QUES: A Quality Estimation System of Arabic to English Translation

Manar Salamah Ali[1], Anfal Alatawi[2], Bayader Alsahafi[3], Najwa Noorwali[4]
Computer Science Department
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Estimating translation quality is a problem of growing importance as it has many potential applications. The quality of translation from Arabic to English is especially difficult to evaluate due to the languages being distant languages: different in syntax and low in lexical similarity. We propose a feature-based framework for estimating the quality of Arabic to English translations at the sentence level. The proposed method works without reference translations, considers both fluency and adequacy of translations, and does not imply assumptions on the source of translation (humans, machines, or post-edited machine translations); thus, making the solution applicable to increasingly more situations. This research solves the translation quality estimation problem by treating it as a supervised machine learning problem. The proposed model utilizes regression algorithms (SVR and Linear Regression) to predict quality scores of unseen translated texts at runtime. This is accomplished by training models on a labeled parallel corpus and mapping extracted features to the quality label. The prediction models succeeded in predicting fluency and adequacy of translations with a Mean Absolute Error of 0.84 and 1.02, respectively. Furthermore, we show that in a similar setting of our approach, fluency of an Arabic to English translated sentence on its own, is an appropriate indication of a translation's overall quality.**

Keywords*—Translation quality estimation; translation adequacy; translation fluency; supervised machine learning

I. Introduction

A good quality translation plays an important role in transferring knowledge and it has great impact on the global economy by allowing businesses to grow globally without the inconveniences of language barriers. In addition, it informs the end-users about the reliability of a translated content and helps in determining if a translated text is ready to be published or if it needs further editing.

Hence, the ability to assess the quality of a translation is critical in order to guarantee its effectiveness in information delivery. The task of assessing the quality of translated content and its appropriateness for use or publishing is usually performed by experts in translation. It is also done automatically using machine translation evaluation systems. The former approach can be very expensive and time-consuming, while the latter requires reference translations in order to perform a comparison operation and evaluate the translation quality. A reference translation is a manually developed translation by an expert translator that is essential in automatic translation evaluation. Automatic evaluation is done by different measures of comparison between the produced translation and the reference translation [14]. Reference translations are expensive and require manual labor and time to produce, thus this approach is also proved costly.

Translation Quality in this research is defined as a function of translation fluency and translation adequacy. As defined in [38], fluency indicates "how well the produced translation is grammatically fluent and natural in the target language" while adequacy indicates "the semantic equivalence between a source sentence and its target translation". In other words, adequacy demonstrates how well the produced translation conveys the same meaning as the original text.

The lack of automated tools that estimate the quality of Arabic to English translation has been the main motivation behind this research. By using such a tool, it will be possible to assess the quality of Arabic translated content and it would be possible to make suitable remedial actions to the translated content accordingly.

This research aims to develop a translation Quality Estimation model (QUES) for Arabic to English translations that requires no access to a reference translation or the source that performed the translation. In addition, this research aims to suggest a measure of the adequacy and fluency of a translation from Arabic to English translations. The goal is to inform an end-user about the reliability of a given translation from Arabic to English at the sentence level.

This paper is organized as follows: In Section II, an overview of translation quality is discussed and is followed by a discussion on translation quality of machine translations in Section III. Human translations are briefly discussed in Section IV. Section V presents the Quality Estimation (QE) of translations without a reference and the QE granularity levels. Related work is discussed in Section VI. The dataset, feature sets, experimentation, and results of our model are discussed in Sections VII, VIII, and IX, respectively. Finally, the conclusions of this paper and future work are discussed in Sections X and XI, respectively.

II. Translation Quality

Translation is the process of replacing or converting a source text (ST) that is written in a source language (SL) into target text (TT) in a target language (TL) [22, 32]. To call a translation an equivalent translation, the TT must be functionally equivalent to the ST [12, 22], which describes a semantically and pragmatically equivalent texts and holds whenever a TT have the same communicative effect as the ST.

