

Semantics based English-Arabic machine translation evaluation

Majdi Beseiso¹, Samiksha Tripathi², Bashar Al-Shboul³, Renad Aljadid⁴

¹Department of Computer Science, Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Salt, Jordan

²Principal Architect – AI, Department of R Digital, R Systems International, Noida, India

³Department of Information Technology, King Abdullah II School of Information Technology, University of Jordan, Amman, Jordan

⁴Department of English Language and Literature, School of Foreign Languages, University of Jordan, Amman, Jordan

Article Info

Article history:

Received Aug 18, 2021

Revised Apr 25, 2022

Accepted May 21, 2022

Keywords:

Arabic machine translation
Bilingual evaluation understudy
Dense sentence embedding
Linguistic knowledge

ABSTRACT

Some classic machine translation (MT) Evaluation methods, such as the bilingual evaluation understudy score (BLEU), have notably underperformed in evaluating machine translations for morphologically rich languages like Arabic. However, the recent remarkable advancements in the domain of word vectors and sentence vectors have opened up new research avenues for low-resource languages. This paper proposes a novel linguistic-based evaluation method for English-translated sentences in Arabic. The proposed approach includes penalties based on length, positions, and context-based schemes such as part-of-speech tagging (POS) and multilingual sentence-BERT (SBERT) models for machine translation evaluation. The proposed technique is tested using pearson correlation as a performance evaluation parameter and compared with state-of-the-art techniques. The experimental results demonstrate that the proposed model evidently outperforms other MT evaluation methods such as BLEU.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Majdi Beseiso

Department of Computer Science, Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University
Salt, Jordan

Email: bsaiso@bau.edu.jo

1. INTRODUCTION

In today's technology-driven landscape, a state-of-the-art approach across machine translation (MT) is imperative. Evaluating the quality of translated text by machine translation systems is one of the prime concerns. There is no denying that the assessment of machine translations by humans is a time-consuming and expensive practice [1]. Moreover, considering the rising trend of machine-based translation, tools to automate MT evaluation are of utmost significance for researchers in natural language processing (NLP) and specifically for MT research. Translation of text has been done traditionally by human translators, which is a very costly, time-consuming and biased process. MT system, however, gained more prominence with the advent of new NLP methods in linguistic evaluations and MT systems improvements [2]. The exertion of excessive resources in the form of time and money on the research, design, and implementation of an MT system necessitates the meticulous evaluation of the entire MT development cycle. It is essential to identify and analyze the possible source of errors and explore potential approaches for addressing these concerns. These approaches must be deployed and tested to observe whether the identified errors lessen without compromising the system performance. If the system performance is not affected or degraded, only then the mechanism is accepted; otherwise, some alternative mechanism is devised [3]. In this work, a novel system for evaluating English to Arabic translation has been presented. This system uses multiple measures, including part-of-speech (POS) and sentence-BERT (SBERT), to ensure the accuracy needed for the translation evaluation of rich and complex

languages like Arabic. With the help of Arabic linguistic experts, the presented method is compared with state-of-the-art MT evaluation systems and has proven to be more accurate.

The rest of the paper is arranged as follows: section 2 covers the recent works on machine translation. Section 3 presents the proposed approach and methodology, while section 4 describes BERT-based sentence similarity computation. Section 5 discusses the experimental setup and evaluation results of the proposed Arabic machine translation system. Lastly, section 6 concludes the paper and identifies potential directions for extending this work in the future.

2. RELATED WORK

Human evaluation of machine translation systems is highly subjective, time-consuming, and cannot be reused [4]. Recently, automatic evaluation methods have gained substantial attention due to their unprecedented benefits. bilingual evaluation understudy score (BLEU) [5] is one of the most well-known and widely employed algorithms for evaluating the quality of a machine-translated text. It calculates n-gram precision and a brevity penalty between the candidate and reference translation. For example, it takes a whole source sentence (n-gram) and compares it with a reference sentence (n-gram) regardless of the words' position. Later, some alterations have been made to the bilingual evaluation understudy (BLEU) metric to introduce the NIST metric, which provides weights to n-gram based on their information level. Besides that, in 2006, Turian *et al.* [6] proposed general text matcher (GTM) based on accuracy measures such as precision, recall, and f-measure. Another package, named ROUGE, was presented back in 2003 for the automatic evaluation of summaries, wherein the computer-generated summary (candidate summary) is evaluated with the human-created summary (reference summary) [7]. For example, the paragraphs summarized using ROUGE are compared with the same paragraphs summarized by a human, and n-gram-wise scores are compared to check the performance.

