IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 31, 2021, accepted November 23, 2021, date of publication December 3, 2021,
date of current version December 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132488

Arabic Machine Translation: A Survey With Challenges and Future Directions

JEZIA ZAKRAOUI, MOUTAZ SALEH, SOMAYA AL-MAADEED, (Senior Member, IEEE), AND JIHAD MOHAMED ALJA'AM

Computer Science and Engineering Department, Qatar University, Doha, Qatar

Corresponding author: Jihad Mohamed Alja'am (jaam@qu.edu.qa)

This work was supported in part by the National Priorities Research Program (NPRP) through Qatar National Research Fund (a member of Qatar Foundation) under Grant 10-0205-170346, and in part by Qatar National Library for open access publication.

ABSTRACT In recent years, computer language area has witnessed important evolvement with applications in different domains. Machine Translation MT technology, considered as a subfield, has received important development with different approaches and techniques. Although, many MT systems and tools that support Arabic already exist; however, the quality of the translation is moderate and needs some improvement. In addition, the high demand for effective technologies to process and translate information from/to Arabic motivated the researchers in Arabic Machine Translation (AMT) to propose new approaches and solutions following the mainstream method, notably neural machine translation (NMT). In this paper, we provide a comprehensive review and compare different NMT approaches mainly for Arabic-English (and English-Arabic) machine translation research works. The discussed approaches address different linguistic and technical challenges and problems while demonstrating great success compared to traditional methods. The results of this work can serve the researchers and professional to be up-to-date and provide them with the necessary resources for modelling and improving of the AMT. These resources include corpora, toolkits, techniques and new models. The obtained results outline various findings, critics, and open issues in this area.

INDEX TERMS Arabic machine translation, Google translation, BLEU, neural machine translation.

I. INTRODUCTION

In recent years, machine translation (MT) become very essential in many applications and achieved advances for almost all languages [1]. Consequently, the recent progress of MT has boosted the translation quality significantly [2], [3], and even approached human translation quality for high-resource language pairs like English-Czech [4] or Chinese-English [5]. Likewise, other research subfields in Natural Language Processing (NLP) systems also achieved a great improvement aiming to satisfy the MT growing needs in several domains and applications such as in multilingual chatbots. Therefore, the demand for quick and accurate machine translations is growing accordingly. However, finding appropriate and optimal translation is not an easy task in any language setting [6], [7], [8]. Several machine translation systems already exist, such as Sakhr,[1] Al-Mutarjim^{TM}Al-Arabey 3.0[2] and SYSTRAN,[3] but the quality of the translation with regards to the Arabic language needs to be further improved due to many issues reported in various research works such as linguistic errors [6], [9]–[13]. In addition, there are web-based MT systems such as Babylon,[4] Bing Translator,[5] Google Translate,[6] and Shaheen[7] that can translate a source text from Arabic to English and vice-versa.

Indeed, machine translation has many challenges, which we split them into two main categories, namely linguistic challenges, and technical challenges. A key linguistic challenge is the nature of the Arabic language as highly ambiguous, linguistic complex and varied compared to other languages. Other characteristics of Arabic such as word order freedom, different diacritization schemas, multitude of dialectal variants along social and geographic dimensions, etc. pose significant linguistic challenges to MT [14]. For

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

[1]http://www.sakhr.com/index.php/en/
[2]https://al-mutarjim-al-arabey.software.informer.com/3.0/
[3]https://www.systransoft.com/
[4]https://translation.babylon-software.com/
[5]https://www.bing.com/translator/
[6]https://translate.google.com
[7]https://mt.qcri.org/api

instance, preprocessing the Arabic source by morphological segmentation [15], [16], syntactical reordering [17], hybridization [7], [18] etc. has been shown to improve the performance of AMT [7], [15]–[19].

A main technical challenge related to AMT rises from the lack of available datasets and lexical resources that can be used as standard benchmark to conduct unified experiments. In fact, researchers tend to collect datasets according to their specific domains and tried to resolve the linguistic issues for Arabic, based on custom datasets such as in the domain of news [9], [20] ignoring hereby many other domains. Other technical challenges such as out-of-vocabulary (OOV), very long sentences, out-of-domain test data, etc. make MT more challenging [21]. For instance, BPE [22], character-level BPE variant [23], hybrid techniques [24] are adopted as good solutions.