Furthermore, a translation can either be literary oriented or linguistically oriented. The literary oriented approach is used to translate literary texts. This is done by manipulating the TT to fit the literary and cultural context of the TL, without giving much emphasis on the relationship between the ST and the TT. Whereas the linguistically oriented approach gives significant emphasis on the relationship between the ST and the TT by considering the functional equivalence [22].

How to determine whether a translation is good or bad is one of the intriguing questions that is connected with any translation. Researchers agree that there is no unified method to measure the quality of a translation [11, 15]. However, long-standing studies and models with different evaluation measurements have been developed over recent decades. In [42], the author states three concepts involved in most of the established models and criteria of Translation Quality Assessment (TQA). The first concept is the quality of the producer (human or machine). The second one is the quality of the process, which includes how the process was predefined, and whether it was followed in order to obtain a good translation. The last one is the quality of the product that includes predefined evaluation standards. Nonetheless, current models aim at focusing on the quality of the product [3].

A plethora of TQA approaches has been developed over the years. Some researchers follow a non-linguistic approach where the features and relations between the original text and the translated text are not considered but rather the focus is on the translation's psychological or behavioral effects on the receivers of the translated text [21]. Linguistic approaches, on the other hand, compare both ST and TT according to various criteria such as coherence, semantics, and syntax. However, these criteria might differ based on the evaluation process. Most of the recent works are taking linguistics-oriented approaches into account [21]. In [3], the author states two main approaches that promise to provide an objective assessment of translation quality: error-based approach and holistic approach. Error-based approaches aim to measure only the errors or defects in a translation. The holistic approach which was first proposed by [43], considers both negative and positive aspects of the translation. Holistic models can be classified into two categories [27]: equivalence-based and nonequivalence-based approaches. In equivalent-based approaches, similarities between the ST and TT are tested and evaluated such as linguistic and narrative structures, overall textual volume and layout, coherence of thematic structures, lexical properties, and grammatical\syntactic equivalence [33].

On the other hand, non-equivalence-based approaches focus on different concepts such as text function and purpose. In [8], the author proposed a model that used assessment parameters or evaluation standards, which were adapted from different linguistic scholars such as [20, 45]. These parameters are as follows: a) the sufficiency and adequacy of the translation based on a semantic and formal language level, b) purpose: whether the translated text is appropriate for the intended purpose, c) context: considering factors such as the target audience, the time and place in which the translation is used, and the text type, and d) language norm: the fluency of the translation such as syntax, grammatical mistakes, spelling mistakes, and punctuation mistakes.

III. Machine Translation Quality

Machine translation (MT) refers to fully automated machines that translate a source text in a natural language into a target text in another natural language. There are three different approaches to machine translations: a) Rule-based approach [17], Statistical MT (SMT) [7], and Neural MT [44].

The central idea behind MT evaluation is assessing the degree of proximity of a translated text to a human translated text. Many methods to evaluate MT were proposed. The most three common methods to evaluate MT performance are discussed here.

The Round Trip Translation Assessment is performed by taking the translated TT and translating it again using the MT system to produce what is called the backward translation, and evaluating the MT system based on how close the backward translation is to the ST [13].

Human evaluations of the MT systems are conducted manually by translation experts. The adequacy and fluency measures are scored through various scales: (1-5), (1-7) or even (1-9) [28]. Although human evaluations of machine translation are expensive, they are very expensive. Therefore the field of automated MT evaluation emerged.

Automated evaluation systems are based on the measure of similarity between a text and a reference human translated text [35]. The most used evaluation measures are BLEU [35], NIST [14], METEOR [4], and TER [34]. For example, BLEU measuring rubric use a weighted average of variable length phrase matches against the reference translations by comparing the n-grams of both the translation and its reference translation[35]. The range of BLEU scores range between 0 and 1 where scores greater than 0.30 means that the translation is understandable while scores greater than 0.50 reflect much better fluent translations [1].

In summary, most current MT evaluation metrics are based on comparisons between machine translations and human references and are based on evaluating the lexical similarity at the n-gram level. There are two challenges for the automatic MT evaluation methods [30]: a) the use of reference translations which are costly from an economic perspective, and b) the focus on fluency of the output text which lacks the integration of semantic information in MT. This has led to MT systems that are illiterate in terms of semantics and meaning [30]. To solve these problems, the authors in [38] proposed to perceive the MT evaluation problem as an adequacy estimation problem and replace the use of reference translation by quality indicators for unseen translated sentences.