In another relevant study, different human judgments are explored with human mediated translation edit rate (HTER) [8]. The HTER is a semi-automatic measure where humans do not score the MT output but generate a new reference translation, which is closer to the MT output and try to retain the fluency and adequacy of the original translated reference [8]. Following that, Callison-Burch *et al.* [9] attempted to correlate the automatic evaluation metrics with human judgments. They tried to determine which automatic system produces the highest quality of translation from the list of nine different automatic evaluation metrics [9] and ranked them with the help of comprehensive human evaluation. They also mentioned the two categorical scales currently used by the human evaluators to represent fluency and adequacy of the MT system.

In 2005, Banerjee and Lavie [10] introduced METEOR, an innovative automatic method for MT evaluation that creates a word alignment between the two sentences, i.e., candidate translation string and reference translation string. The alignment is done through word mapping such as i) Stem matching, ii) Exact matching, and iii) Synonym matching. In recent years, an extension of the METEOR translation evaluation metric was presented to the phrase level in METEOR NEXT metric [11]. Previously, METEOR needed human-based judgments in target languages. However, METEOR Universal learns function word lists and paraphrase tables to provide language-specific evaluations. Furthermore, METEOR Universal depicted improved performance for Hindi and Russian languages since it uses a universal parameter set learned through pooling [12].

Several attempts, with reference translations [13]-[18] and without reference translations [16]-[19], are made over the years to involve ML techniques in MT evaluation. In Corston *et al.* [16], employed decision trees for evaluating the well-formedness of the MT output and building classifiers that learn to distinguish human translations from MT. In contrast, Akiba *et al.* [20] approached this as a multi-class classification task and trained the decision-tree classifiers on multiple edit distance features that include lexical, morpho-syntactic, and lexical-semantic information. Moreover, Kuleza and Shieber [17] trained a support vector machine (SVM) to serve the purpose.

In Quirk [19], argued that human references are not mandatory for MT evaluation. They applied numerous supervised algorithms, including Decision trees, SVMs, and linear regression. All these statistical techniques worked well, but linear regression exhibited exceptional performance. Following that, Russo-Lassner *et al.* [21] developed a linear regression model that used stemming, WordNet synonymy, verb class synonymy, noun phrase heads matching, and matching proper names. Some other studies [13]-[15] also suggested linear regression-based models for the metric combination that outperformed most of the existing approaches. Gamon *et al.* [22], introduced a new approach that involved the training of a large corpus of domain-specific data instead of modeling the output on a target language. They also added perplexity scores to improve the sentence-level language model. Ye *et al.* [23] and Duh [24] approached the sentence-level MT evaluation as a ranking problem. They used n-grams, dependency, and translation perplexity of the reference language model (LM) as features for ranking the SVM algorithm. Gautam and Bhattacharyya [25] proposed a layered approach for MT evaluation based on lexical, syntactic, and semantic layers. They used BLEU as a

baseline metric for the lexical layer while Hamming score, Kendall Tau distance score, and the spearman rank score were considered for the syntactic layer.

In [26], the authors explored language divergences and ambiguities in English to Arabic machine translation. Keeping different features of Arabic in mind, Abu-Ayyash [27] worked on errors and non-errors made by MT systems. He also investigated the extent to which MT systems can deliver when the language has different rules and grammatical representations. A comparative study between various MT systems using BLEU and METEOR was done in [28]. It aimed at identifying the most suitable metric for the Arabic to English translation system, which helps the developers enhance the effectiveness of these systems. The authors examined the translation accuracy of two known machine translation programs, Google translator and Babylon, translating the exact Arabic text into English. These methods were used to measure the quality of MT and determine the scheme closest to human ratings. They declared that BLEU is the best method for human rating judgments.

Moreover, different approaches for Arabic to English translations are reviewed in [29], which depicted that neural machine translation approaches demonstrated greater accuracy than the other alternatives. Furthermore, the emerging attention-based approach is found to be remarkably effective at improving NMT's performance for all languages. Guzman *et al.* [30] presented Kendall's τ scores obtained from five n-gram-based metrics and observed findings while training neural networks with embeddings of different representations as input. They added different lexical and morpho-syntactic features of languages and compared the performance of BLEU, NIST, METEOR, and 1-TER, along with AL-BLEU. They observed that both NIST and METEOR obtained approximately the same performance for this task. In another study, Shimanaka *et al.* [31] explored the utility of universal sentence representations to measure machine translation quality, while training sentence representations using a small translation dataset is a challenging task. They also introduced the regressor using sentence embeddings (RUSE) metric during a workshop on machine translation'18 (WMT18) session. RUSE uses sentence embeddings and can capture global information that n-gram based models fail to capture. It has also been concluded that universal sentence embeddings trained on a limited or small in-domain dataset are less effective compared to the ones trained on a large-scale dataset.

