led to the lowest correlations in 50% of the cases.

Table 2. The Relationship between Edit Distance Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text A)

	1 Reference	2 References	3 References	4 References	5 References
-WER	0.7756	0.7814	0.7811	0.70434	0.6947
-PER	0.8759	0.8883	0.8817	0.8092	0.8031
-TER	0.8172	0.8169	0.8159	0.6383	0.6384
-TER_{base}	0.8089	0.8116	0.8106	0.6310	0.6302
-TER_P	0.8391	0.8358	0.8340	0.6540	0.6626
-TER_PA	0.8969	0.9011	0.8986	0.7927	0.8062

Similarly, in the case of Text B, in 50% of the cases, using just 2 reference translations has led to the highest correlations [-TER, -$TER_{base,}$ and -TER_P], while using 3 [-WER], 4 [-PER] and 5 [-TER_PA] reference translations have each led to the highest correlations in 16.66% of the cases. While, using 1 [-WER, -PER, and -TER_PA] and 5 [-TER, -$TER_{base,}$ and -TER_P] reference translations have each led to the lowest correlations in 50% of the cases.

Table 3. The Relationship between Edit Distance Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text B)

	1 Reference	2 References	3 References	4 References	5 References
-WER	0.8386	0.8536	0.8537	0.8531	0.8521
-PER	0.8918	0.9031	0.9058	0.9060	0.9057
-TER	0.8543	0.8716	0.8577	0.8579	0.8505
-TER_{base}	0.8562	0.8743	0.8592	0.8598	0.8528
-TER_P	0.8867	0.9032	0.8924	0.8933	0.8865
-TER_PA	0.8722	0.9187	0.9232	0.9253	0.9267

Precision Oriented Lexical ATQEUMs

The relationship between precision oriented lexical ATQEUMs and expert human scores for

Texts A and B is represented in Tables 4 and 5 respectively in terms of Pearson correlation coefficient accompanied by 95% confidence intervals using bootstrap resampling. The results obtained with regard to Text A again contradicted the researcher's hypothesis, i.e. in 52.94% of the cases, use of just 2 reference translations [BLEU-1, BLEU-2, BLEU-3, BLEU-4, NIST-3, NIST-4, NIST-5, NISTi-4, and Pl] and in 5.88% of the cases use of 3 reference translations [NISTi-3], have led to the highest correlations. However, in 41.17% of the cases, use of 5 reference translations [BLEUi-2, BLEUi-3, BLEUi-4, NIST-1, NIST-2, NISTi-2, and NISTi-5] has led to the highest correlations. Interestingly, using 4 reference translations has led to the lowest correlations in 76.47% of the cases [BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, and Pl]. Nevertheless, in 17.64% of the cases [NISTi-3, NISTi-4, NISTi-5] using 1 reference translation has led to the lowest correlations, and in just 5.88% of the cases [BLEUi-4], using 3 reference translations have led to the lowest correlations.

Table 4. The Relationship between Precision Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text A)

	1 Reference	2 References	3 References	4 References	5 References
BLEU-1	0.9082	0.9090	0.9046	0.8381	0.8638
BLEU-2	0.8475	0.8539	0.8492	0.7869	0.8178
BLEU-3	0.8030	0.8109	0.8078	0.7562	0.7850
BLEU-4	0.7732	0.7799	0.7755	0.7361	0.7614
BLEUi-2	0.6945	0.7005	0.6977	0.6907	0.7180
BLEUi-3	0.6864	0.6889	0.6838	0.6831	0.7058
BLEUi-4	0.6787	0.6786	0.6730	0.6780	0.6951
NIST-1	0.9145	0.9169	0.9144	0.9074	0.9189
NIST-2	0.9033	0.9070	0.9034	0.8899	0.9071
NIST-3	0.8991	0.9033	0.8995	0.8840	0.8999
NIST-4	0.8977	0.9016	0.8978	0.8814	0.8969
NIST-5	0.8963	0.8997	0.8961	0.8785	0.8953
NISTi-2	0.8175	0.8243	0.8213	0.7966	0.8265
NISTi-3	0.7161	0.7538	0.7546	0.7240	0.7289
NISTi-4	0.6773	0.7220	0.7215	0.6908	0.7057
NISTi-5	0.6436	0.6524	0.6500	0.6439	0.6936
Pl	0.6336	0.6799	0.6773	0.6208	0.6209

With regard to Text B, in 58.82% of the cases, using just 2 reference translations has led to the highest correlations [BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, NIST-2, NIST-3, NIST-4, and NIST-5], while using 3 reference translations has led to the highest correlations in 23.52% of the cases [BLEU-1, NISTi-3, NISTi-4, and NISTi-5]. However, in 11.76% of the cases [NIST-1, and Pl], using 2 reference translations and in just 5.88% of the cases [NISTi-2], using 4 reference translations have led to the highest correlations. Moreover, using 1 reference translation has led to the lowest correlations in 94.11% of the cases [BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, and NISTi-5]. In contrast, in just 5.88% of the cases, using 5 reference translations has led to the lowest correlation [Pl].