IV. Human Translation

Human Translation (HT) is the process of translating source text in one language into target text in another language which is performed by humans. The separation between HT and MT is increasingly indistinct nowadays due to the availability and widespread of Computer-Aided Tools and accordingly it is not possible to distinguish between a translated text produced by humans, machines, or post-edited machine translations [11]. Consequently, researchers argue that approaches and measures for evaluating translation quality

could be unified [11]. Some work has been dedicated to investigating the correlation between MT and HT of the same source text. It has been found that there is a strong correlation in English-to-Spanish [10] and English-to-Arabic [2]. Hence, in this research, assumptions about the source of translation are omitted and a non-equivalence based approach is followed which can be applied to HT and MT at the same time.

V. Translation Quality Estimation

Translation Quality Estimation (QE) is an automatic evaluation framework that avoids the use of reference translations. It aims to provide a quality indicator for machine-translated sentences at various levels (word level, sentence level, document level) [38]. Translation quality estimation is generally addressed using Machine Learning (ML) techniques to predict quality scores [9, 23, 24, 46]. The most common method in these approaches is considering the problem as a supervised learning task using standard regression or classification algorithms to predict various quality labels. QE solves the challenges in MT evaluation by adapting cross-lingual semantic inference capabilities and judging a translation [30]and utilizing machine learning to infer the relationship between texts and their corresponding quality label from the training data. Attempting to extract features that represent the adequacy of a translation, rather than the fluency of the target text alone.

Adequacy in the context of translation is defined as "semantic equivalence between source sentence and target translation", in other words, an adequate translation is a translation that preserves the meaning of the input text and does not add any information to it [38]. Fluency, on the other hand: is the grammatical correctness of a target translation, in other words, a fluent translation, is said to be a grammatical and naturally occurring text in the target language. Mostly, both adequacy and fluency are the two most desirable features for a correct translation [30]. MT Evaluation metrics rely entirely on the fluency of the produced text (target text), which proved to be a weak point. In an effort for more robust quality estimation systems, the author in [38] proposed considering adequacy in estimating the quality of a translated text.

QE is done on various granularity levels, on the word level, the sentence level, or on the document level. Granularity-level refers to the type of portion of text the QE system is trained on and therefore is expected to evaluate. Sentence-level QE was the first form of QE [6], the QE system is trained on translated sentence pairs in order to evaluate sentence pairs at runtime. As for document-level QE, it consists of predicting the quality of text sizes larger than sentences: document at a time. For word-level QE, the system is trained and expected to run on individual words. The use of word-level QE is to highlight the specific words that need editing or to inform readers which parts of the sentences are not reliable, among other uses.

VI. Related Work

In this section, works related to the quality estimation problem are discussed. The approaches may vary in many aspects such as the source and target languages, the criteria for evaluation, the machine learning models, or the text granularity level.

A fair amount of research and progress in QE has been led by the shared task competitions proposed at WMT (the Workshop on Statistical Machine Translation) [6]. The WMT started in the year 2006, and every year, introduced different tasks centered on Statistical Machine Translation (SMT) topics that vary in purpose. Some of the tasks are translation tasks, challenging participants to produce SMT systems that produce results better than the baseline SMT system provided. Other tasks challenge participants to produce new MT evaluation metrics (like BLEU, TER, etc.). At the 6th round of WMT: WMT12, a new task was proposed (which is the interest of this research) motivated by the recent work in considering adequacy in translation quality prediction [38]. The task was named Quality Estimation, and it is regarded as the first emergence of QE as it is now known. The tasks started with estimating the quality of the translation produced by MT systems. Utilizing features describing how the MT system works, for example, its confidence levels in the translation and the language model used. But the task of QE evolved over time, after proving the ability to achieve great results needless of the details of the MT system that produced the translation [5]. The irrelevance of the source of the translation allows the QE framework to solve more general problems, thus allowing it to be more applicable in increasingly more situations. For example, the case where more than one MT produced a given text, or that the text was translated by humans.