During the survey, several other evaluation metrics have been investigated, and it was found that many of these metrics have released their upgraded versions. Moreover, it was observed that some of the existing metrics could finely correlate with manual evaluations; however, not all of these have this capability. They are incapable of performing well with all the languages, particularly the morphologically rich ones. The summary of the literature review is given in Table 1.

Table 1. Summary of literature review

Study	Technique	Limitations	Languages
[5]	BLEU	It does not handle morphologically rich languages well and does not consider sentence structure directly.	Language independent
[6]	GTM	The target language is overlooked since the significant focus stays on the mother tongue.	Chinese to English Arabic to English
[7]	ROUGE	It does not provide a conclusive understanding of how well summaries perform in comparison with human summaries.	Arabic to English
[8]	HTER	It is a strictly quantitative metric and weights all errors equally	Arabic to English Hungarian to English
[32]	Correlation between automated evaluation and human judgement	A few investigations have shown a poor correlation between annotators utilizing this strategy	German to English Spanish to English Czech to English
[11]	METEOR	Requires lots of human effort for translation	Arabic to English Chinese to English
[19]	Linguistic model	Evaluation metrics were computationally expensive and could only be used if only a few different hypotheses needed to be tested	Spanish to English
[21]	Paraphrase based Linear regression model	Inability to understand local slang	Arabic to English
[22]	SVM based Linguistic model	It limits the amount of data that can be processed.	French to English
[13]-[15]	Regression-based learning model	Can lead to erroneous and misleading results	Chinese to English Arabic to English
[23]	SVM learning algorithm	Limitation on the quantity of data	Chinese to English
[25]	NLP layers based	High computational cost	German to English Spanish to English French to English
[27]	PNMT based	It has grammatical mistakes and no quality control	Arabic to English
[26]	ANN and rule-based MT system	It demands a tremendous amount of time and linguistic resources	Arabic to English
[29]	Google translation based MT system	Involves the use of creative linguistic tools	Arabic to English English to Arabic
[30]	Neural network-based MT evaluation model	The source-text sentences must be evident and cohesive.	Arabic to English

3. THE PROPOSED METHOD

The main drawbacks of performing manual evaluation include resource consumption, time utilization and task recurrence [33]. On the other hand, automatic evaluation has many advantages, such as good performance in some languages compared to others, primarily when English is used as a target language [5]. It can be attributed to the rich data resources available for English and the rich resources users can use for aiding evaluation, such as dictionaries, and thesauruses. However, it exhibits poor performance when English is used as a source language against another low-resource language, such as Arabic or Urdu. One of the main reasons is the lack of appropriate data for evaluation. A few metrics use several linguistic features that make it harder to generalize them for other languages. Moreover, other metrics, for instance BLEU [5], utilizes context-independent features using an n-gram precision score. Some researchers believe that a high BLEU score does not necessarily indicate better translation [5]. In this study, a novel MT evaluation method for the Arabic language is proposed inspired by [1]. Arabic is a morphologically dynamic language because of its word-to-word syntactical independence [34]. Even though Arabic is a widely spoken language; however, it is still considered a low resource language due to the unavailability of adequate Arabic data resources. Metrics like BLEU cannot be used directly for languages like Arabic because the structure of Arabic scripts is different from English and European languages. The widely used BLEU metric implements a brevity penalty [35] for short sentences; however, longer sentences are improperly penalized. To overcome this issue, a sentence length penalty factor is introduced to penalize the shorter and longer sentences in comparison to the reference translations. There are three types of length penalties:

- a) When the candidate sentence (translation) length is the same as the reference sentence, then there is no penalty, and LP is one.
- b) When the candidate sentence length is less than the reference sentence, the penalty is computed using (1):

$$LP = \exp(1 - r/c) \quad c < r \quad (1)$$

- c) When the length of the candidate sentence is greater than the reference sentence, the penalty is computed using (2).

$$LP = \exp(1 - c/r) \quad c > r \quad (2)$$

Where c and r represent the lengths of both candidate sentence and reference sentence, respectively

