IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2019, accepted September 8, 2019, date of publication September 12, 2019,
date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941161

Arabic–Chinese Neural Machine Translation: Romanized Arabic as Subword Unit for Arabic-sourced Translation

FARES AQLAN[1], XIAOPING FAN[1,2], ABDULLAH ALQWBANI[1], AND AKRAM AL-MANSOUB[3]
[1]School of Computer Science and Engineering, Central South University (CSU), Changsha 410083, China
[2]Academy of Financial and Economic Big Data, Hunan University of Finance and Economics (HUFE), Changsha 410205, China
[3]School of Computer Science and Engineering, South China University of Technology (SCUT), Guangzhou 510006, China

Corresponding author: Xiaoping Fan (xpfan@csu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61876190 and Grant 61802120.

ABSTRACT Morphologically rich and complex languages such as Arabic, pose a major challenge to neural machine translation (NMT) due to the large number of rare words and the inability of NMT to translate them. Unknown word (UNK) symbols are used to represent out-of-vocabulary words because NMT typically operates with a fixed vocabulary size. These rare words can be effectively encoded as sequences of subword units by using algorithms, such as byte pair encoding (BPE), to tackle the UNK problem. However, for languages with highly inflected and morphological variations, such as Arabic, the aforementioned method has its own limitations that make it not effective enough for translation quality. To alleviate the UNK problem and address the inconvenient behavior of BPE when translating the Arabic language, we propose to utilize a romanization system that converts Arabic scripts to subword units. We investigate the effect of our approach on NMT performance under various segmentation scenarios and compare the results with systems trained on original Arabic form. In addition, we integrate Romanized Arabic as an input factor for Arabic-sourced NMT compared with well-known factors, namely, lemma, part-of-speech tags, and morph features. Extensive experiments on Arabic-Chinese translation demonstrate that the proposed approaches can effectively tackle the UNK problem and significantly improve the translation quality for Arabic-sourced translation. Additional experiments in this study focus on developing the NMT system on Chinese-Arabic translation. Before implementing our experiments, we first propose standard criteria for the data filtering of a parallel corpus, which helps in filtering out its noise.

INDEX TERMS Arabic morphology, Arabic romanization, BPE, data filtering, linguistic feature, morphological segmentation, neural machine translation.

I. INTRODUCTION

Neural machine translation (NMT) has obtained impressive results in previous years [1] by outperforming traditional phrased-based statistical machine translation (PBSMT) approaches on various language pairs [2]. State-of-the-art NMT systems rely on an encoder-decoder architecture with an attention mechanism to model a soft word alignment; the model subsequently encodes the source sentence into a fixed-length vector and then decodes word by word to output the target string from vector representations [1].

Arabic and Chinese are two of the most spoken languages in the world, with approximately 319 and 918 million native Arabic and Chinese speakers, respectively[1]. However, the translation between both languages is disproportionally understudied. Thus, building an accurate NMT system, especially for these languages, may gain a high impact on the economic, cultural, and social level. Although research on machine translation (MT) focuses on building algorithms that are independent of languages and can be applied to any language pairs, each language has features that are interesting to investigate.

At present, research on Arabic-Chinese NMT is almost inexistent. To bridge this gap, we propose an approach that basically consists of five components: 1) data filtering,

The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou.

[1]https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

2) morphological segmentation, 3) Arabic romanization, 4) data-driven subword units, and 5) linguistic feature integration. This study describes our approach's details and discusses the evaluation results. To the best of our knowledge, our work is the first in-depth investigation on NMT for this specific domain.

NMT and statistical machine translation (SMT) models are sensitive to the noise of training data [3], especially NMT that quickly memorizes bad examples. Many studies have proposed filtering out the noise in parallel corpora [4]. Data filtering can reduce data size and noise and therefore, can improve training effectiveness in terms of time and quality.

However, finding optimal criteria to filter data is challenging, especially among low-resource languages, such as Arabic and Chinese. The reason is due to the lack of a known high-quality data set, which is required in most well-known filtering and selection approaches, such as modified Moore and Lewis approach [5] and Zipporah [6]. In this work, we propose standard criteria to filter parallel corpora. During the filtering process, we do not rely on the training with a high-quality data set but perform several types of cleaning steps to remove problematic sentences from a parallel corpus. The proposed criteria help remove bad sentences and improve the quality of Arabic–Chinese and Chinese-Arabic PBSMT and NMT models.

NMT typically operates with a fixed vocabulary for input and output sequences by using a limited vocabulary of 30,000 to 50,000 words, but the translation is an open-vocabulary problem [7]. Thus, mistranslated rare words or those denoted with UNK symbols exist. Such symbols are used to represent every possible out-of-vocabulary (OOV) word. In addition to the inability of NMT to translate rare words, using UNK symbols to represent OOV words increases the ambiguity of sentences because UNKs break the structure of those sentences. Thus, the translation and reordering of in-vocabulary words are negatively affected.

This problem becomes further challenging when translating to or from morphologically rich languages (MRLs) such as Arabic, because these languages consist of a large set of morphological features that lead to many rich surface forms. The increase in surface forms produces large vocabularies and high sparseness, adversely affecting the performance of MT, especially that of NMT systems.

Many Arabic morphological segmenters have been proposed in the past two decades. Segmenters split words into morphemes to reduce data sparseness, enhance word alignment, and improve translation quality. However, even with this technique,rare and unknown words still occur in NMT. The basic unit of the input sequence is still represented by a word or its morpheme. Thus, when training on large-scale data, word-level unit compared with subword- and character level units can result in data sparseness among rare words. That is, named entities, numbers, time, and data infrequently occur, thus producing a large vocabulary. These infrequent words are also denoted as UNKs in NMT.

Sennrich *et al.* [7] proposed to segment words into subword units on the basis of data compression by using Byte pair encoding (BPE) algorithm. BPE segmentation shows promising improvement in alleviating limited vocabulary and UNK. It also performs better than back-off translation model. BPE performs better than morphological and character-based segmentation in Arabic–English NMT; nevertheless, the combination of morphological and BPE segmentations further improves the translation [8].

BPE still has certain limitations despite its advantages. For instance, when a root word occurs in various morphological forms, BPE often makes different segmentations, thereby increasing data ambiguity and causing translation errors. Moreover, BPE may split a rare or an unknown word into either not meaningful subword units or semantically different known units, which can output semantically incorrect translations [8]. These cases appear evidently when translating the Arabic language because it has a rich and complex inflectional and cliticization morphology system.

In this study, we propose a new approach to tackle these shortcomings for MRLs while making use of Arabic case studies. This approach adapts a romanization system to transform Arabic texts into Latin-script form. The way this system handles Arabic scripts allows for easy and intuitive handling of morph-phonetic changes and implementing of BPE segmentation algorithm to convert Arabic scripts into subwords, which may alleviate the rare word problem.

We expect that using Romanized Arabic (Latin-script representation) instead of original Arabic scripts can increase the overlap in Arabic vocabulary. Romanized Arabic can also allow the model to linearize Arabic syllables into a sequence of phonemes. Hence, alphabets can be connected with their sound properties. Employing the BPE algorithm on Romanized Arabic can provide flexibility to BPE segmentation in extracting suitable BPE rules, thereby preventing the misleading issue of BPE segmentations and subsequently improving translation quality. Furthermore, BPE on Romanized form can easily recognize inflected words that are prevalent in the Arabic language, thereby producing more frequent words than original Arabic scripts and alleviating the UNK problem. To the best of our knowledge, this work is the first attempt to utilize Romanized Arabic as subword units for Arabic-sourced NMT.

To use Arabic morphological segmentation that splits words into morphemes, we create an approach consisting of four steps (morphological segmentation, romanization, BPE segmentation, and translation), which we call the MRBT approach. We initially conduct morphological analysis to segment Arabic texts. Then, we use a romanization system to convert the output into Romanized form. We further segment the resulting Romanized texts into subword units by using BPE algorithm. Finally, we integrate them into NMT system.

The extensive experiments conducted on Arabic–Chinese NMT demonstrate that: 1) using Romanized Arabic as translation units without any segmentation can reduce vocabulary size and subsequently reduce UNKsin translation;

2) converting the output of morphological segmentation to Romanized Arabic can reduce UNK even further; 3) using BPE to encode the Romanized Arabic can optimize the efficiency of BPE segmentation and thus improve translation quality; 4) applying the MRBT approach leads to substantial improvement on translation quality.

To exploit the power of our approach, we utilize Romanized Arabic as an input feature of Arabic–Chinese factored NMT system to provide additional information on Arabic words. In this manner, we can further improve the translation performance. We compare such a performance with popular linguistic features, namely, lemma, part-of-speech (POS) tags, and morph features. In all our experiments, we represent the baseline model by a standard word-level NMT system conducted without Arabic segmentation.