TABLE I. Review of Relted Work

Reference	Language Pair	Source of translation	Granularity Level	Quality criteria	Features	ML	Metrics
[26]	English to Spanish	MT	Sentence	Post editing effort	Baseline-Latent Semantic Indexing	Support Vector Machine	HTER
[46]	English to Chinese	HT	Sentence	Fluency - Adequacy	Monolingual features, bilingual features, language modelling features, and bilingual embedding	Support Vector Machines-Relevance Vector machines	scale of 60 points
[24]	English to German	MT	Word\sentence\phrase	-	POS taggers	Predictor-Estimator Neural Model + stack propagation	BAD\OK
[23]	English to French\Spanish\Russian	Neural MT (NMT)	Sentence\document	-	Black box features + Baseline features from Quest++	Bi-directional Recurrent Neural network (bi-RNN)	HTER for sentence level & BLUE for document level

As mentioned in Section V, one common approach is to treat the problem as a supervised learning task. Recently, neural networks have been used to improve the performance of QE. A review of some recent works is listed in Table I.

VII. QUES SETUP

In the following subsections, the experimental setup of QUES is demonstrated in detail.

A. Dataset

It is very important for non-English universities to provide high quality translated web content in English to attract international students and scholars. In addition, educational web ranking institutions such as Webometrics and USNews are concerned with the English content on the websites of universities. The aforementioned issues have motivated the choice of the dataset domain of this work. In this work, we are interested in evaluating the quality of translations from Arabic to English in university web pages.

Part of the corpus is collected from the Open Parallel Corpus collection (OPUS) [22, 41]. The OPUS is a collection of translated text sentences from various resources that provide parallel corpora that are aligned and linguistically annotated. The Wikipedia Corpus is a collection of translations published by the Wikimedia Foundation and their article translation system [41]. In order to align the collected data with the domain of this research, the data was filtered by the educational domain. The filtering was performed automatically using keywords which indicate an educational content in the sentences, and hence only sentence pairs (corpus instances) including educational content are selected. In order to enrich the dataset, a second source of corpus was collected by volunteers from the King Abdulaziz University's E-portal website. The total number of sentence pairs in both data sets is 5571 instances.

In this work, both adequacy and fluency are the features used to estimate the quality of correct translations, and they both were used for labeling the data. In [38], researchers introduced adequacy in a 4-point scale (1-4) measure. However, this scale was complex for a learning model to distinguish between 3 & 4, and 2 & 3, which requires more complex features. Therefore, in this research, adequacy, and fluency are measured using a 5-point scale [25]. The definition of adequacy scale is defined as follow:

5. All Meaning expressed in the ST appears in the TT.
4. Most Meaning of the ST is expressed in the TT
3. Much Meaning of the ST is expressed in the TT
2. Little Meaning of the ST is expressed in the TT.
1. None of the meaning expressed in the ST is expressed in the TT.

While fluency scale is defined as follows:

5. Flawless English: no grammar errors, good word choice, and syntactic structure in the TT.
4. Good English: few terminology or grammar errors that do not impact the overall understanding of the meaning.
3. Non-native English: about half of the translation contains errors.
2. Non-fluent English: wrong word choice, poor grammar, and syntactic structure.
1. Incomprehensible: absolutely ungrammatical and for the most part doesn't make any sense.

The first attempt for labeling the datasets was conducted by senior students in the translation department. But due to the low quality of the work, the datasets were then labeled by professional translators. The labeling task was provided to the labelers through Dataturks tool, which is an online platform that provides collaborative data labeling, annotation, and segmentation.

B. Feature Sets

After comprehensive research in the literature, the following features have been selected for this work: Black box features (BBF) [16], Baseline Features (BF) [39], Fluency features (FF) [38], Adequacy Features (AF) [38], and Word Vectors (WV) [18].

The black box features are extracted from the source and target texts only such as sentence length, punctuation, type-token ration, and PoS tagging on source and target texts. This feature set is a group of 61 black box features which applies to the Arabic and English language pair.