Furthermore, this work introduced another penalty for the difference between positions of different *n-grams*, where candidate sentences are penalized based on comparing different word positions with respect to the reference sentence. If all the word positions are the same, the penalty is applied, and its value is 1. The penalty value varies between 0 and 1. When no positions match at all, the maximum penalty (of value 0) is applied. In (3) is used to compute the position-based penalty.

$$npd_score = \exp|PD| \quad (3)$$

Where PD denotes position difference and npd_score represents the position difference score. The position difference penalty is built on the length penalty and they are in direct proportion to each other. An increase in length penalty will lead to an increase in PD penalty. The proposed method also computes the common word's score by measuring the ratio between word overlap (i.e., N_c) between the reference and candidate sentence to the total number of unique words (i.e., T_w), as shown in (4).

$$Cw_score = N_c/T_w \quad (4)$$

Parts of speech (POS) tags show the syntactic relation between two sentences. The POS tags score for Arabic, defined in [36], is added to the proposed method for capturing the syntactic structure of the translated sentence with the reference sentence. Arabic is a linguistically rich language. Due to the linguistic richness of Arabic, a context can be represented in different ways. Therefore, when the candidate sentence is contextually similar to the translated sentence, but the words in the candidate sentence are not the same at some positions but are synonyms, it will result in a position penalty. Thus, POS tags are used in this study to normalize that penalty term. The POS tags will help the model penalize the penalty term by comparing it with the POS tags of the reference sentence. Finally, the length penalty, position difference penalty, and the POS score are aggregated, as shown in Figure 1. The final score is computed using (5). Table 2 show some samples used in the proposed evaluation model.

$$Final_score = LP * npd_score * cw_score * pos_accuracy \quad (5)$$

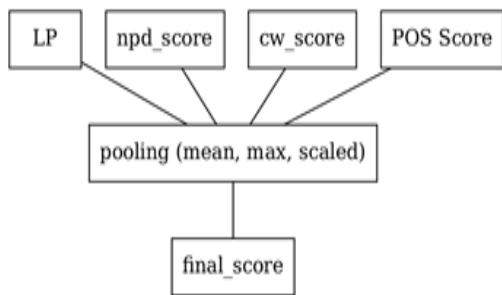


Figure 1. Model architecture

Table 2. Results of LP, position difference and POS tags model

Candidate Sentence	Reference Sentence	Final score
الطائرة تطير The plane is flying	كانت الطائرة تحلق The plane was flying	0.7
السماسرة يعملون brokers are working	السماسرة يتعاملون brokers deal in	1
السفينة تبحر ship sailing	الطائرة تطير بسرعة كبيرة The plane is flying very fast	0

4. BERT BASED SENTENCE SIMILARITY

In 2018, Devlin *et al.* [37] introduced the bi-directional encoder representation for transformers (BERT) to compute context-based word vectors based on the encoder part of [38], i.e., an attention-based neural MT model. It captures word vectors based on the context where the word is used. Masked language modeling (MLM) and next sentence prediction (NSP) are the two pre-training tasks of BERT which help the model learn good syntactic and semantic language representation.

The BERT model is set as state-of-the-art for many NLP tasks, which makes it important to utilize it in this evaluation as well. A considerable amount of data and computation power is required for language model pre-training. The language model cannot be trained from scratch because of the lack of computing resources and data. In this study, the distilled multilingual BERT model is used that supports over 100 languages. This model does not have high accuracy in low resource languages such as Arabic and Sindhi [39]. To overcome that, the multilingual BERT model is fine-tuned on a corpus of Arabic texts containing 1 million clean Arabic sentences. The data is scraped from different Arabic News resources and blogs, and spacy is used for the necessary pre-processing of data. The data is normalized to remove diacritics and other unnecessary HTML tags, and URLs. Moreover, sentence tokenizer available in spacy is used to split documents into sentences.

Once pre-training is complete, the model is fine-tuned on a custom task such as sentiment analysis, POS, and named entity recognition (NER). Since the goal is to compute the similarity between candidate and reference sentences, therefore, the data is hand-curated for this task. The Arabic linguistic experts validated the data for sentence similarity tasks by looking at the syntactic and semantic features of the sentence. A total of 10,000 Arabic cleaned sentences are collected and tweaked so that their context remains the same, but they can be represented with different words or helping verbs. Moreover, 10,000 different sentences are also selected, and random sentences are put against them. The sentences with little tweaking are labelled similar, and the random sentences are labelled not similar, as shown in Table 3. This fine-tuning used a Siamese triple loss network. The architecture of the Siamese network is depicted in Figure 2.