For the Chinese–Arabic direction, we explore the impact of morphological segmentation and BPE separately. Subsequently, we test the performance of using both methods in tandem. To explore the performance of Chinese–Arabic factored NMT, we use Chinese POS tags as an input factor.

In summary, we contribute the following:

- We build the first optimized NMT models between Arabic and Chinese languages in both directions.
- We propose standard criteria for the data filtering of Arabic–Chinese parallel corpus.
- We compare the impact of different segmentation strategies on Arabic–Chinese and Chinese–Arabic NMT systems.
- We propose a subword transformation solution for Arabic-sourced NMT by using Romanized Arabic.
- We integrate Romanized Arabic as an input factor and use three additional factors to augment NMT systems.
- We provide a qualitative analysis on the translation results.

The rest of the paper is arranged as follows. In Section II, related works are reviewed on Arabic–Chinese MT, data filtering, and UNK problem. In Section III, the NMT framework is described in general. In Section IV, we detail the segmentation schemes and the proposed Romanization-based approach. In Section V, the linguistic input features integrated into our work are discussed. In Section VI, the data filtering steps and experimental settings are summarized, and the results analysis is carried out. And finally, in Section VII, we conclude and provide avenues for future work.

II. RELATED WORK

In this section, we review the related studies on Arabic–Chinese MT, previous research on data filtering, and the UNK problem of NMT.

A. ARABIC–CHINESE MT

Only a few studies on Arabic–Chinese MT exist despite being two of the most-spoken languages in the world. Initial studies, Habash and Hu [9], solved the Arabic–Chinese challenge by pivot technique using English obtains better results than a direct translation in SMT. Ghurab *et al.* [10] collected a corpus from the documentation of the United Nations and built direct PBSMT models between Chinese and Arabic. In [11], Zalmout et al. compared the effects of different tokenization schemes in enhancing the PBSMT performance of Arabic when translating it into several languages, such as Chinese. Recently, Aqlan *et al.* [12] used linguistic input features to improve PBSMT quality between the two languages.

Different from these studies, which involve SMT systems, Junczys-Dowmunt *et al.* [13] compared 30 translation directions, including Arabic–Chinese and Chinese–Arabic PBSMT and NMT models. However, they have trained standard models without conducting any individual preprocessing or linguistic features for Arabic.

B. DATA FILTERING

Many researchers have filtered out the noise in a parallel corpus. For example, Xu and Koehn [6] presented Zipporah as a trainable tool for data selection to select a high-quality data subset from a large amount of noisy corpus. They indicated that Zipporah could improve translation performance by 2.1 BLEU points. However, known high-quality data are required for training when using Zipporah tool. Wolk [14] proposed an approach that relies on the translating source sentences of a parallel corpus by using online MT engines and then compares them with the target sentences. This approach is expensive to perform on a large corpus that may contain tens of millions of sentence pairs. In [15], Khadivi et al. introduced a filtering method on the basis of the models of word alignment. However, the training in this method also requires high-quality data.

To the best of our knowledge, the most similar work to our proposed filtering approach is [16], in which Rikters proposed filtering criteria by performing several types of cleaning steps to remove problematic sentences from parallel corpora. The author reported that in cases where most given parallel corpora are noisy and a small fraction of high-quality corpus exists, cleaning boosts NMT performance. However, in our approach, we perform additional steps related to the nature of the Arabic language and adapt the filtering in accordance with alignment scores between both sides. We evaluate the effect of filtering approach on PBSMT and NMT systems; the filtering steps are presented in our experimental section directly.

C. UNK PROBLEM

Different techniques have been presented to mitigate the UNK word problem. Luong *et al.* [17] handled the OOV problem in a post-processing step by translating each OOV word using a back-off dictionary method. They reported substantial improvement of up to 2.8 BLEU points in experiments on the English–French translation task of WMT14.

In [18], Li et al. proposed an approach that replaces rare words in a testing sentence with similar in-vocabulary words to tackle UNK symbols in a pre-processing step, in which the similarity model is learned from monolingual data. On the

basis of their experimental results on the Chinese–English translation task, a significant improvement of using this method is presented compared with a standard NMT system.

Sennrich *et al.* [7] proposed to segment words via a variant of BPE that can encode open vocabularies with a compact symbol vocabulary of variable-length subword units. In accordance with their experiments on English–German and English–Russian translation tasks of WMT'15, BPE-based models are revealed to outperform back-off models significantly.

To alleviate the UNK problem during the decoding process, an extensive study has been presented by using a large target vocabulary for NMT [19], [20]. The main idea is to select a subset of data from a large-scale target vocabulary to generate a target word in the decoding step. The results of experiments on several language pairs shown that the proposed methods can speed up the translation and alleviate the UNK problem. Luong and Manning [21] proposed a hybrid word-character system to achieve open vocabulary NMT. In their work, the hybrid model is built to translate at the word-level unit, and the character components are consulted for rare words. The experiments on the English–Czech translation task of WMT'15 revealed that the proposed approach performs better than other systems that already handle UNK words.

III. NEURAL MACHINE TRANSLATION

We follow the NMT architecture introduced in [1] as a de facto standard implementation of an attentional encoder-decoder network with recurrent neural networks. The encoder is a bidirectional neural network with gated recurrent units that encodes the source-side sentence $\mathrm{x} = (x_1, \ldots, x_M)$ as one-hot vector sequence and calculates the forward sequence of hidden states $\left(\overrightarrow{h}_1, \ldots, \overrightarrow{h}_M\right)$ and $\left(\overleftarrow{h}_1, \ldots, \overleftarrow{h}_M\right)$ as a backward sequence of hidden states, where $\overrightarrow{h}_m$ and $\overleftarrow{h}_m$ are concatenated to form the final annotation vector h_m as in (1):

$$h_m = \left[\overleftarrow{h}_m, \overrightarrow{h}_m\right] \tag{1}$$

The decoder is also a recurrent neural network that predicts a target sequence $\mathrm{y} = (y_1, \ldots, y_N)$ word by word, where each word y_n is predicted on the decoder hidden state s_n , the previously predicted word y_{n-1}, and source-side context vector c_n , as in (2):

$$p(y|x) = \prod_{n=1}^{N} p(y_n|x, y_{<n}) \tag{2}$$

where

$$p(y_n|x, y_{<n}) = g(y_{n-1}, s_n, c_n) \tag{3}$$

Here, g is a non-linear activation function, which can be defined as LSTM or GRU, to output the probability of y_n, meanwhile, context vector c_n is computed as a weighted sum of annotations h_m . The weight of each annotation is computed by the attention mechanism that models the alignment between y_n and x_m as in (4):

$$\alpha_{n,m} = \frac{exp\left(e_{n,m}\right)}{\sum_{k=1}^{M} exp\left(e_{n,k}\right)} \tag{4}$$

where

$$e_{n,m} = a\left(s_{n-1}, h_m\right) \tag{5}$$

$e_{n,m}$ is the alignment model that shows the probability that input word m is aligned to output word n. Therefore, this mechanism makes the decoder focus on related inputs. The alignment model is a feedforward neural network that is jointly trained by backpropagation.

The factored NMT was initially proposed by Sennrich and Haddow [22], where the encoder is represented by a combination of features. In particular, the encoder receives the linguistic features (lemma, POS, morph, and dependency labels) of source-side sentences in addition to source-side words.

The encoder converts each feature into its embedding vector, and then generates hidden states $\overrightarrow{h_m}$ from the embedding vectors of source-side words X and the linguistic feature's embedding vectors as follows:

$$\vec{h}_m = g\left(\vec{W}\left(\Big\|_{k=1}^{|F|} E_k x_{m,k}\right) + \vec{U}\vec{h}_{m-1}\right) \tag{6}$$

where $\|$ is the vector concatenation, $E_k \in \mathbb{R}^{m_k \times K_k}$ are the matrices of feature embedding, with $\sum_{K=1}^{|F|} m_k = m$, and is the vocabulary size of the k_{th} feature, and $|F|$ is the number of features included in the feature set F [22]. In other words, we roughly split the embedding vectors between the word feature and the linguistic features, which are then concatenated to bring the total embedding size, and the other parts of the model remain unchanged. A detailed description can be found in Sennrich and Haddow [22].

We can use any form of knowledge as a feature to enrich the NMT system with additional information. In [22], Sennrich et al. use lemmas, POS tags, morph features, and dependency labels. To augment Arabic–Chinese NMT in our work, we use lemmas, POS tags, morph features, and Romanized Arabic as a new proposed feature. As for Chinese–Arabic factored NMT; we integrate the input with POS tags (see Section V).

IV. SEGMENTATIONS AND SUBWORD APPROACHES

Morphological segmentation is crucial for MT from and to Arabic and is typically applied as a pre-processing step by using language-specific analyzing tools. An alternative option to morphological segmentation is to integrate language-agnostic subword units into the training algorithm. Next, we describe both options and our proposed approaches for the pre-processing step and subword elements in Arabic-sourced NMT.