With regards to the task of evaluating MT quality, which also encounters many difficulties like the lack of benchmark evaluation datasets, a good study [25] reviewed relevant MT evaluation research works. Each studied work [25] constructed custom datasets and applied specific evaluations metrics. Meanwhile, it becomes more important for the AMT and Arabic NLP community to develop open datasets for researchers to benchmark against, learn from, and extends upon, as recently shown in the following works [26]–[28].

1) MOTIVATION OF THE SURVEY

In the previous section, we have introduced the relevant literature in AMT trend and provided an overview interpretation for current mainstream technology in NMT. We highlighted the effectiveness of NMT in boosting MT quality for high-resource languages. That would be of great interest to inspire young researchers to take part for the mainstream task and investigate different new techniques for Arabic too. However, the feasibility of such accurate MT systems requires a combination of different powerful techniques from different NLP research areas to produce accurate results. This survey lists the most important tools and resources that are available for building and testing new Arabic MT systems. Nevertheless, this survey shed light on the unresolved problems in the development of AMT itself. This survey is intended also for researchers and scholars who want to have an up-to-date comprehensive overview of what has been accomplished so far in the field of Arabic MT. It also provides a quick and broad overview of the Arabic language characteristics and translation difficulties.

2) CONTRIBUTION OF THE SURVEY

A number of surveys were conducted by Arwa *et al.* [8], Elsherif and Soomro [29], Ameur *et al.* [30] where the issues of machine translation of Arabic into other languages was discussed in details. All these previous reviews and works concluded that finding a suitable MT system meeting human requirements is very hard [8]. However, none of the mentioned reviews have explored the linguistic and technical challenges of AMT at glance. Especially, the recent advances in AMT have reshaped the way we can handle these challenges efficiently. Nonetheless, the AMT assessments can be adapted accordingly to fit the recent advances in MT evaluation process. Essentially, we also propose a new classification of the research studies developed for Arabic MT focused on the intended linguistic and/or technical main challenge. Our survey is a comprehensive framework, as shown in Figure 1, more practical for the reader and considers leading research papers from reputable peer-reviewed journals and conferences. Papers that include commercial descriptive statistics or discuss the use of MT in other applications like online post-editing services were excluded. The following exploratory questions form the basis for this paper:

FIGURE 1. A framework to review AMT research works.

What are the challenges, linguistic and technical aspects that are addressed by research works in AMT?

What are the recent research studies that have addressed these challenges in AMT and what are the datasets, tools, evaluation efforts, metrics, and results?

What are the research findings, critics and areas of improvement that can be further studied for AMT and its evaluation?

The rest of the paper is organized as follows: Sections 2 depicts our framework. Section 3 describes the approaches of machine translations. Section 4 studies the linguistics and technical challenges with related works to AMT. Section 5 concludes the review with several ideas and findings. Finally, Section 6 mentions critics and some open issues.

II. SURVEY FRAMEWORK

Arabic is one of the most spoken languages in the world, with approximately 423 million Arabic speakers.[8] United Nations adopted Arabic as the sixth official language of the organization including the release of UN parallel corpus.[9]

Arabic is morphological a very rich language that belongs to a distinct language family, namely the Semitic languages [14]. The Arabic morphology and syntax in addition to other linguistic aspects has made the automatic translation from and to Arabic more challenging. Therefore, it is possible for the size of vocabularies to reach a huge number [7] causing a main computational constraint. Although, there has been a significant improvement due to the recent advances in MT approaches such as statistical, neural methods and hybrid models, these linguistic aspects are still

[8]https://en.wikipedia.org/wiki/List_of_Arab_countries_by_population
[9]https://conferences.unite.un.org/UNCorpus/en/DownloadOverview

causing many difficulties. Thus, when translating between Arabic and Western languages using MT, many problems arise that make it difficult for machines to produce accurate outputs [8], [14], [15], [31], [32].

In this context, we propose a framework to review the main research problems AMT and investigate specific challenges addressed in exiting research works developed for AMT. Further, we explore the most important datasets, tools and resources that are available for developing, testing, and evaluating new AMT systems. Moreover, the framework discusses the current state of the arts with analysis and synthesis. It provides insights on future directions and areas of improvements.

The framework is based on a process flow structure, and consists of four tasks as depicted in Figure 1. The blue-colored texts in the framework denote key elements of the corresponding task.