Table 5. The Relationship between Precision Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text B)

	1 Reference	2 References	3 References	4 References	5 References
BLEU-1	0.9053	0.9453	0.9466	0.9439	0.9446
BLEU-2	0.8503	0.9066	0.9331	0.9351	0.9396
BLEU-3	0.8045	0.8598	0.8915	0.8956	0.8996
BLEU-4	0.7711	0.8220	0.8528	0.8546	0.8597
BLEUi-2	0.7707	0.8190	0.8325	0.8347	0.8351
BLEUi-3	0.7287	0.7706	0.7880	0.7901	0.7907
BLEUi-4	0.7030	0.7416	0.7530	0.7528	0.7538
NIST-1	0.9137	0.9484	0.9442	0.9404	0.9418
NIST-2	0.9010	0.9393	0.9475	0.9464	0.9485
NIST-3	0.8987	0.9365	0.9464	0.9463	0.9482
NIST-4	0.8977	0.9352	0.9459	0.9457	0.9477
NIST-5	0.8972	0.9346	0.9455	0.9449	0.9470
NISTi-2	0.8117	0.8686	0.8958	0.8971	0.8970
NISTi-3	0.7616	0.8159	0.8465	0.8422	0.8358
NISTi-4	0.7416	0.7875	0.8223	0.8124	0.8115
NISTi-5	0.7079	0.7615	0.7932	0.7770	0.7761
Pl	0.7112	0.7203	0.6347	0.6312	0.6273

Recall Oriented Lexical ATQEUMs

The relationship between recall oriented lexical ATQEUMs and expert human scores for Texts A and B is represented in Tables 6 and 7 respectively in terms of Pearson correlation coefficient accompanied by 95% confidence intervals using bootstrap resampling. The results obtained with regard to Text A confirmed the

researcher's hypothesis, i.e. in 88.88% of the cases, use of 5 reference translations has led to the highest correlations [ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, and ROUGE-W]. However, in just 11.11% of the cases, use of 2 reference translations [Rl] has led to the highest correlations. Interestingly, using 4 reference translations has led to the lowest correlations in 88.88% of the cases [ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, and ROUGE-W]. Nevertheless, in 11.11% of the cases [Rl] using 5 reference translations has led to the lowest correlations.

Table 6. The Relationship between Recall Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text A)

	1 Reference	2 References	3 References	4 References	5 References
ROUGE-1	0.8474	0.8442	0.8461	0.8399	0.8505
ROUGE-2	0.7743	0.7755	0.7735	0.7695	0.7881
ROUGE-3	0.7143	0.7180	0.7122	0.7095	0.7371
ROUGE-4	0.6931	0.6959	0.6900	0.6876	0.7098
ROUGE-L	0.8220	0.8192	0.8200	0.8144	0.8296
ROUGE-S*	0.7880	0.7874	0.7865	0.7823	0.7958
ROUGE-SU*	0.8019	0.7996	0.7975	0.7954	0.8115
ROUGE-W	0.7931	0.7910	0.7892	0.7866	0.8046
Rl	0.8465	0.8535	0.8485	0.7807	0.7803

The results obtained with regard to Text B, however, contradicted the researcher's hypothesis once more, i.e. in 55.55% of the cases, use of just 1 reference translation has led to the highest correlations [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S*, and ROUGE-W]. However, using 2 [ROUGE-SU*, and Rl] and 5 [ROUGE-3, and ROUGE-4] reference translations have each led to the highest correlations in 22.22% of the cases. Interestingly, using 5 reference translations has led to the lowest correlations in 66.66% of the cases [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S*, ROUGE-SU*, and ROUGE-W]. Nevertheless, in 33.33% of the cases [ROUGE-3, ROUGE-4, and Rl], using 1 reference translation has led to the lowest correlations.

Table 7. The Relationship between Recall Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text B)

	1 Reference	2 References	3 References	4 References	5 References
ROUGE-1	0.8753	0.8713	0.8407	0.8448	0.8394
ROUGE-2	0.8220	0.8116	0.7823	0.7747	0.7613
ROUGE-3	0.7663	0.8234	0.8611	0.8587	0.8651
ROUGE-4	0.7277	0.7731	0.8030	0.8043	0.8121
ROUGE-L	0.8717	0.8711	0.8449	0.8472	0.8416
ROUGE-S*	0.8246	0.8142	0.7855	0.7835	0.7683
ROUGE-SU*	0.8509	0.8566	0.8325	0.8299	0.8217
ROUGE-W	0.8329	0.8256	0.8041	0.7968	0.7878
Rl	0.8725	0.8851	0.8777	0.8790	0.8804