Table 3. Sentence similarity dataset

Sentence1	Sentence2	Label
كان يركض was running	هو يركض he runs	Similar
إنه طبيب he's a doctor	الحقيقة صعبة The truth is hard	Not Similar

This architecture computes the Softmax loss for both the labels, which are similar and not similar, where u and v are the two vectors (each generated from candidate and reference sentence), fed to the similarity algorithm to calculate the similarity between the two. After that, the proposed model is fine-tuned on the semantic textual similarity (STS) dataset [40]. This dataset is freely available on the Stanford text similarity benchmark dataset. For this study, the dataset is translated from English to Arabic and the wrong translations are manually corrected with the help of linguistic experts. This dataset is available for regression tasks, and the score is assigned from 0 to 5 for each sentence pair.

Next, the last Softmax layer is changed, and the model is fine-tuned on the sentence similarity task using cosine similarity. The labels are normalized between -1 and 1 for computing the cosine score, where 1

refers to completely similar and -1 denotes the opposite. This finalizes the proposed model for computing cosine similarity for Arabic sentences. The reported results are provided in Table 4.

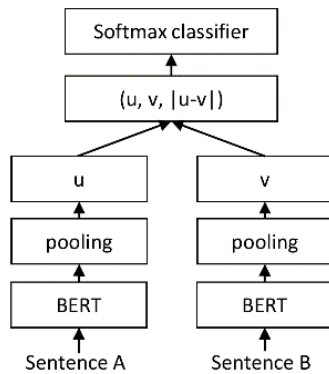


Figure 2. Siamese network for softmax

Table 4. BERT sentence similarity results

Candidate Sentence	Reference Sentence	Cosine Similarity
الانتقال من مكان إلى آخر Moving from one place to another	انتقل من مكان إلى آخر Move from one place to another	0.8
زنزانة للسجناء العنيفين A dungeon for violent prisoners	سجن للسجناء العنيفين Prison for violent prisoners	0.82
قم بالجرى بوتيرة سريعة بشكل معتدل Run at a moderately fast pace	اركض بوتيرة سريعة على الفور Run at a fast pace instantly	0.61

5. EXPERIMENT SETUP AND RESULTS

All the experiments were performed on a powerful GPU machine having an NVIDIA 24 GB GPU and CPU having RAM of 128 GB. The POS model training took 4 hours, whereas BERT fine-tuning for 200k steps is completed in five days. Moreover, the BERT fine-tuning on sentence similarity tasks took 3 hours. The samples used in this research are extracted from 250 paragraphs collected from online sources that provide bilingual corpora for the English-Arabic language pairs. These are as follows:

- Reverso, which provides a huge number of bilingual texts derived from real-life contexts; and
- The UN Parallel Corpus, particularly the English-Arabic texts.

Since colloquial Arabic has more than 25 different dialects, the sample paragraphs used in this work are taken from sources that adopt Modern Standard Arabic, which is widely used in official contexts. It is worth noting that all machine-based translations are generated using Google Translator, while the reference translations are based on the sources mentioned hereinabove after the review by a translation expert. The reported results are based on MT using the Google Translator service for translating English to Arabic sentences. Results of the proposed model are compared against state-of-the-art BLEU and METEOR utilizing Pearson Correlation as a performance evaluator parameter. Pearson correlation coefficient is given by (6):

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (6)$$

where, \bar{x} , \bar{y} are mean values and s_x , s_y are square roots of variance.

The reported results show that the proposed model outperforms others, in terms of accuracy and linguistic validation, during the evaluation of machine translation tasks for English-Arabic sentence pairs. The evaluation is performed using a corpus of 2,000 English and Arabic sentence pairs. Evaluation scores are computed for all the methods and their scores are correlated with the actual scores given by human translators. A detailed comparison of varied metrics is given in Table 5. The human evaluators evaluated the machine-translated sentences against the ones provided by the human translators and provided the evaluation scores. The average of all the evaluators for the MT system came out as 0.6954. Among the various automatic evaluation of MT systems, the POS + BERT (score: 0.6549) results came closest to the human evaluations in evaluating the accuracy of MT systems. The proposed system of POS + BERT evidently outperformed the standard BLEU and METEOR metrics for machine translation evaluation in the case of the English-Arabic language pair. Moreover, adequacy and fluency are checked for translated sentences against the human labeled sentences. English sentences and their corresponding Arabic translations were presented to human experts who were asked to evaluate these sentences on a 5-point scale, i.e. (1: no sense, 2: non-acceptable, 3: acceptable, 4: good, 5: ideal). To ensure that human judgments are not biased towards a particular sentence, reference translations employed in the automatic evaluation were not disclosed to the humans. These evaluations were gathered from three different native Arabic-speaking subjects.