A. MORPHOLOGICAL SEGMENTATION

Given that Arabic is an MRL, a multitude of word forms created from the same root fragments the data and bring sparse data problems. Several studies have discussed Arabic statistical segmentation to split words into their individual morphemes, the basic semantic/syntactic units. This subtask plays an important role in several NLP applications. For instance, MT is quite robust to the type of input representation and requires consistency between both training and test data.

In other words, by segmenting Arabic words, we split the clitics that are written attached to their morphemes. The reason is that those clitics increase ambiguities, making the proper detection of word boundaries difficult to achieve, especially in MT, where each morpheme in source language can be aligned to a specific word in the target language.

Such an alignment is explained by the example illustrated in FIGURE 1, in which an Arabic sentence consisting of only one word is aligned to a five-word English sentence (for ease of readability, Arabic examples are transliterated using the Habash-Soudi-Buckwalter scheme [23]). Morphological segmentation (MORPH SEG) process handles this issue by splitting various clitics with several verbosity levels, thus reducing data sparsity, perplexity, and OOV words. Furthermore, MORPH SEG normalizes the inconsistently typed characters in accordance with the default setting of segmentation tools.

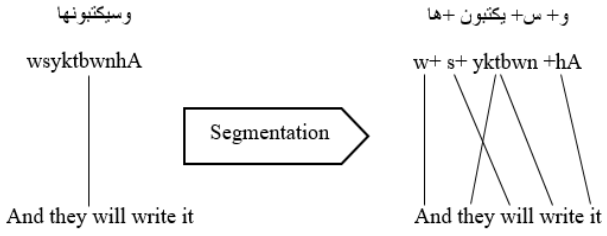

FIGURE 1. Example of Arabic segmentation impact on word alignment.

In this work, we explore the use of two state-of-the-art segmenters that gain above 98% of segmentation accuracy, namely, MADAMIRA [24] and Farasa [25]. MADAMIRA morphological analysis produces a list of possible analyses (independent of context) for each token. Original texts and analyses are added to a feature modeling component to produce predictions for the token's morphological features on the basis of SVM and language models. On the other hand, Farasa is a fast SVM-based segmenter that ignores contexts and uses various features and lexicons to segment words.

B. BYTE PAIR ENCODING

A common standard approach for word segmentation in NMT research and production is BPE [7]. BPE is a data compression algorithm that segments words into small units by merging character sequences on the basis of the frequency of character pairs. Hence, during BPE segmentation, every word is treated as a sequence of characters. The most commonly appearing character pair is then combined into one consisting of those combined characters. The algorithm stops when a predefined number of operations is reached or when no token pair occurs more than once.

This technique has been proven to be an efficient solution for rare words and UNK problems by splitting rare words and leaving frequent words unsegmented (UNSEG). In this manner, the requirements of NMT segmentation is met because NMT typically employs a fixed size of vocabulary.

In this study, we train subword-level NMT systems by using BPE as a pre-processing step to overcome data sparsity issue and further eliminate the UNK rate. We explore the effect of BPE algorithm on UNSEG data. Moreover, we test the use of morphological structure in subword segmentation by performing MORPH SEG prior to BPE method.

However, BPE segmentation faces difficulties in disambiguating word forms in Arabic, given that the language exhibits a complicatedly productive but consistent morphological system with a variety of affixes, clitics, spelling ambiguities, and the root-and-pattern morphology of Semitic languages. This limitation of BPE may cause misleading and incorrect translations for rare words, negatively affecting translation quality.

To increase BPE efficiency, prevent the misleading issue during segmentation, and alleviate the UNK problem, we propose an approach on the basis of Arabic romanization. This approach is detailed hereafter.

C. PROPOSED APPROACH: ROMANIZATION-BASED

Transliteration is the conversion from a given source language text into another language in accordance with the approximate phonetic or spelling equivalents. This well-defined technique has been utilized in many areas, such as MT and information retrieval (IR). For instance, using transliteration to address OOV words in MT systems [26].

Transliteration is also considered a romanization method. In this work, we define romanization as the process of mapping characters into Latin script (ASCII) in accordance with the approximate pronunciation of the original text. Meanwhile, the transliteration system strictly using Latin ASCII characters represents source language orthography one-to-one.

Nowadays, Romanized Arabic is widely adopted and used almost everywhere, such as in messaging, forums, social posts, product and movie ads, mobiles, and TV, because of its ease of reading and typing [27]. However, to the best of our knowledge, no prior research has utilized Romanized Arabic as subword units in the NMT system. Thus, our work is the first attempt in this direction.

NMT operates at the word level with a fixed vocabulary size. As such, several inflected forms of a presented root are treated as UNKs, although certain variations of the given root are occurring in the corpus. This challenge negatively affects NMT performance, especially when translating the Arabic language, which is written from right to left using an abjad writing system that consists of 28 letters and eight diacritical marks with highly inflected and morphological variations.

As we discussed earlier, BPE is a segmentation technique that merges the most frequent sequences of characters; it often separates root words from their affixes. Thus, the vocabulary of NMT only has morphemes, leaving space for infrequent words that are not included in the vocabulary. However, BPE segmentation is only effective on prefixes and suffixes without considering the morphology, particularly on languages with a high degree of inflection, such as Arabic. Hence, BPE has infixes that are not properly segmented, thereby leading to incorrect translations.

To address this inconsistent behavior of BPE, we use Romanized Arabic as subword units instead of the original Arabic form. We expect that this approach can increase the vocabulary overlap, which results in many common subwords and provides good flexibility to BPE segmentation when extracting BPE rules. These rules can prevent the misleading issue of BPE segmentations and hence improve translation quality. Additional advantages of our approach are the easy readability and the proper segmentation of named entities compared with the case when using original Arabic form.

According to our statistics on the corpus we utilize in this study, in the case of using Romanized Arabic without any segmentation, the word-level NMT model easily recognizes inflected words, leading to an evident reduction in vocabulary size. The reduction indicates a decrease in rare words and UNK.

Furthermore, using BPE algorithm to encode Romanized Arabic can provide a subword transformation solution in Arabic-sourced NMT and help alleviate the limitation of BPE, especially the inconsistent segmentation of inflected words. Hence, translation quality is optimized.

To reduce data sparseness and produce the morphological structure of words, we segment words by conducting a MORPH SEG prior to romanization. Subsequently, we break down the Romanized Arabic into subword units via BPE algorithm. These steps constitute our proposed approach MRBT. FIGURE 2 displays the diagram of this approach.

To summarize, we evaluate the effect of our proposed approach on Arabic-sourced NMT by using subword units under six different configurations. We apply BPE algorithm on each of them separately to determine the best subword unit granularity for Arabic-sourced NMT. The configurations are as follows:

1) **UNSEG**: The NMT system takes the unsegmented Arabic original form as input. Only separating punctuation marks is performed as a preprocessing step for the unsegmented text.
2) **UNSEG-Romanized**: The NMT system takes the Romanized Arabic form of the unsegmented text as input.
3) **MADAMIRA.SEG**: The NMT system takes the morphologically segmented text by MADAMIRA tool as input.
4) **MADAMIRA.SEG-Romanized**: The NMT system takes the Romanized Arabic form of the morphologically segmented text by MADAMIRA tool as input.

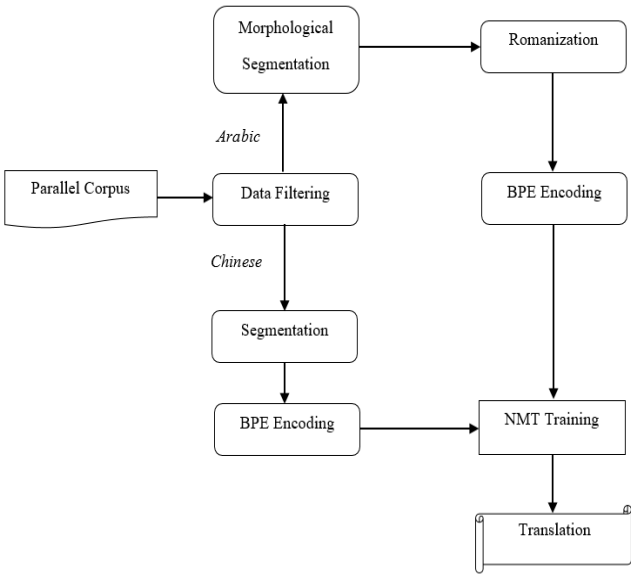

FIGURE 2. Diagram of our proposed MRBT approach.

TABLE 1. Example of preprocessing configurations conducted in this work.