F-Measure Oriented Lexical ATQEUMs

The relationship between F-measure oriented lexical ATQEUMs and expert human scores for Texts A and B is represented in Tables 8 and 9 respectively in terms of Pearson correlation coefficient accompanied by 95% confidence intervals using bootstrap resampling. The results obtained with regard to Text A again contradicted the researcher's hypothesis, i.e. in 66.66% of the cases, use of just 1 reference translation has led to the highest correlations [GTM-2, GTM-3, METEOR-ex, METEOR-st, METEOR-sy, and METEOR-pa]. Moreover, in 22.22% of the cases, use of 2 reference translations [Fl, and Ol] has led to the highest correlations. In contrast, using 5 reference translations has led to the highest correlations in just 11.11% of the cases [GTM-1]. Interestingly, using 5 reference translations has led to the

lowest correlations in 44.44% of the cases [METEOR-ex, METEOR-st, METEOR-sy, and METEOR-pa]. Furthermore, using 3 [GTM-2, and GTM-3] and 4 [Fl and Ol] reference translations have each led to the lowest correlations in 22.22% of the cases. Nevertheless, in 11.11% of the cases [GTM-1] using 1 reference translation has led to the lowest correlations.

Table 8. The Relationship between F-measure Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text A)

	1 Reference	2 References	3 References	4 References	5 References
GTM-1	0.8768	0.8883	0.8834	0.8834	0.9116
GTM-2	0.7161	0.7072	0.6931	0.7025	0.7134
GTM-3	0.6959	0.6873	0.6714	0.6851	0.6944
METEOR-ex	0.8456	0.8347	0.8011	0.7564	0.7399
METEOR-st	0.8458	0.8364	0.8027	0.7554	0.7410
METEOR-sy	0.8515	0.8419	0.8066	0.7603	0.7457
METEOR-pa	0.8666	0.8562	0.8164	0.7707	0.7540
Fl	0.8289	0.8396	0.8363	0.74485	0.74488
Ol	0.7943	0.8053	0.7991	0.7184	0.7224

The results obtained with regard to Text B again contradicted the researcher's hypothesis, i.e. using 2 [METEOR-pa, Fl, and Ol] and 3 [GTM-1, GTM-3, and METEOR-ex] reference translations have each led to the highest correlations in 33.33% of the cases. Moreover, in 22.22% of the cases, use of 5 reference translations [GTM-2, and METEOR-st] has led to the highest correlations, while using 4 reference translations has led to the highest correlations in just 11.11% of the cases [METEOR-sy]. Interestingly, using 1 reference translation has led to the lowest correlations in 100% of the cases [GTM-1, GTM-2, GTM-3, METEOR-ex, METEOR-st, METEOR-sy, METEOR-pa, Fl, and Ol].

Table 9. The Relationship between F-measure Oriented Lexical ATQEUMs and Expert Human Scores in Terms of Pearson Correlation Coefficient (Text B)

	1 Reference	2 References	3 References	4 References	5 References
GTM-1	0.8870	0.9487	0.9537	0.9504	0.9517
GTM-2	0.7420	0.7761	0.7834	0.7821	0.7842
GTM-3	0.7242	0.7544	0.7611	0.7587	0.7599
METEOR-ex	0.7994	0.8189	0.8262	0.8260	0.8259
METEOR-st	0.8039	0.8230	0.8308	0.8307	0.8312
METEOR-sy	0.8190	0.8384	0.8425	0.8435	0.8432
METEOR-pa	0.8298	0.8520	0.8505	0.8507	0.8516
Fl	0.8608	0.8764	0.8664	0.8666	0.8651
Ol	0.8072	0.8320	0.8122	0.8112	0.8101

Discussion

The researchers of the study had hypothesized that using more reference translations in the case of lexical ATQEUMs as the benchmark to decide upon the quality (both accuracy and fluency) of human translations would result in more objective and reliable scores in comparison to the scores given by the expert human scorers. This hypothesis is based on the logic that human translators are creative in such translation processes as problem solving and decision making while selecting the best equivalent possible in the target text. Therefore, deciding upon the quality of the translated text based on just 1 reference translation does not seem sufficient to encompass all the possible and equally accurate equivalents. However, addition of more reference translations has not led to higher correlations with the scores provided by expert human scorers in all cases.

In the case of Text A, using just 1 reference translation has led to the highest correlations in 19.51% of the cases. This number has increased

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 3, Autumn 2020, Page 618 to 629

to 39.02 % with the addition of the second reference translation which proves the hypothesis. In contrast, addition of the third reference translation has led to a great decrease in the correlation level, i.e. only 2.43% of the cases. This decrease has reached to 0.0% of the cases with the addition of the fourth reference translation. Nonetheless, addition of the fifth reference translation has compensated for all these deductions, i.e. the number has increased to 39.02% of the cases. In other words, using 5 reference translations has had an exact result as using 2 reference translations, i.e. the highest correlations in 39.02% of the cases.