Scores given by the subjects are considered average as the overall human judgment for adequacy (comprehensiveness) and fluency (naturalness) as well. The respective scales for Adequacy are, none: 1, little meaning: 2, much meaning: 3, most meaning: 4, all meaning: 5; and for fluency are, incomprehensible: 1,

disfluent: 2, non-native: 3, good: 4, flawless: 5. In order to assess the performance of the proposed method in terms of adequacy and fluency; 40 paragraphs are chosen randomly and provided to human evaluators to score based on fluency (naturalness) and adequacy (comprehensiveness). After that, the scores were plotted against the BERT similarity scores and proposed metric scores. Finally, the Pearson's correlation coefficients is calculated to correlate fluency/adequacy scores and the proposed metric with POS/proposed metric with POS+BERT. The obtained results are included in Table 6. The relevant graphs are shown in Figures 3 and 4.

The obtained results validated that the proposed method outperforms BLEU and METEOR with reference to capturing the Adequacy and Fluency in candidate sentences. The proposed method considers syntactic and semantic features which other metrics lack. Furthermore, the linguistic features are considered to compute the similarity of candidate sentences with reference sentences. In contrast, BLEU and METEOR do not consider the syntactic and semantic properties in an embedding space.

Table 5. Results of correlation for different metrics: system-level score and correlations with human judgments

Metric	Scores	Coeff. of corr.
BLEU	0.5273	0.3929
METOR	0.5627	0.6142
Proposed with POS	0.6142	0.6280
Proposed with BERT	0.6522	0.6842
Proposed with POS+BERT	0.6549	0.6763
Human	0.6154	-

Table 6. Pearson correlation comparison with adequacy and fluency

	Pearson's correlation coefficient	
	Proposed metric with POS	Proposed metric with POS+BERT
Fluency score	0.6933	0.7211
Adequacy score	0.6824	0.7123

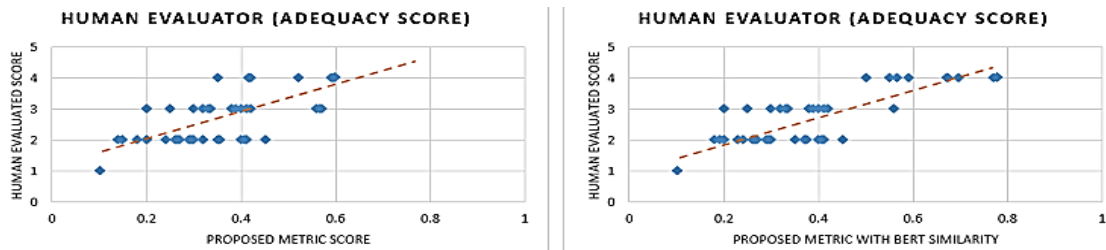


Figure 3. Adequacy score of proposed metric vs human evaluator and proposed metric with BERT similarity

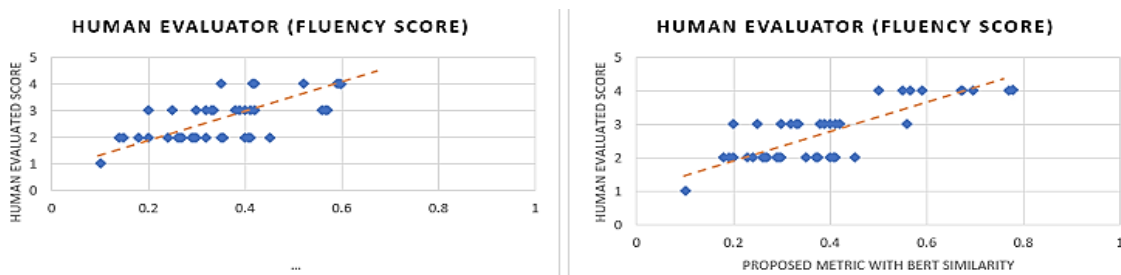


Figure 4. Fluency score of proposed metric vs human evaluator and proposed metric with BERT similarity

6. CONCLUSIONS AND FUTURE WORK

Despite the numerous challenges associated with the evaluation of MT into low-resource and morphologically rich languages, the concerns are not well-addressed in the current literature. This study is believed to set a new direction in machine translation evaluation for morphologically rich languages like Arabic. In this work, a context-based approach is utilized that goes well into the semantics of the language and understands the similarity between machine translation and human translation. The experimental results demonstrate that the proposed method surpasses the previous state-of-the-art in English to Arabic translation.