1. Explore main categories of MT approaches (e.g., blue arrow); we focus on NMT which is emerged as a new paradigm and quickly replaced SMT as the mainstream approach to MT. There are many techniques that contributed to the success of NMT such as the breakthrough of deep learning, attention mechanism [3], transformer models [2], GNMT [33], etc. In practice, NMT also becomes the key technology behind many commercial MT systems such as in Google Translate.
2. Discuss both linguistic and technical challenges in AMT (e.g., red arrow); we consider main challenges that have great impact on the adequacy and fluency of AMT. Besides we mention secondary issues which are less investigated and need to be considered more in the future such as computational complexity.
3. Conduct extensive AMT evaluation studies (e.g., yellow arrow); we also summarize the resources and tools that are useful for AMT evaluation.
4. Perform critical analysis to draw findings and set future directions (e.g., green arrow). We report main findings addressed by researchers and summarize observations, open issues and future research directions.

A. MACHINE TRANSLATION (MT)

Machine Translation MT is a computer application that translates texts or speech from one natural language (i.e., the source) to another (i.e., the target). MT generates a sentence, by translating a source sentence, that gives its meaning in the target language [34]. The translation process which deals with semantic, syntactic, morphological, and additional varieties of grammatical complexities of both languages is hard and complex. In fact, the problem is harder in languages where the source and the target languages have a wide array of linguistic dissimilarities. For example, the Arabic language differs from the target language, such as English, at the phonological, orthographical, morphological, syntactical, and lexical levels [14].

First, phonologically, Arabic contains 28 consonants, 3 short vowels, 3 long vowels, 2 diphthongs. Arabic spelling is mostly phonemic corresponding to its letter sound [14]. Second, orthographically, the Arabic script is an alphabet with allographic variants, optional zero-width diacritics, and common ligatures [14]. The shape of a single Arabic letter may change slightly depending on its position within the Arabic word (beginning, middle, or at the end). Third, morphologically, Arabic morphology has two main functions, notably inflection and derivation. Inflectional function, which is mostly concatenative, modifies features of words such as tense, number, person, etc. Derivational function, which is mostly templatic, creates new words by inserting one or multiple affixes [14]. Fourth, Syntactically, Arabic syntax is heavily related to its morphological level. Thus, several syntactic aspects are not expressed uniquely via word order but also through morphology. The Arabic admits two types of sentences: verbal and nominal. Arabic is case marking: nominative, accusative, genitive, and almost-free word order. A detailed study of Arabic language is reviewed by Habash [14].

Moreover, AMT poses the challenge of resolving Word Sense Disambiguation (WSD) problem where words can have more than one meaning. Arabic is classified as the language with higher ambiguous words [14]. Indeed, many research works [14], [30], [31], [35] have demonstrated the linguistic characteristics and challenges of Arabic language. Other works [15], [20] proposed to resolve these linguistic challenges to improve the quality of AMT. Few works such as [19], [25] studied the robustness of these systems by measuring their performance on many different datasets while works like [27] participated to improve the AMT computation.

B. MT EVALUATION

MT can be assessed manually or automatically. While the experts evaluation is very costly and steady, automatic evaluation is considered objective and cheap, however less comprehensive than human evaluation [25]. Indeed, considering the benefits of automatic evaluation, like speed and free cost, most of the researchers use automatic evaluation metrics to optimize the performance of their systems. Notably, most researchers use Bilingual Evaluation Understudy (BLEU) method [36] to measure the quality of an MT output. BLEU measures the closeness of the candidate output of a machine translation system to a reference translation, in standard case a professional human, of the same text to determine the quality of the MT.

There are several studies in the literature presenting enhanced BLEU methods such as [37]. Depending on different aspects, researchers use also different metrics such as Word Error Rate [38], METEOR [38], AL-BLEU [39] metric which extends BLEU to deal with Arabic rich morphology. Following the same perspective, AL-TERp [19] is proposed as a new metric that supports Arabic language. In addition, a study conducted by Sai *et al.* [38] tried to evaluate several automatic metrics. However, there were no conclusions to

admit that one metric outperforms the others among the studied ones in the works [25], [38].