In the case of Text B, using just 1 reference translation has led to the highest correlations in 12.19% of the cases. This number has increased to 24.39 % with the addition of the second reference translation which proves the hypothesis. In contrast, addition of the third reference translation has led to a decrease in the correlation level, i.e. 19.51% of the cases. This decrease has reached to 7.31% of the cases with the addition of the fourth reference translation. Nonetheless, addition of the fifth reference translation has compensated for all these deductions, i.e. the number has increased to 36.58% of the cases.

Furthermore, it is noteworthy to mention that the difference between the correlations has sometimes been as small as 0.0001, i.e. too small to reject the whole logic of using more reference translations. However, the issue can also be defined in the light of the fact that some of the lexical ATQEUMs themselves take account of such concepts as synonymy, paraphrase, and stemming as well, including -TER with its four variants (i.e. -TER default, -TER_{base}, -TER_{P}, and -TER_{P}A) and METEOR with its four variants (i.e. METEOR-ex, METEOR-st, METEOR-sy, and METEOR-pa). In other words, considering these concepts itself alleviates the need of using more reference translations to a great extent which can explain the cases where using more reference translations has not led to higher correlations.

5. Conclusion

This empirical, applied, evaluative and formative research study was an attempt to examine the effects of using more reference translations as the benchmark to decide upon the quality of human translations while applying the lexical ATQEUMs on the reliability of the scores provided. To assess the reliability of the lexical ATQEUMs' scores, their correlation with the scores given by the expert scorers was calculated. The higher the correlation, the more reliable the scores are interpreted to be. All these metrics work according to a basic common rule: the similarity level of the texts compared. Therefore, it was hypothesized that using more reference translations would result in higher correlations with the scores provided by the expert scorers which can be interpreted as higher reliability. On the whole, considering the results obtained from both Texts A and B, it can be concluded that using 5 reference translations has led to the highest correlations in 37.80% of the cases, which is more than the number for any other situation considered (i.e. using 4 reference translations (3.65% of the cases), 3 reference translations (10.97% of the cases), 2 reference translations (31.70% of the cases), and 1 reference translation (15.85% of the cases)). This confirms the hypothesis considered by the researchers which means that using more reference translations will lead to higher reliability when the comparison is between 1, 2 and 5 reference translations. Nevertheless, addition of the third and fourth reference translations does clearly reject the hypothesis. This can be explained by the quality of the reference translations, the coverage of such concepts as synonymy, paraphrase, and stemming by some of the metrics. However,

researches of greater scopes are needed to find the standard of the number of the reference translations that are sufficient to come upon the most reliable scores possible.

References

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

Beikmohammadi, Maryam, Alavi, Seyyed-Mohammad, Kaivanpanah, Shiva (2020). Learning-oriented Assessment of Reading: A Mixed Methods Study of Iranian EFL University Instructors' Perceptions and Practices. Journal of Foreign Language Research, 10 (2), 316-329.

Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. Meta: Translators' Journal, 46(2), 345-364.

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman, The Cambridge handbook of expertise and expert performance, First Edition (pp. 21-29).

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. Proceedings of the 2nd International Conference on Human Language Technology, (pp. 138–145).

Gonzàlez, M., & Giménez, J. (2014). Asiya: An open toolkit for automatic machine translation (meta-)evaluation, Technical Manual, Version 3.0. Retrieved from TALP Research Center Project Management.

Hoffman, R., Ward, P., Feltovich, P. J., Dibello, L., Fiore, S. M., & Andrews, D. H. (2014). Accelerated expertise, training for high proficiency in a complex world. New York: Taylor & Francis.

House, J. (1997). Translation quality assessment. A model revisited. Gunter Narr Verlag: Tubingen.

Kiraly, D. (2000). A social constructivist approach to translator education, empowerment from theory to practice. London and New York: St. Jerome Publishing.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 8(10), 707–710.

Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics.

Màrquez, L. (2013). Automatic evaluation of machine translation quality. Invited talk at Dialogue 2013. Bekasovo Resort, Russia: TALP Research Center, Technical University of Catalonia (UPC).

Melamed, I. D., Green, R., & Turian, J. (2003). Precision and recall of machine translation. Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL).

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC).

Olohan, M. (2004). Introducing corpora in translation studies. New York: Routledge.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), (pp. 311-318). Philadelphia.

Saldanha, G., & O'Brien, S. (2014). Research methodologies in translation studies. London and New York: Routledge, Taylor and Francis Group.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA) (pp. 223–231).

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

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability. TOEFL iBT® Research Report. ETS, Georgia State University, Atlanta.