Consequently, the proposed method encourages the utilization of automatic and semantic-based methods for machine translation. However, the integration of more linguistic features and semantic embedding along with the proposed method can be explored in the future to improve the MT evaluation system.




REFERENCES

- [1] S. Tripathi and V. Kansal, "Using linguistic knowledge for machine translation evaluation with Hindi as a target language," *Computación y Sistemas*, vol. 21, p. 717–724, 2017, doi: 10.13053/cys-21-4-2869.
- [2] A. H. Aliwy and A. A. Ahmed, "Development of arabic sign language dictionary using 3D avatar technologies," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 609–616, 2021, doi: 10.11591/ijeecs.v21.i1.pp609-616.
- [3] F. Nurifan, S. Riyanto, and C. S. Wahyuni, "Developing corpora using word2vec and wikipedia for word sense disambiguation," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 3, pp. 1239–1246, 2018, doi: 10.11591/ijeecs.v12.i3.pp1239-1246.
- [4] B. Kituku, L. Muchemi, and W. Nganga, "A review on machine translation approaches," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 1, no. 1, pp. 182–190, 2016, doi: 10.11591/ijeecs.v1.i1.pp182-190.
- [5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, doi: 10.3115/1073083.1073135.
- [6] J. P. Turian, L. Shea, and I. D. Melamed, "Evaluation of machine translation and its evaluation," New York Univ Ny, 2006, doi: 10.21236/ADA453509.
- [7] E. H. Hovy, C.-Y. Lin, L. Zhou, and J. Fukumoto, "Automated summarization evaluation with basic elements," in *LREC*, 2006.
- [8] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "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*, 2006.
- [9] C. Callison-Burch, C. S. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) evaluation of machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, doi: 10.3115/1626355.1626373.
- [10] S. Banerjee and A. Lavie, "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*, 2005.
- [11] M. Denkowski and A. Lavie, "Extending the METEOR machine translation evaluation metric to the phrase level," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.
- [12] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, doi: 10.3115/v1/W14-3348.
- [13] J. S. Albrecht and R. Hwa, "Regression for machine translation evaluation at the sentence level," *Machine Translation*, vol. 22, pp. 1–27, 2008, doi: 10.1007/s10590-008-9046-1.
- [14] J. Albrecht and R. Hwa, "A re-examination of machine learning approaches for sentence-level MT evaluation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.
- [15] J. Albrecht and R. Hwa, "Regression for sentence-level MT evaluation with pseudo references," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.
- [16] S. Corston-Oliver, M. Gamon, and C. Brockett, "A machine learning approach to the automatic evaluation of machine translation," in *Proc. of the 39th Ann. Meet. of the Assoc. for Comp. Ling.*, 2001, doi: 10.3115/1073012.1073032.
- [17] A. Kulesza and S. Shieber, "A learning approach to improving sentence-level MT evaluation," in *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004.
- [18] D. Liu and D. Gildea, "Syntactic features for evaluation of machine translation," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- [19] C. Quirk, "Training a sentence-level machine translation confidence measure," in *LREC*, 2004.
- [20] Y. Akiba, K. Imamura, and E. Sumita, "Using multiple edit distances to automatically rank machine translation output," in *Proceedings of the MT Summit VIII*, 2001.
- [21] G. Russo-Lassner, J. Lin, and P. Resnik, "A paraphrase-based approach to machine translation evaluation," 2005, doi: 10.21236/ADA448032.
- [22] M. Gamon, A. Aue, and M. Smets, "Sentence-level MT evaluation without reference translations: Beyond language modeling," in *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, 2005.
- [23] Y. Ye, M. Zhou, and C.-Y. Lin, "Sentence level machine translation evaluation as a ranking," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, doi: 10.3115/1626355.1626391.
- [24] K. Duh, "Ranking vs. regression in machine translation evaluation," in *Proceedings of the Third Workshop on Statistical Machine Translation*, 2008, doi: 10.3115/1626394.1626425.
- [25] S. Gautam and P. Bhattacharyya, "Layered: Metric for machine translation evaluation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014, doi: 10.3115/v1/W14-3350.
- [26] M. Akeel and R. B. Mishra, "Divergence and ambiguity control in an English to Arabic machine translation," *the International Journal of Engineering Research and Applications*, vol. 3, pp. 1670–1679, 2013.
- [27] E. A. S. Abu-Ayyash, "Errors and non-errors in English-Arabic machine translation of gender-bound constructs in technical texts," *Procedia Computer Science*, vol. 117, pp. 73–80, 2017, doi: 10.1016/j.procs.2017.10.095.
- [28] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative study between METEOR and BLEU methods of MT: Arabic into english translation as a case study," *International Journal of Advanced Computer Science and Applications*, vol. 6, pp. 215–223, 2015, doi: 10.14569/IJACSA.2015.061128.
- [29] J. Zakraoui, M. Saleh, S. Al-Maadeed, and J. M. AlJa'am, "Evaluation of Arabic to English machine translation systems," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, doi: 10.1109/ICICS49469.2020.239518.
- [30] F. Guzmán, H. Bouamor, R. Baly, and N. Habash, "Machine translation evaluation for Arabic using morphologically-enriched embeddings," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.
- [31] H. Shimanaka, T. Kajiwara, and M. Komachi, "Ruse: Regressor using sentence embeddings for automatic machine translation evaluation," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018, doi: 10.18653/v1/W18-6456.
- [32] C. Callison-Burch, C. S. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," in *Proceedings of the third workshop on statistical machine translation*, 2008, doi: 10.3115/1626394.1626403.
- [33] M. S. Maučec and G. Donaj, "Machine translation and the evaluation of its quality," in *Recent Trends in Computational Intelligence*, IntechOpen, 2019.