Arabic:
لن يتراجع عن دعمه للسلام بين الأطراف المتصارعة
Chinese:
他不会放弃支持冲突各方之间的和平
English Gloss:
Will not back down from his support for peace between the conflicting parties
UNSEG:
لن يتراجع عن دعمه للسلام بين الأطراف المتصارعة
UNSEG (BPE):
لن يت@@ راجع عن دعمه للسلام بين الأطراف المتصف@@ ارعة
UNSEG-Romanized:
ln ytraj' 'n d'mh llslam byn alatraf almtsar'a
UNSEG-Romanized (BPE):
ln y@@ traj' 'n d'mh llslam byn alatraf almtsa@@ r'a
MADAMIRA.SEG:
لن يتراجع عن دعم +ه ل+ السلام بين الاطراف المتصارعة
MADAMIRA.SEG (BPE):
لن يترا@@ جع عن دعم +ه ل+ السلام بين الاطراف المتص@@ ارعة
MADAMIRA.SEG-Romanized:
ln ytraj' 'n d'm +h l+ alslam byn alatraf almtsar'a
MADAMIRA.SEG-Romanized (BPE):
ln y@@ traj' 'n d'm +h l+ alslam byn alatraf almtsar'a
Farasa.SEG:
لن يتراجع عن دعم +ه ل+ السلام بين الاطراف المتصارعه
Farasa.SEG (BPE):
لن ي@@ تراجع عن دعم +ه ل+ السلام بين الاطراف المتص@@ ار@@ عه
Farasa.SEG-Romanized:
ln ytraj' 'n d'm +h l+ alslam byn alatraf almtsar'h
Farasa.SEG-Romanized (BPE):
ln y@@ traj' 'n d'm +h l+ alslam byn alatraf almtsar'h

5) **Farasa.SEG**: The NMT system takes the morphologically segmented text by Farasa tool as input.
6) **Farasa.SEG-Romanized**: The NMT system takes the Romanized Arabic form of the morphologically segmented text by Farasa tool as input.

TABLE 1 shows an example extracted from our corpus to illustrate the segmentation output under those different configurations performed later on to train the NMT systems.

As presented in this table, BPE can be applied to either Arabic scripts or Romanized Arabic. This advantage helps reduce OOVs in the data. However, the segmentation accuracy differs from one setting to another. Other details are provided as follows:

- **UNSEG Setting:** In the case of unsegmented Arabic scripts, BPE encodes the rare word يتراجع (ytrAjʕ) ''back down'' to يت (yt), which is not a meaningful subword, and راجع (rAjʕ), which means ''review.'' Thus, the segmentation of this infrequent word is incorrect in terms of semantic meaning. The second rare word المتصارعة (AlmtSArʕħ) ''the conflicting'' is encoded to المتص (AlmtS) and رعة (rʕħ). Both subwords are not meaningful and thus may output semantically incorrect translations or be denoted as UNKs.

  By transforming UNSEG Arabic to Romanized form and then applying the BPE method, the first rare word (ytraj') ''back down'' is encoded to (y) subword, which is a suffix added to the verb to indicate its present tense, and (traj') subword, which means ''back down.'' This segmentation is correct in terms of semantic meaning. Hence, using Romanized form instead of Arabic scripts is helpful in disambiguating word forms. However, the segmentation of the second rare word (almtsar'a) ''the conflicting'' in Romanized form remains incorrect, which can be caused by the presence of homophones for the two subword units.
- **MADAMIRA.SEG Setting:** MADAMIRA segmentation first splits all orthographic clitics, except from the article (Al+), and normalizes them on the basis of the default (Alif/Ya) normalization. By applying the BPE method, it encodes rare words (ytrAjʕ) ''back down'' and (AlmtSArʕħ) ''the conflicting'' to totally not meaningful subword units. MADAMIRA reduces data sparsity and improves word alignment by splitting words into morphemes, but MADAMIRA may bring further ambiguities for translation.

  By converting the output of MADAMIRA.SEG to Romanized form prior to BPE algorithm, the first rare word (ytraj') ''back down'' is correctly encoded similar to the scenario of UNSEG setting. BPE defines the second rare word (almtsar'a) ''the conflicting'' here as a frequent common word. Therefore, using Romanized scripts after MORPH SEG helps in the disambiguation of inflected words, resulting in additional common words. This configuration takes advantage of MORPH SEG that splits words into morphemes, providing rich information about word forms. Romanized scripts have a good vocabulary overlap and flexibility for BPE rules.
- **Farasa.SEG Setting:** Farasa morphologically ranks possible segmentations of words and normalizes the context according to the default normalizer. For rare words written in Arabic scripts, BPE correctly encodes (ytrAjʕ) ''back down,'' which can be attributed to the higher quality of Farasa segmentation than that of MADAMIRA. However, BPE still produces an incorrect segmentation of the second rare word (almtsar'a) ''the conflicting.''

  By converting into Romanized form, BPE correctly encodes (ytraj') ''back down'' and defines (almtsar'a) ''the conflicting'' as a frequent common word, which can improve the translation performance, reduce the UNK, and prove the efficiency of our proposed model MRBT by using MORPH SEG prior to romanization step and subsequently applying BPE algorithm.

V. LINGUISTIC INPUT FEATURES

Factored models are often utilized in PBSMT [28] by adding further linguistic information into the source and/or target side, leading to improved translation performance. Several recent studies have proven that linguistic information is beneficial to NMT, where a combination of features represents the encoder input of the factored NMT.

In our work, we propose another feature by using Romanized Arabic as an input factor for Arabic-sourced NMT to explore the effect of linguistic features on the performance of NMT between Arabic and Chinese apart from well-known features, namely, lemma, POS tags, and morph features. Here, we discuss the individual input features in further detail.

A. LEMMA

Lemma captures semantic similarities between all word forms that have the same base form, which allows the sharing of information. Recent studies on Arabic IR systems exhibit the benefits of representing Arabic words at the lemma level in many applications, including SMT [29]. Moreover, using lemma as an input feature in factored Arabic–Chinese PBSMT shows promising results [12].

The need of lemmas can be attributed to the rich derivational and inflectional morphology of Arabic, where the final version of Arabic words is generated by attaching a set of patterns consisting of prefixes, suffixes, infixes, and vowels to the root. For example, the word وسيكتبونها (wsyktbwnhA) ''and they will write it'' has the prefix (ws) ''and will'' and the suffix (wnhA) ''they it,'' both of which are attached to the root (ktb), which has the basic meaning of writing. The lemma of this word is (ktb) ''the concept of writing.''

Different words may share the same base form. For example, words كتاب (ktAb) ''book,'' يكتب (yktb) ''write,'' كاتب (kAtb) ''writer,'' مكتوب (mktwb) ''written,'' مكتب (mktb) ''office,'' and مكتبة (mktbħ) ''library'' share the lemma form كتب (ktb) ''the concept of writing.''

In principle, adding lemma-level representation to the encoder input increases data efficiency, and NMT models can learn that words with inflectional variants are semantically related, leading to precise knowledge. In the continuous vector space, models represent those words as similar points. To extract Arabic lemmas, we use MADAMIRA lemmatization system, which is based on lexeme models and feature ranking. For Chinese–Arabic factored NMT; we do not utilize lemma input factor because the inflection in Chinese is limited, suggesting that using lemmas has no benefit.

B. POS TAGS

POS tagging is important for various NLP tasks, such as information extraction, IR, parsing, and MT. These tags help in disambiguation by computationally determining the activated POS of a word according to its use in the context, which can be a noun, verb, adjective, adverb, and so on.

Arabic is a strongly structured, grammatically ambiguous, and inflectional and derivational language. Furthermore, it is written from right to left in a horizontal form, following a typical order of verb–subject–object (VSO) and a mixed distribution of SVO and VOS. By contrast, Chinese is written from left to right with an SVO word order. Thus, we expect that the annotation of Arabic input to associate each word with its POS tag can help in disambiguating and reducing data sparseness problems.

Previous studies on Arabic–Chinese factored PBSMT exhibit the best translation quality by using POS tags as an additional input feature in the Arabic source side [12]. In our work, we explore the effect of POS tags on the performance Arabic-to-Chinese and Chinese-to-Arabic NMT models. To annotate Arabic corpus, we use MADAMIRA tagger that is considered a state-of-the-art Arabic POS tagging system. For Chinese POS tagging, we utilize Stanford POS tagger via Stanford CoreNLP toolkit [30].

C. MORPH FEATURES

Different word types in Arabic have various types of morphological features. For instance, verbs have person, number, mood, voice, and person. Nouns have case, gender, gloss, state, number, and poclitic. In this work, we use MADAMIRA morphological analysis to generate different types of morph analysis for each Arabic token, including the following.