The baseline features are a set of 17 features that quantify the complexity of ST and TT such as the number of tokens in the sentences, number of punctuation marks, language model probability in source and target language. Those features have been proven to produce good results on many language pairs [39]. Hence, the goal in this work is to test if it performs well on Arabic and English text pairs.

Fluency features are extracted from target text such as translated sentence length and coherence. For example, one of the features calculates the absolute difference between no tokens and source and target normalized by source length.

Adequacy features are features extracted from target and source text such as the ratio of the number of tokens in source and target text and the ratio of the percentage of numbers and non-content words in source and target text.

Finally, word embedding, which is an approach used in natural language processing to map words in a text to vectors of real numbers is used. Word embedding is used here to represent sentences in a format that can be fed into a machine learning model while preserving the word order and the meaning of the sentences. Working with word embedding gives us the advantage of measuring the contextual similarity between two corresponding sentences, which is a great indication for both adequacy and fluency.

For feature extraction, the first framework used is QUEST++ [40]. QUEST++ can extract a set of 130 system-independent features given the source text, target texts, and a set of auxiliary tools like POS tagger and Language Models. Features are extracted from target text to measure text fluency and from both the source and target texts to measure adequacy. While for word embedding, FastText[1] is used, which is an

[1] https://fasttext.cc/docs/en/crawl-vectors.html

open-source library from Facebook that provide word2vec models of continuous bag of words (CBOW) and Skip-gram. It also provides pre-trained models for 147 languages, including English and Arabic. In addition, the library has an aligned version of 44 word2vec models that are aligned in the same vector space, which means that the representation of each vocabulary in one language would be very similar to the representation of the vocabulary translation in a different language [18, 31]. In this work, pre-trained CBOW aligned models are used for Arabic and English with a vector dimension of 300; since all the provided vectors of the aligned models are of equal length.

VIII. EXPERIMENTATION

The general architecture of a QE system shown in Fig. 1. In this research, the different ML algorithms are treated as black-box algorithms. The algorithms used are predefined in the SKLearn library [37]. The model parameters are to be chosen experimentally, using grid search and cross-validation. Different ML algorithms such as Support Vector Regression SVR and linear regression are used in order to compare and contrast results, and to produce meaningful insights on how the features correlate with the scores.

The different combinations of the QUES components are tested and evaluated, and multiple experiments are conducted. In each experiment, a different combination of features is extracted which produces a development data set of selected features extracted from selected data sources. After that the development data is split into a ratio of 80% for training and 20% for testing (evaluation). Then an ML algorithm is trained on the training data.

QUES is built using the Jupyter notebook environment. To facilitate the development of a machine learning model, the state of the art open-source library Sci-kit Learn [37] is used, due to the availability of different ML algorithm implementations, and the frequent use of Sci-kit Learn in the development of QE systems [6]. In addition, the following Python libraries are used: Gensim, NLTK, SKLearn, TensorFlow, Pandas, and Numpy.

In order to perform feature extraction, a variety of feature extractors from source texts, translated texts, external resources, and tools are used from QUEST++[40]. QUEST++ requires additional tools for the processing of the Arabic language, the additional libraries (the POS tagger, language model, etc.). Hence, Stanford OpenNLP [29] is used for segmentation, language model, POS tagging, dependency parsing and English NER,. And for Arabic NER, Madamira [36] is used. Named Entities Recognition (NER) parses unstructured text and classifies named entity mentions into predefined categories such as names, organizations, locations, etc [19].

The preservation of Named Entities (NE) is one of the desirable characteristics of a correct translation[38]. Especially when translating domain-specific texts in which it is crucial to preserve named entities. Some of the features that the framework of QE allows to consider are based on matching the number and categories of named entities in the source and target sentences.