- [34] M. Z. Dendane, "An Overview of Verb Morphology in Arabic" *Revue Maghrébine des Langues*, vol. 5, no. 1, pp. 338-361, 2007.
- [35] M. Neishi, J. Sakuma, S. Tohda, S. Ishiwatari, N. Yoshinaga, and M. Toyoda, "A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, 2017.
- [36] Y. E. Hadj, I. Al-Sughayeir, and A. Al-Ansari, "Arabic part-of-speech tagging using the sentence structure," in *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2009*.
- [37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [38] A. Vaswani, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017.
- [39] J. C. B. Cruz and C. Cheng, "Evaluating language model finetuning techniques for low-resource languages," *arXiv preprint arXiv:1907.00409*, 2019.
- [40] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "SEM 2013 shared task: Semantic textual similarity," in *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 2013.

BIOGRAPHIES OF AUTHORS






Majdi Beseiso    received his BSc degree from Jordan University of Science & Technology, Jordan. He obtained his MSc degree in Computer Science from the University of Jordan, Jordan in 2006. In 2013, he received his Ph.D. degree in Information & Communication Technology from Tenaga National University, Malaysia. Currently, he is an assistant professor in the Computer Science Department at Al-Balqa Applied University, Jordan. Dr. Beseiso's research interests include natural language processing, machine learning, and information retrieval. Dr. Beseiso is the dean assistant for planning, development and quality at Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology. He can be contacted at email: bsaiso@bau.edu.jo.






Samiksha Tripathi    is PhD in Computer Science specializing in Artificial Intelligence and NLP domains. She is a 'Principal Architect – Data Science and Cognitive Analysis' in R Systems International's R&D department. She previously worked as an Innovation Consultant at Cotiviti Labs, Atlanta, USA (R & D). Samiksha has been a senior research associate at Language Technology Lab, IIT Kanpur specializing in Machine Translation for Indian Languages. She can be contacted at email: samiksha.tripathi@gmail.com.



Bashar Al-Shboul    is an Associate Professor with the Department of Information Technology/The University of Jordan. His main research interests include: information retrieval, natural language processing, data science and artificial intelligence. He worked as a consultant and a regional developer for government/international projects, as well as the private sector. Dr. Al-Shboul received his B.Sc. degree in Information Technology from Al-Balqa'a Applied University (Jordan) in 2003, the M.Sc. degree in Computer Science from The University of Jordan in 2006, and the Ph.D. degree in Information and Telecommunications Technology Engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2012. He can be contacted at email: bashar.shboul@gmail.com.



Renad Aljadid    received her BA degree in English Language & Translation in 2018 from Prince Sultan University in Riyadh, Saudi Arabia. She then obtained her MA degree in Translation in 2021 from the University of Jordan, Jordan. She has been working in the professional translation industry using translation technology for 4 years so far. Her research interests include translation technology, machine translation, and translation theories. She can be contacted at email: renad.aljadid@gmail.com.