- Aspect: describes the aspect of a verb and consists of three values: perfective, imperfective, and imperative
- Case: case of Arabic nominal: nominative, accusative, or genitive
- Number: one of the three values to describe the number of a word: singular, dual, or plural
- Gender: two binary values indicate the gender of a word: masculine or feminine
- Mood: with any of the three moods: indicative, subjunctive, or jussive
- State: definite, indefinite, or construct
- Voice: indicates the voice with a binary value: active or passive
- Person: binary feature to indicate if a word is marked for a person or not.
- Politics and Enclitics: clitics attached to a stem

The encoder learns more morph information than the decoder [31]. By incorporating the morph features of each word into the encoder as a feature value for Arabic–Chinese NMT, we hypothesize that the embedding informed by these features can have an advantage on this word, thus optimizing translation quality.

D. PROPOSED FEATURE: ROMANIZED ARABIC

Apart from popular linguistic features (lemma, POS, and morph), we propose a new feature for Arabic-sourced factored NMT models by using Romanized Arabic as an input factor. To illustrate this factored NMT system, an example is provided below.

alahd|الأحد ywm|يوم'qdha|عقدها almqrr|المقرر aljlsa|الجلسة

In this example, the decoder takes the Arabic word and its Romanized form as input, in which the right of the vertical bar ''|'' indicates the Arabic word written in Arabic alphabets, whereas the factor on the left of ''|'' represents the corresponding Romanized Arabic written in Latin script.

We expect that the Romanized Arabic can provide the encoder of Arabic-sourced NMT additional information about the original Arabic word that can further improve translation performance.

VI. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of our proposed approaches, we conduct several groups of experiments on Arabic–Chinese and Chinese–Arabic PBSMT and NMT systems.

A. DATA SELECTION AND FILTERING

Arabic–Chinese belongs to low-resource language pairs; to our knowledge, only two corpora are freely available for this language pair, namely, MultiUN [32] and the United Nations Parallel Corpus v1.0 (UN corpus) [33]. One main advantage of UN corpus is the availability of official validation and test data sets for MT tasks that were created from the documents released in 2015 and excluded from the training corpus.

Unfortunately, training Arabic–Chinese NMT systems with UN corpus as a large data set is not feasible, given our time constraints. Furthermore, the corpus exhibits corrupted parts that negatively affect the quality of the systems and models that learn from the corpus; data selection and filtering on this corpus improve MT performance in terms of training time and translation quality [34]. Hence, further data selection and filtering steps are applied for UN corpus.

Given the scarcity of high-quality data sets between Arabic and Chinese, we propose a standard approach to select and filter bad sentences, reduce training time, and improve the quality of Arabic–Chinese and Chinese–Arabic SMT and NMT models.

We select the first 2M sentences from the UN corpus, then perform the filtering process consisting of the following steps.

- data ambiguity filtering (remove the diacritic marks of Arabic language)
- bad encoding filtering (remove sentences consisting of corrupt symbols)
- unique parallel sentence pairs (remove duplicate sentence pairs)
- filter repeating sentence pairs, where the same sentence of the source side is aligned with various sentences of the target side, or various sentences of the source

side are aligned with the same sentence of the target side
- incorrect language filtering (remove sentences that are neither Arabic nor Chinese according to the results of language identification tool [35])
- bad alignment filtering (remove bad sentence pairs according to the alignment scores produced by the fast-align toolkit[2])

The filtering approach removes 109,484 bad and noise sentence pairs. To evaluate the efficiency of this approach, we train standard PBSMT and NMT systems on the filtered and unfiltered corpora. To make a fair evaluation, focus on the effects of filtering approach, and further speed up the training by allowing a large batch size, we perform standard cleaning steps on the filtered and unfiltered corpora before training. We use Moses [36] scripts to normalize punctuations, eliminate non-printing characters, and drop sentence pairs shorter than 3 or longer than 70 tokens. In all experiments, we use the official validation and test data sets of UN corpus, each consisting of 4,000 sentences. TABLE 2 summarizes the statistics of unfiltered parallel corpus before and after cleaning, and the filtered-cleaned corpus that we utilize in the rest of our experiments. For Chinese, size refers to the segmented words.

TABLE 2. Statistics of the Arabic–Chinese parallel corpus before and after filtering.

Corpus	Sentences	Arabic words	Chinese words
Unfiltered	2,000,000	48,790,761	48,275,437
Unfiltered-cleaned	1,835,654	44,078,203	43,507,315
Filtered-cleaned	1,726,170	42,592,208	41,884,593
Validation	4,000	108,131	104,728
Test	4,000	108,110	104,694

B. EXPERIMENTAL SETTINGS

1) SEGMENTATION

For word-level systems, UNSEG Arabic text is tokenized via Moses default tokenizer that separates punctuation marks. This system represents the baseline of our experiments. In the case of MORPH SEG, we implement the ATB segmentation scheme of MADAMIRA and Farasa tools individually. ATB separates all clitics, except for the definite article, and normalizes the text according to the tool's default settings.

When using MORPH SEG on Arabic as a target language in Chinese–Arabic NMT, we detokenize the outputs via MADA detokenizer before evaluation.

In the experiments with subword-level systems, we use BPE with 40K of merge operations to separately learn and apply segmentation on the source and target training data. Segmentation into subword units is applied after any other preprocessing step. In all scenarios, Chinese texts are segmented by initially applying Jieba[3] segmenter.

2) ROMANIZATION

We utilize Uroman Romanization system [37] to convert Arabic script into Romanized Arabic. This tool romanizes Arabic and myriads of languages on the basis of the pronunciation of the original text. Romanized Arabic is used instead of Arabic script in several experiments of Arabic-sourced NMT, in which the Arabic texts of the training, validation, and test data sets are romanized.

To understand the strength of this proposed model, we evaluate it on word- and subword-level models. For word-level, we separately romanized the UNSEG data and the data segmented by MORPH SEG, then train NMT systems on both scenarios. For the subword level, we test BPE on the Romanized form of UNSEG data then implement MRBT approach presented in this work by employing BPE on the Romanized Arabic of the data already segmented by MORPH SEG. The results of these models are compared with those of models trained on original Arabic form.

We also examine the transliteration method by using Buckwalter transliteration scheme[4] that converts Arabic to Latin ASCII characters. This system strictly represents one-to-one Arabic orthography, and it has been employed in many NLP studies and Linguistic Data Consortium resources. However, we do not see a significant improvement; thus, we hypothesize that the problematic symbols and marks used in Buckwalter scheme may be the cause. This scheme leads to additional noise in Arabic data. Other transliteration schemes, such as Habash–Soudi–Buckwalter, can work well for Arabic-sourced NMT. We suggest future studies to explore this topic.

3) LINGUISTIC FEATURES

To further improve translation quality, we experiment with factored NMT by adding linguistic factors to the input of Arabic-to-Chinese and Chinese-to-Arabic translations. The factors used to enrich the Arabic side are lemma, POS, and morph features, which are obtained using MADAMIRA. Furthermore, we use our proposed feature (Romanized Arabic) as an input factor that adds rich information to Arabic words in their original form. For Chinese input features, we add POS tags obtained by Stanford CoreNLP tagger. The feature representation in our experiments is obtained after BPE segmentation.

4) PHRASE-BASED SMT SETTINGS

We build a standard PBSMT system using Moses MT toolkit, and GIZA++ [38] to extract word alignment. The grow-diag-final-and strategy is used for alignment symmetrization, whereas the msd-bidirectional-fe configuration is employed for lexical reordering. KenLM [39] is used to train a 5-gram language model on the target side of the training data, and Minimum Error Rate Training (MERT) is applied to perform the tuning.

[2]https://github.com/clab/fast_align
[3]https://github.com/fxsjy/jieba‘
[4]http://www.qamus.org/transliteration.htm

TABLE 3. Vocabulary sizes of Arabic-sourced in world-level systems.

Word-level system	Vocabulary size	Ratio (%)
UNSEG	327782	-
UNSEG-Romanized	312152	95.23
MADAMIRA.SEG	190964	58.26
MADAMIRA.SEG-Romanized	183467	55.97
Farasa.SEG	181138	55.26
Farasa.SEG-Romanized	176674	53.90

5) NEURAL MT SETTINGS

We conduct NMT experiments by using Sennrich *et al.* [40] toolkit that is based on TensorFlow framework implemented by attention-based encoder-decoder architecture and GRUs. We set the word embedding dimension of size 500, hidden layers of size 1024, mini batches of size 80, and a maximum sentence length of 50. The vocabulary size for input and output is set to 40K, decoding is performed with a beam size of 12, and the default hidden and embedding dropout are applied. Models are trained using Adam optimizer, thus reshuffling the training corpus among epochs. We validate the model every 10,000 mini batches via BLEU on the validation set and save the model every 30,000 iterations. Early stopping is based on BLEU score with a patience set to 10.