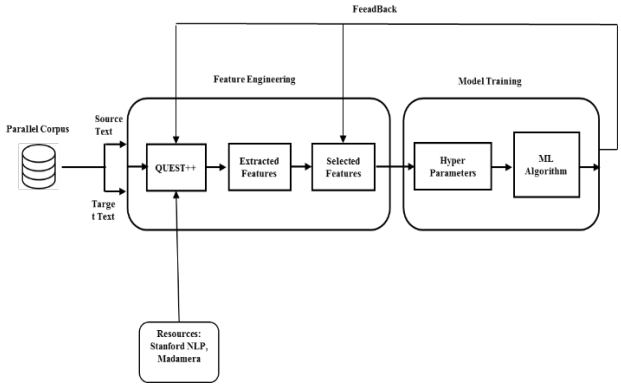

Fig. 1. General Architecture of QUES.

The extracted features were run through different ML algorithms to train different models. The models are then evaluated. Each evaluation represents an experiment, and each experiment tested the possible features' performance in predicting both the Adequacy and Fluency labels. The evaluations are measured in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as defined in Equations 1 and 2, respectively[11].

$$\text{MAE} = \frac{\sum_{i-1}^{N} |H(s_i) - V(s_i)|}{\text{N}} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(H(s_i) - V(s_i)\right)^2}{N}} \quad (2)$$

Where

N= is the number of test instances

$H(s_i)$ is the predicted score for instance s_i

$V(s_i)$ is the labeler score for s_i

Two sets of experiments were conducted, the first batch of experiments was conducted on individual features sets (section VII). The second batch of experiments was conducted on combined feature sets. The results of experiments and their evaluations in MAE and RMSE are listed in Tables II and III.

The second batch of experiments was set as follows:

1) Combining (BF+FF) features in an attempt to measure if increasing the number features representing fluency (in addition to a base set with features representing basic attributes both fluency and adequacy) would improve the accuracy of the predictor.

2) Combining (BF + AF) features to investigate if increasing the number of features representing adequacy (to a base set with features representing basic attributes both fluency and adequacy) has an effect on predicting translation quality of sentences.

3) Combining (AF + FF) to evaluate the model where features represent solely the adequacy and fluency of a translation pair.

4) Experimenting with all the applicable black-box features, in a group named BBF. This was done to investigate the effect of features representing all possible linguistic

attributes of a translation text pair on the ability to predict translation quality.

5) The final combination set is the CF (Correlated Features), which is a set of 14 features that showed the highest correlation with the two labels.

IX. Results

In this section, the results of the experiments and the observations are discussed. First, all the sets of features produce similar results with slight variations. Second, it is observed that when only features representing fluency are extracted from the sentences, the model produces accurate results in predicting the adequacy of a translated sentence pair. This leads to the conclusion that fluency and adequacy are highly correlated in Arabic to English translations. That is also the case when using features that represent only adequacy to predict fluency. As shown by the AF and the BF entries in Table I, where both experiments produce accurate and similar results for both labels. Further calculations have been conducted to measure the correlation between the two labels, and it has been found that it reaches as high as to 0.8.

While regarding the best performing models, the results are interpreted as such: an MAE of 1.0 means that the model is of the true label by 1.0. For example, let assume that for a sentence with a true fluency label of 3.0, the system predicts a prediction label of 4.0 or 2.0. That is an acceptable result, since the labels are continuous values, and a 2.0 or 4.0 label is not far off in meaning from a label of 3.0. Third, it is observed that the best indicator for a sentence pair's overall quality is the target sentence's fluency, as the entry FF in Table I shows. The explanation behind this is the tendency for well-translated sentences to be fluent in our data set, and therefore in the real world. This observation also shows that the sentences that are dis-fluent tend to be inadequate. Fourth, regardless of the slightly better results produced by the FF group, it's observed that different feature combinations in Table III produced very similar results. The researchers believe that this is due to the size of data, which is not large enough to detect noticeable differences in performance between the different feature sets. Finally, it appears that training the models on features of word vectors doesn't perform well in Arabic-English pairs. The attempt to vectorize long sentences using a 300 long vector for each word, made the sentence and its translation reach a vector length of 53100. This caused a data sparsity problem.