For factored NMT, we keep the total size of the embedding layer fixed to 500 as in [22] to ensure that the improvements on translation performance are not simply due to an increase in the number of model parameters. We set the embedding sizes of the lemma, POS, morph features, and Romanized Arabic to 165, 10, 10, and 250, respectively. Each factored system uses a single feature, in which the word embedding size is set to bring the total layer size to 500.

The models are trained for approximately one month by using 8 Tesla P100 GPUs. An ensemble of the four last saved models is used to report the results because the ensemble technique helps smooth the variance among single models.

We compute BLEU scores via *multi-bleu.perl* script included with Moses, and calculate statistical significance using a bootstrap resampling significance test [41] to determine whether the differences in BLEU scores between unfiltered and filtered systems are statistically significant ($p < 0.05$). The same technique is used to compare the results of other systems trained under different configurations versus the baseline system (UNSEG). All results are reported on the official validation and test sets.

C. STATISTICS OF VOCABULARY SIZE

TABLE 3 illustrates the vocabulary sizes of the training data in different Arabic-sourced word-level NMT systems.

In this table, ratio (%) refers to the percentage of the vocabulary size in a word-level NMT system over that of the baseline system. We observe a substantial reduction in vocabulary sizes by using MORPH SEG of MADAMIRA and Farasa. Transforming into Romanized form leads to a further reduction, which indicates the decrease of rare words and therefore improves performance.

TABLE 4. Results of bleu on Arabic–Chinese PBSMT and NMT. "∗" indicates that the translation performance is significantly better.

	System	Validation	Test
PB-SMT	UNSEG-unfiltered	19.01	19.37
	UNSEG-filtered	**19.40***	**19.63***
NMT	UNSEG-unfiltered	23.02	24.05
	UNSEG	23.22	24.29
	UNSEG (BPE)	**23.91***	**24.90***
	UNSEG-Romanized	23.16	24.13
	UNSEG-Romanized (BPE)	23.84	24.76
	MADAMIRA.SEG	23.81	24.37
	MADAMIRA.SEG (BPE)	23.96	25.05
	MADAMIRA.SEG-Romanized	23.83	24.42
	MADAMIRA.SEG-Romanized (BPE)	**24.12***	**25.16***
	Farasa.SEG	23.87	24.61
	Farasa.SEG (BPE)	24.20	25.12
	Farasa.SEG-Romanized	23.86	24.74
	Farasa.SEG-Romanized (BPE)	**24.42***	**25.43***
	Factor. Lemma	24.46	25.54
	Factor. POS	23.96	25.12
	Factor. Morph	23.82	25.05
	Factor. Romanized Arabic	**24.66***	**25.85***

TABLE 5. Results of BLEU on chinese → Arabic PBSMT and NMT.

	System	Validation	Test
PB-SMT	UNSEG-unfiltered	13.28	12.98
	UNSEG-filtered	**18.18***	**18.52***
NMT	UNSEG-unfiltered	19.39	20.82
	UNSEG	22.75	23.89
	UNSEG (BPE)	**23.47***	**24.33***
	MADAMIRA.SEG	23.52	24.44
	MADAMIRA.SEG (BPE)	**24.20***	**25.17***
	Factor. POS	**24.41***	**25.18***

D. RESULTS AND ANALYSIS

TABLE 4 and TABLE 5 show the results for Arabic → Chinese and Chinese → Arabic translations, respectively.

From these tables, we observe the following:

1) EFFICIENCY OF THE FILTERING APPROACH

The PBSMT and NMT systems trained on unfiltered data are worse than the baseline systems of Arabic → Chinese and notably worse on Chinese → Arabic. The proposed filtering approach significantly improves PBSMT by 0.39 (19.01 → 19.40) and 0.26 (19.37 → 19.63) BLEU points on the validation and test sets for Arabic → Chinese, respectively. A significant and great improvement of 4.9 (13.28 → 18.18) and 5.54 (12.98 → 18.52) BLEU points on the validation and test sets for Chinese → Arabic are respectively observed compared with PBSMT system trained on unfiltered data.

The effect of the filtering process on NMT system is apparent. Arabic → Chinese NMT is improved by 0.20 (23.02 → 23.22) and 0.24 (24.05 → 24.29) BLEU points on the validation and test sets, respectively. A remarkable improvement is observed in Chinese → Arabic NMT, in which the result is improved by 3.36 (19.39 → 22.75)

and 3.07 (20.82 → 23.89) BLEU points on the validation and test sets compared with NMT trained on unfiltered data, respectively. Moreover, the unfiltered corpus costs additional training iterations.

These results indicate that noise and ambiguity exist in the unfiltered parallel corpus. The filtering process we propose helps remove bad sentences and enhance translation time and quality, especially when translating into MRLs as Arabic.

2) PERFORMANCE OF WORD-LEVEL NMT

MORPH SEG schemes (MADAMIRA.SEG, Farasa.SEG) help improve translation quality by handling tokens that do not exist in the training corpus. The results confirm that using a proper segmentation for Arabic greatly benefits NMT. The effects become noticeable when translating into Arabic. Similar results are observed in [42] for Arabic → English and English → Arabic NMT systems.

MADAMIRA.SEG respectively improves Arabic → Chinese NMT performance by 0.59 (23.22 → 23.81) and 0.08 (24.29 → 24.37) BLEU points on the validation and test sets compared with UNSEG NMT as a baseline system. Farasa.SEG improves the result by 0.65 (23.22 → 23.87) and 0.32 (24.29 → 24.61) BLEU points on the validation and test sets, respectively.

Farasa.SEG outperforms MADAMIRA.SEG due to the rich output of MADAMIRA (tokenization, lemma, POS, gloss, and almost all inflected features), whereas Farasa focuses on specified outputs and produces high-quality segmentation performance.

The significant improvement is gained when using MADAMIRA.SEG for Chinese → Arabic NMT. The result improves by 0.77 (22.75 → 23.52) and 0.55 (23.89 → 24.44) BLEU points on the validation and test sets compared with baseline, respectively. We do not test Farasa.SEG on Chinese → Arabic NMT, leaving it for future work.

By using Romanized Arabic instead of its original form on Arabic → Chinese NMT, UNSEG-Romanized performs comparably with baseline system trained on Arabic scripts. MADAMIRA.SEG-Romanized gains better results than MADAMIRA.SEG, whereas findings between Farasa.SEG-Romanized and Farasa.SEG are comparable with each other.

However, vocabulary sizes in Romanized systems are less than those in Arabic script. These results motivate us to apply BPE segmentation to both scenarios to split words into subword units to further improve translation quality.

3) PERFORMANCE OF SUBWORD-LEVEL NMT

Using BPE algorithm as an alternative to MORPH SEG in Arabic → Chinese NMT leads to better translation quality, whereas using both methods in tandem result in further improvement. In contrast, MORPH SEG performs better than BPE in Chinese → Arabic NMT. These results are consistent with previous work on Arabic → English and English → Arabic NMT [8]. Using both methods to train MADAMIRA.SEG (BPE) system provides further improvement. We refer this improvement to the necessity of MORPH SEG when translating into Arabic. In MORPH SEG, words are divided into morphemes, thus reducing data sparsity.

The best results of Arabic → Chinese subword-level systems are achieved by utilizing our proposed approach MRBT, in which we initially transform the output of MORPH SEG to Romanized Arabic and then apply BPE encoding on the resulting text.

Farasa.SEG-Romanized (BPE) approach significantly improves NMT performance by 0.51 (23.91 → 24.42) and 0.53 (24.90 → 25.43) BLEU points on the validation and test set compared with UNSEG (BPE) system trained on original Arabic scripts, respectively. And by 0.22 (24.20 → 24.42) on the validation set and 0.31 (25.12 → 25.43) on the test set compared with Farasa.SEG (BPE). Similar significant improvements are gained on MADAMIRA.SEG-Romanized (BPE) compared with their counterparts trained on original Arabic scripts. Comparable performances are found between UNSEG-Romanized (BPE) and UNSEG (BPE) models, with an advantage to UNSEG (BPE).

Again, MRPT approach takes advantage of the analysis of MORPH SEG, which splits words into morphemes. The output is then transformed into Romanized form. The approach provides better flexibility to BPE segmentation when extracting BPE rules that prevent misleading issues during segmentation. Another advantage is the proper segmentation of names, entities, dates, and times in the Romanized context.

Moreover, the vocabulary sizes of Romanized-based systems are less than those of Arabic script systems. The results indicate that Romanized-based systems not only decrease vocabulary sizes and rare words but also improve translation performance. The promising results of this approach inspire us to further understand this topic. If we utilize Romanized script as an input factor, then we may further improve translation quality. Thus, we train factored NMT systems by using the most well-known factors and the Romanized Arabic factor.