X. Conclusion

This research studied the problem of translation quality estimation. The system proposed in this work has succeeded in predicting two important quality measures, fluency, and adequacy, with the best models producing a mean absolute error of 0.84 and 1.02, respectively. That proves that the features that extract the best quality indicators from text are the features representing fluency. That is due to the observation that adequate sentences tend to be fluent. This means that, generally, when a sentence is translated from Arabic to English, a low-quality translation tends to be dis-fluent in the English language. On the opposite side, high-quality translations tend to be fluent. The results of the experiment show that it is rarely the case that sentences that are adequate lack fluency. Therefore, this research concludes that in a similar setting of this work, fluency of a translated sentence on its own is an appropriate indication of a translation's overall quality.

XI. Future Work

One of the areas of improvement in this research is to increase the size of the data set. The collection, filtration, and labeling of data is a costly process, and the researches reached an economic limitation as a result of it. The researchers believe that acquiring more data will produce more accurate results as is often the case for machine learning systems. As well as allow the testing of deep learning techniques, as it solves the problem of data sparsity. Another area of improvement is to combine the labels of Adequacy and Fluency, in one general quality measure, as this research concludes that they are highly correlated, and features for one label predict the other label efficiently.

TABLE II. Experiment Results with Individual Feature Sets

Feature Set	ML Algorithm	Evaluation			
		Fluency		Adequacy	
		MAE	RMSE	MAE	RMSE
BF	Linear Regression	0.96	1.19	1.11	1.40
	SVR	0.97	1.36	1.10	1.41
AF	Linear Regression	0.96	1.187	1.11	1.43
	SVR	0.98	1.37	1.14	1.49
FF	Linear Regression	1.03	1.22	1.10	1.41
	SVR	0.84	1.43	1.02	1.42
WV	Linear Regression	2066.20	2717.49	1141.38	1485.04
	SVR	0.83	1.47	1.11	1.41

TABLE III. Experiment Results with Combined Feature Sets

Feature Set	ML Algorithm	Evaluation			
		Fluency		Adequacy	
		MAE	RMSE	MAE	RMSE
BBF	Linear Regression	0.96	1.21	1.16	1.45
	SVR	1.00	1.35	1.17	1.49
BF+AF	Linear Regression	0.95	1.19	1.06	1.36
	SVR	0.94	1.18	1.03	1.32
BF+FF	Linear Regression	0.95	1.20	1.07	1.36
	SVR	0.95	1..30	1.10	1.40
AF+FF	Linear Regression	0.96	1.18	1.11	1.42
	SVR	0.98	1.37	1.12	1.50
CF	Linear Regression	1.01	1.21	1.13	1.42
	SVR	0.89	1.33	1.14	1.47

REFERENCES

[1] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating English to Arabic machine translation using BLEU," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 1, 2013.

[2] S. A. Almazroei, H. Ogawa, and D. Gilbert, "Investigating Correlations Between Human Translation and MT Output," Machine Translation Summit XVII, p. 11, 2019.

[3] M. O. L. Almutairi, "The objectivity of the two main academic approaches of translation quality assessment: Arab Spring presidential speeches as a case study," University of Leicester, 2018.

[4] 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, pp. 65-72.

[5] E. Biçici, "Referential translation machines for quality estimation," in Proceedings of the eighth workshop on statistical machine translation, 2013, pp. 343-351.

[6] O. Bojar et al., "Findings of the 2014 workshop on statistical machine translation," in Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 12-58.

[7] P. F. Brown et al., "A statistical approach to machine translation," Computational linguistics, vol. 16, no. 2, pp. 79-85, 1990.

[8] L. Brunette, "Towards a terminology for translation quality assessment: A comparison of TQA practices," The Translator, vol. 6, no. 2, pp. 169-182, 2000.

[9] C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011, pp. 22-64: Association for Computational Linguistics.

[10] M. Carl and M. C. T. Báez, "Machine translation errors and the translation process: a study across different languages," Journal of Specialised Translation, vol. 31, pp. 107-132, 2019.

[11] S. Castilho, S. Doherty, F. Gaspari, and J. Moorkens, "Approaches to human and machine translation quality assessment," in Translation Quality Assessment: Springer, 2018, pp. 9-38.

[12] J. C. Catford, A linguistic theory of translation: An essay in applied linguistics. Oxford University Press, 1965.