4) PERFORMANCE OF FACTORED NMT

Linguistic features, namely, lemma and POS greatly improve the quality of Arabic → Chinese NMT system. However, when using the morph features, translation performance decreases in the validation and test data sets compared with other features. We infer that this behavior is caused by the many morph features integrated with each word, leading to confusion by sharing similar misclassified tags. We hypothesize that using a single morph feature can lead to good performance.

Chinese → Arabic NMT using POS factor performs better than word- and subword-level models trained without factors. POS tags help in disambiguating and providing additional information about the role of each word in the data.

Among all systems of Arabic → Chinese NMT, the highest performance is obtained by utilizing Romanized Arabic as an input factor, which provides the best support in the training set. Factored NMT system integrated with Romanized Arabic significantly improves the UNSEG baseline system

TABLE 6. Number of unk symbols in the translations of Arabic–Chinese word-level NMT systems.

System	Validation		Test	
	UNKs	Ratio (%)	UNKs	Ratio (%)
UNSEG	1519	-	1493	-
UNSEG-Romanized	1356	89.27	1303	87.27
MADAMIRA.SEG	1232	81.11	1190	79.71
MADAMIRA.SEG-Romanized	1204	79.26	1175	78.70
Farasa.SEG	1189	78.28	1127	75.49
Farasa.SEG-Romanized	1124	74.00	1032	69.12

by 1.44 (23.22 → 24.66) on the validation set and by 1.56 (24.29 → 25.85) absolute BLEU points on the test set. Therefore, this proposed feature is proven helpful for the Arabic-sourced NMT.

E. EVALUATION OF UNK

Apart from reporting the BLEU scores, we also evaluate the influence of using Romanized Arabic in alleviating data sparsity by measuring the number of UNKs in the translations of word-level NMT systems. TABLE 6 summarizes our findings.

"Ratio (%)" in TABLE 6 refers to the reduction rate in the number of UNK symbols in Arabic → Chinese word-level systems over the baseline system. We include the systems trained on original Arabic form and those trained on the transformed Romanized data. From this table, we observe the following:

- Baseline generates many UNKs in the translation due to the sparsity issue of Arabic words, where clitics are written attached to Arabic words, thus increasing their ambiguity and leading to further difficulties in properly detecting word boundaries.
- The MORPH SEG of MADAMIRA and Farasa help in splitting clitics with several levels of verbosity, which helps reduce data sparsity, perplexity, and UNK symbols.
- By transforming Arabic scripts of the baseline to Romanized form, vocabulary size decreases, resulting in the further reduction of rare words and UNK symbols in translations. Using MORPH SEG then converting the output to Romanized form increases the vocabulary overlap between words, thereby producing many frequent words and reducing UNKs in translations.

F. TRANSLATION OF NAMED ENTITIES

In addition to prefix/suffix derived words and new words (unseen at training time), named entities (NEs) form the majority of OOVs. NEs are words or phrases for real-world objects, such as the name of persons, locations, and organizations. Moreover, NEs sometimes include numeric expressions, such as time, date, monetary values, and percent expressions.

Numerous research studies have been published on the task of named entity recognition (NER), which is a popular technique in information extraction that seeks to identify and classify the NEs into predefined categories. NER has become an important part of many NLP applications, such as MT and IR.

TABLE 7. Examples of named entity translation under different configuration.

Types	Example with (English gloss)
Person	***Arabic:*** أخيم كاسو ***Reference:*** 阿希姆· 卡索 (Achim Kassow) ***Baseline-BPE:*** 祖哈尔· 卡萨索 (Zuhar Kasasow) ***MRPT Approach:*** 阿姆· 卡索 (Am Kassow) ***Factor. Morph:*** 阿卡姆. 卡索 (Akam Kassow) ***Factor. Romanized Arabic:*** 阿希. 卡索 (Achi Kassow)
Organization	***Arabic:*** المنظمة العالمية لرابطات التثقيف السابق للولادة ***Reference:*** 世界产前教育协会组织 (World Organization of Prenatal Education Associations) ***Baseline-BPE:*** 产前教育协会的普遍性组织 (Universal organization of prenatal education associations) ***MRPT Approach:*** 世界产前教育协会组织 (World Association of Prenatal Education Association) ***Factor. Morph:*** 世界产前教育协会 (World Association of Prenatal Education) ***Factor. Romanized Arabic:*** 世界产前教育协会组织 (World Association of Prenatal Education Association)
Date and Time	***Arabic:*** 10 : 00 الساعة , 2015 فبراير / شباط 18 يوم الأربعاء ***Reference:*** 2015 年 2 月 18 日 星期三 上午 10 时 (Wednesday, February 18, 2015 10 AM) ***Baseline-BPE:*** 2000 年 2 月 18 日 星期三 上午 10 : 00 (Wednesday, February 18, 2000 10:00 AM) ***MRPT Approach:*** 2015 年 2 月 18 日星期三上午 10 时 (Wednesday, February 18, 2015 10 AM) ***Factor. Morph:*** 2015 年 2 月 18 日 星期三 上午 10 时 (Wednesday, February 18, 2015 10 AM) ***Factor. Romanized Arabic:*** 2015 年 2 月 18 日 星期三上午 10 : 00 (Wednesday, February 18, 2015 10:00 AM)
Monetary value	***Arabic:*** 277.9 مليون دولار ***Reference:*** 2.779 亿美元 ($277.9 million) ***Baseline-BPE:*** 2.71 亿美元 ($271 million) ***MRPT Approach:*** 2.770 亿美元 ($277 million) ***Factor. Morph:*** 27.9 亿美元 ($2.79 billion) ***Factor. Romanized Arabic:*** 2.79 亿美元 ($279 million)

There are two main approaches for developing NER systems, the rule-based approach that requires a set of grammar rules [43], and the machine learning approach that uses prediction methods, such as the Conditional Random Field, it depends on an annotated data to extract a set of features [44], [45]. The extracted features include gazetteer features (predefined NEs) and morph features that have produced by the morphological analyzer.

However, Arabic NER is a challenging task, due to the lack of capitalization in Arabic orthography to mark NEs, the highly ambiguous nature of Arabic NEs, and the lack of resources (that are freely available) [43].

On the other hand, in order to improve the NEs translation, some previous studies have used the transliteration method as a preprocessing step and gained impressive results, such

as Jiang *et al.* [46] who proposed Maximum Entropy model for NEs transliteration from English to Chinese.

In this section, we investigate our system's performance on the translation of NEs, given that our approach is based on three techniques that have been proven useful for this issue, namely Romanization (transliteration is considered a romanization method), the morphological analysis of the Arabic language by MADAMIRA and Farasa tools, and the integration of additional morphological features.

We manually selected different types of NEs from our evaluation data, and investigate the translation results under different configurations. TABLE 7 presents the findings.

TABLE 7 shows four examples produced from the baseline system (UNSEG) with BPE, and from some of the proposed approaches in this work. The first example shows that the baseline system provides poor translation quality for NEs; translating the person name to a different name. A similar result is found in the third example, in which the year 2015 is translated to 2000.

However, we can see that the proposed systems have the capability of making NEs translations more fluent. For instance, the translation results of (MRPT) approach and (Factor. Romanized Arabic) in the second and third examples exactly match the references.

These examples show the capability of Romanized-based systems to make the translation more adequate compared with the baseline system and factored NMT annotated with morph features. We expect to gain better results by integrating predefined NEs; we leave the investigation of this task for future work.

VII. CONCLUSION

This paper presents the first enhanced work on NMT between Arabic and Chinese in both directions. We propose a new approach as a subword transformation solution for Arabic-sourced NMT, that is, we use morphological segmentation schemes to segment Arabic words then employ a romanization system to convert the output into subword units. BPE algorithm is applied on Romanized Arabic to enhance the efficiency of BPE segmentation, reduce the number of rare words, and alleviate the problem of UNKs. Furthermore, we create four factored NMT for Arabic–Chinese; one integrates Romanized Arabic as a proposed input factor, and the others use lemma, POS, and morph features, respectively. From the experiments on Arabic–Chinese, we observe the following: 1) At the word level, the Romanized-based models reduce vocabulary sizes and UNK rates and achieve better or comparable translation results compared with their counterparts in Arabic under various segmentation scenarios. 2) An advantage of romanization at the subword level is that Latin encoding provides great flexibility in extracting proper BPE rules during segmentation, further reducing rare words and improving translation quality. 3) Integrating Romanized Arabic as an input factor can provide extra information that disambiguates input words, leading to the best translation quality among baseline and all other systems. For Chinese–Arabic NMT, we explore the effect of using BPE as an alternative to morphological segmentation and examine both methods in tandem to achieve good translation performance. Moreover, we create factored Chinese–Arabic NMT by using Chinese POS tags as an input factor. Before we implement our experiments, we first propose standard criteria for the data filtering of the parallel corpus, which helps in filtering out its noise and optimizing translations in terms of time and quality.