[13] S.-w. Chan, A dictionary of translation technology. Chinese University Press, 2004.

[14] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the second international conference on Human Language Technology Research, 2002, pp. 138-145.

[15] J. Drugan, Quality in professional translation: Assessment and improvement. A&C Black, 2013.

[16] M. Felice and L. Specia, "Linguistic features for quality estimation," in Proceedings of the Seventh Workshop on Statistical Machine Translation, 2012, pp. 96-103: Association for Computational Linguistics.

[17] L. Gerber and J. Yang, "Systran MT dictionary development," in Machine Translation: Past, Present and Future. In: Proceedings of Machine Translation Summit VI. October, 1997, pp. 211-218.

[18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," arXiv preprint arXiv:1802.06893, 2018.

[19] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

[20] J. House, Translation quality assessment: A model revisited. Gunter Narr Verlag, 1997.

[21] J. House, "Translation quality assessment: Past and present," in Translation: A multidisciplinary approach: Springer, 2014, pp. 241-264.

[22] J. House, Translation: The Basics. Routledge, 2017.

[23] J. Ive, F. Blain, and L. Specia, "DeepQuest: a framework for neural-based quality estimation," in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3146-3157.

[24] H. Kim, J.-H. Lee, and S.-H. Na, "Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation," in Proceedings of the Second Conference on Machine Translation, 2017, pp. 562-568.

[25] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between european languages," in Proceedings on the Workshop on Statistical Machine Translation, 2006, pp. 102-121.

[26] D. Langlois, "Loria system for the wmt15 quality estimation shared task," in Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 323-329.

[27] S. Lauscher, "Translation quality assessment: Where can theory and practice meet?," The translator, vol. 6, no. 2, pp. 149-168, 2000.

[28] A. Lavie, "Evaluating the output of machine translation systems," AMTA Tutorial, vol. 86, 2010.

[29] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55-60.

[30] Y. Mehdad, M. Negri, and M. Federico, "Match without a referee: evaluating MT adequacy without reference translations," in Proceedings of the Seventh Workshop on Statistical Machine Translation, 2012, pp. 171-180: Association for Computational Linguistics.

[31] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," arXiv preprint arXiv:1712.09405, 2017.

[32] J. Munday, Introducing translation studies: Theories and applications. Routledge, 2016.

[33] P. Newmark, Approaches to translation (Language Teaching methodology senes). Oxford: Pergamum Press. https://doi. Org/10.1017, 1981.

[34] J. Olive, "Global autonomous language exploitation (GALE)," DARPA/IPTO Proposer Information Pamphlet, 2005.

[35] 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 on association for computational linguistics, 2002, pp. 311-318: Association for Computational Linguistics.

[36] A. Pasha et al., "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," in LREC, 2014, vol. 14, pp. 1094-1101.

[37] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825-2830, 2011.

[38] L. Specia, N. Hajlaoui, C. Hallett, and W. Aziz, "Predicting machine translation adequacy," in Machine Translation Summit, 2011, vol. 13, no. 2011, pp. 19-23.

[39] L. Specia, K. Shah, J. G. De Souza, and T. Cohn, "QuEst-A translation quality estimation framework," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2013, pp. 79-84.

[40] L. Specia, G. Paetzold, and C. Scarton, "Multi-level translation quality prediction with quest++," in Proceedings of ACL-IJCNLP 2015 System Demonstrations, 2015, pp. 115-120.

[41] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in Lrec, 2012, vol. 2012, pp. 2214-2218.

[42] S. Vandepitte, "Translation product quality: A conceptual analysis," in Quality aspects in institutional translation: Language Science Press, 2017.

[43] C. Waddington, "Should translations be assessed holistically or through error analysis?," HERMES-Journal of Language and Communication in Business, no. 26, pp. 15-37, 2001.

[44] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.

[45] X. Yinhua, "Equivalence in translation: Features and necessity," International Journal of Humanities and Social Science, vol. 1, no. 10, pp. 169-171, 2011.

[46] Y. Yuan, S. Sharov, and B. Babych, "Mobil: A hybrid feature set for automatic human translation quality assessment," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016: Leeds.