One avenue of future work is to apply the Romanized-based approach of Arabic-sourced NMT when translating into English and other languages written in Latin characters. We expect further improvement by learning the joint BPE encoding to increase the segmentation consistency between Romanized Arabic and those languages.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable and constructive comments.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–15.
[2] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," in *Proc. EMNLP*, Austin, TX, USA, 2016, pp. 257–267.
[3] H. Khayrallah and P. Koehn, "On the impact of various types of noise on neural machine translation," in *Proc. 2nd Workshop Neural Mach. Transl. Gener.*, Melbourne, VIC, Australia, 2018, pp. 74–83.
[4] H. Khayrallah, H. Xu, and P. Koehn, "The JHU parallel corpus filtering systems for WMT 2018," in *Proc. 3rd Conf. Mach. Transl., Shared Task Papers*, Brussels, Belgium, 2018, pp. 896–899.
[5] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Edinburgh, U.K., 2011, pp. 355–362.
[6] H. Xu and P. Koehn, "Zipporah: A fast and scalable data cleaning system for noisy Web-crawled parallel corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2945–2950.
[7] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, Berlin, Germany, 2015, pp. 1715–1725.
[8] H. Sajjad, F. Dalvi, N. Durrani, A. Abdelali, Y. Belinkov, and S. Vogel, "Challenging language-dependent segmentation for arabic: An application to machine translation and part-of-speech tagging," in *Proc. ACL*, Vancouver, BC, Canada, 2017, pp. 1–7.
[9] N. Habash and J. Hu, "Improving Arabic-Chinese statistical machine translation using english as pivot language," in *Proc. 4th Workshop Stat. Mach. Transl.*, 2009, pp. 173–181.
[10] M. Ghurab, Y. Zhuang, J. Wu, and M. Y. Abdullah, "Arabic-Chinese and Chinese-Arabic phrase-based statistical machine translation systems," *Inf. Technol. J.*, vol. 9, no. 4, pp. 666–672, 2010.
[11] N. Zalmout and N. Habash, "Optimizing tokenization choice for machine translation across multiple target languages," *Prague Bull. Math. Linguistics*, vol. 108, no. 1, pp. 257–269, 2017.
[12] F. Aqlan, X. Fan, A. Alqwbani, and A. Al-Mansoub, "Improved Arabic–Chinese machine translation with linguistic input features," *Future Internet*, vol. 11, no. 1, p. 22, 2019.
[13] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "Is neural machine translation ready for deployment? A case study on 30 translation directions," in *Proc. IWSLT*, Seattle, WA, USA, 2016, pp. 1–8.
[14] K. Wołk, "Noisy-parallel and comparable corpora filtering methodology for the extraction of bi-lingual equivalent data at sentence level," 2015, *arXiv:1510.04500*. [Online]. Available: https://arxiv.org/abs/1510.04500
[15] S. Khadivi and H. Ney, "Automatic filtering of bilingual corpora for statistical machine translation," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Berlin, Germany: Springer, 2005, pp. 263–274.

[16] M. Rikters, "Impact of corpora quality on neural machine translation," 2018, *arXiv:1810.08392*. [Online]. Available: https://arxiv.org/abs/1810.08392
[17] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, 2015, pp. 11–19.
[18] X. Li, J. Zhang, and C. Zong, "Towards zero unknown word in neural machine translation," in *Proc. 25th Int. Joint Conf. Artif. Intell.* New York, NY, USA, 2016, pp. 2852–2858.
[19] G. L'Hostis, D. Grangier, and M. Auli, "Vocabulary selection strategies for neural machine translation," 2016, *arXiv:1610.00072*. [Online]. Available: https://arxiv.org/abs/1610.00072
[20] H. Mi, Z. Wang, and A. Ittycheriah, "Vocabulary manipulation for neural machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 1–6.
[21] M.-T. Luong and C. D. Manning, "Achieving open vocabulary neural machine translation with hybrid word-character models," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 1054–1063.
[22] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proc. 1st Conf. Mach. Transl.*, Berlin, Germany, 2016, pp. 83–91.
[23] N. Habash, A. Soudi, and T. Buckwalter, "On Arabic transliteration," in *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*, A. Soudi, A. van den Bosch, and G. Neumann, Eds. Dordrecht, The Netherlands: Springer, 2007, pp. 15–22.
[24] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. LREC*, Reykjavik, Iceland, vol. 14, 2014, pp. 1094–1101.
[25] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics, Demonstrations*, San Diego, CA, USA, 2016, pp. 11–16.
[26] N. Habash, "Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation," in *Proc. 46th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol., Short Papers*, 2008, pp. 57–60.
[27] A. Chalabi and H. Gerges, "Romanized Arabic transliteration," in *Proc. 2nd Workshop Adv. Text Input Methods*, 2012, pp. 89–96.
[28] P. Koehn and H. Hoang, "Factored translation models," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007, pp. 868–876.
[29] A. El Isbihani, S. Khadivi, O. Bender, and H. Ney, "Morpho-syntactic Arabic preprocessing for Arabic-to-English statistical machine translation," in *Proc. Workshop Stat. Mach. Transl.*, 2006, pp. 15–22.
[30] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.
[31] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass, "What do neural machine translation models learn about morphology?" 2017, *arXiv:1704.03471*. [Online]. Available: https://arxiv.org/abs/1704.03471
[32] A. Eisele and Y. Chen, "MultiUN: A multilingual corpus from united nation documents," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 2868–2872.
[33] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations parallel corpus v1. 0," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, Portorož, Slovenia, 2016, pp. 3530–3534.
[34] N. Durrani, F. Dalvi, H. Sajjad, and S. Vogel, "QCRI machine translation systems for IWSLT 16," 2017, *arXiv:1701.03924*. [Online]. Available: https://arxiv.org/abs/1701.03924
[35] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proc. ACL Syst. Demonstration*, 2012, pp. 25–30.
[36] P. Koehn, M. Federico, B. Cowan, H. Hoang, N. Bertoldi, W. Shen, A. Birch, C. Moran, C. Callison-Burch, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting ACL Interact. Poster Demonstration Sessions*, 2007, pp. 177–180.
[37] U. Hermjakob, J. May, and K. Knight, "Out-of-the-box universal romanization tool uroman," in *Proc. ACL, Syst. Demonstrations*, 2018, pp. 13–18.
[38] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
[39] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, Short Papers*, vol. 2, 2013, pp. 690–696.
[40] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. M. Barone, J. Mokry, and M. Nadejde, "Nematus: A toolkit for neural machine translation," 2017, *arXiv:1703.04357*. [Online]. Available: https://arxiv.org/abs/1703.04357
[41] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Barcelona, Spain, 2004, pp. 388–395.
[42] A. Almahairi, K. Cho, N. Habash, and A. Courville, "First result on Arabic neural machine translation," 2016, *arXiv:1606.02680*. [Online]. Available: https://arxiv.org/abs/1606.02680
[43] K. Shaalan, "A survey of arabic named entity recognition and classification," *Comput. Linguistics*, vol. 40, no. 2, pp. 469–510, 2014.
[44] J. T. Zhou, M. Fang, H. Zhang, C. Gong, X. Peng, Z. Cao, and R. S. M. Goh, "Learning with annotation of various degrees," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2794–2804, Sep. 2019.
[45] J. T. Zhou, H. Zhang, D. Jin, X. Peng, Y. Xiao, and Z. Cao, "RoSeq: Robust sequence labeling," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[46] L. Jiang, M. Zhou, L.-F. Chien, and C. Niu, "Named entity translation with Web mining and transliteration," in *Proc. IJCAI*, vol. 7, 2007, pp. 1629–1634.

FARES AQLAN received the B.S. and M.S. degrees in computer science and technology from Central South University, Changsha, China, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree in computer science and technology. His research interests include natural language processing, deep learning, and data mining.

XIAOPING FAN received the B.S. degree in automation from Nanchang University, in 1981, the M.S. degree in traffic information engineering and control from Central South University, in 1984, and the Ph.D. degree in control science and engineering from the South China University of Technology, in 1995. He is currently a Full Professor with the School of Computer Science and Engineering, Central South University, and the Vice President of the Hunan University of Finance and Economics. His main research interests include wireless sensor networks, robotics, data mining, and intelligent transportation systems.

ABDULLAH ALQWBANI received the B.S. degree in electronic information engineering and the M.S. degree in computer science and technology from Central South University, Changsha, China, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree in computer science and technology. His research interests include deep learning, natural language processing, and data mining.

AKRAM AL-MANSOUB received the B.S. degree in mathematics and computer science from Ibb University, Ibb, Yemen, in 2002, and the M.S. degree in computer science and technology from Central South University, Changsha, China, in 2013. He is currently pursuing the Ph.D. degree in computer science and technology with the South China University of Technology, Guangzhou, China. His research interests include virtual machine, deep learning, and data mining.