While automatic metrics such as BLEU capture the average case for how well a MT model translates sentences, they do not give insight into which linguistic aspects MT models struggle with producing fluent output. Some research investigated MT samples with native speakers so they could review the linguistic aspects of MT errors [13], [40] other research works used neural networks to detect errors [41] or to correct them [42]. However, as the quality of the MT output improves over the time, MT evaluation becomes fully integrated in many frameworks such as Multidimensional Quality Metric (MQM),[10] Dynamic Quality Evaluation Mode (DQM)[11] and Natural Language Generation (NLG).[12] These frameworks are prepared as efforts to standardize translation quality evaluation and were used to assess the MT systems in terms of holistic adequacy and fluency scale [40].

III. APPROACHES TO MACHINE TRANSLATION

Machine translation has two main categories: rule-based (also known as knowledge-driven) approaches (RBMT) and data-driven approaches which in turn comprises example-based (EBMT), statistical (SMT), and neural machine translation (NMT) [8]. MT systems that use multiple MT approaches, to compensate for the weakness of each approach, are called hybrid MT. Figure 2 visualizes the evolution and the general classification of the MT approaches. Following, we briefly review the development of these approaches.

A. RULE-BASED MACHINE TRANSLATION (RBMT)

Rule-Based machine translation is considered as a traditional approach to MT as it involves a set of linguistic rules to translate text from the source to the target languages [8], [31], [34], [43]. The rules are usually constructed by a language expert [32]. This approach also relies on bilingual or multilingual lexicons including Arabic and other languages.

This approach is divided into direct method, transfer method, and Interlingua (IL) method [8], [32], [43]. The strength of RBMT approach is that it can deeply analyze both syntax and semantic levels as mentioned in [8] and it is still effective for languages with limited parallel data, namely low-resource language pairs [43]. However, it is impossible to write rules that cover all languages, as this requires great linguistic knowledge with a large and good dictionary. A good dictionary is even expensive to build and may not cover all the words. In addition, linguistic experts are needed to build comprehensive rules that cover the morphological, syntactic, and semantic mapping between languages pairs [44], [45].

[10]http://www.qt21.eu/quality-metrics/
[11]Dynamic Quality Framework (TAUS). https://dqf.taus.net/
[12]https://github.com/Maluuba/nlg-eval

B. EXAMPLE-BASED MACHINE TRANSLATION (EBMT)

EBMT is a translation method that retrieves similar examples of source phrases, sentences, or texts and their translations from a database of examples adapting the examples to translate new input languages [8], [31], [34], [43]. In general, EBMT works on four stages namely, example acquisition, example base management, example application, and target sentence synthesis.

The strength of this approach is that it is easily upgraded because it has no rules, thus improvement is achieved simply by adding appropriate examples to the database. However, when the database of examples becomes large, the translation quality does not improve and there might be cases where the performance starts to decrease. In this case, the retrieval from the example database will be slow [46]. Worth mentioning works that also tackled the EBMT method are discussed in [47], [48].

C. STATISTICAL MACHINE TRANSLATION (SMT)

While RBMT involves a set of linguistic rules to translate text from the source to the target language, SMT [8], [31], [34] on the other hand, builds statistical models from a collection of datasets constituting of sentence-aligned parallel corpus so-called multilingual resource. Phrase-based SMT (PBSMT) models give the state-of-the-art performance for most languages [49] and outperform the simple word-to-word translation methods [49]. The key component of a PBSMT model is a phrase-based lexicon, which pairs phrases in the source language with phrases in the target language. The lexicon is built from the training data set that is a bilingual corpus.

PBSMT takes place in three main phases, namely language model phase, translation model phase and decoder model phase. First, the translation model can be trained on the bilingual corpus to estimate the probability of the source sentence being a translated version of the target sentence. Second, the language model can be trained on monolingual corpora and used to improve the fluency of the output translation. Finally, in the decoder phase, the maximum probability of product of both the language model and the translation model is computed which gives the most probable sentence in the target language. Given a source sentence s along with its translation t, the highest probability sentence is chosen as the best translation after applying Bayes theorem using the following equation [50]:

$$\hat{t} = \underbrace{argmax}_{t} P(s \mid t) \times P(t) \tag{1}$$

where $P(t)$ is the language model that ensures the fluency of the generated target output and $P(s \mid t)$ is the translation model for computing that ensures the accuracy of the translation. The decoder provides a dynamic programming solution by applying the beam search algorithm [50].

There are three models in SMT namely, word-based, phrase-based, and syntax-based SMT [1], [43]. Briefly, word-based SMT translates the source language word by word or more in the target language [1]. However, PBSMT translates

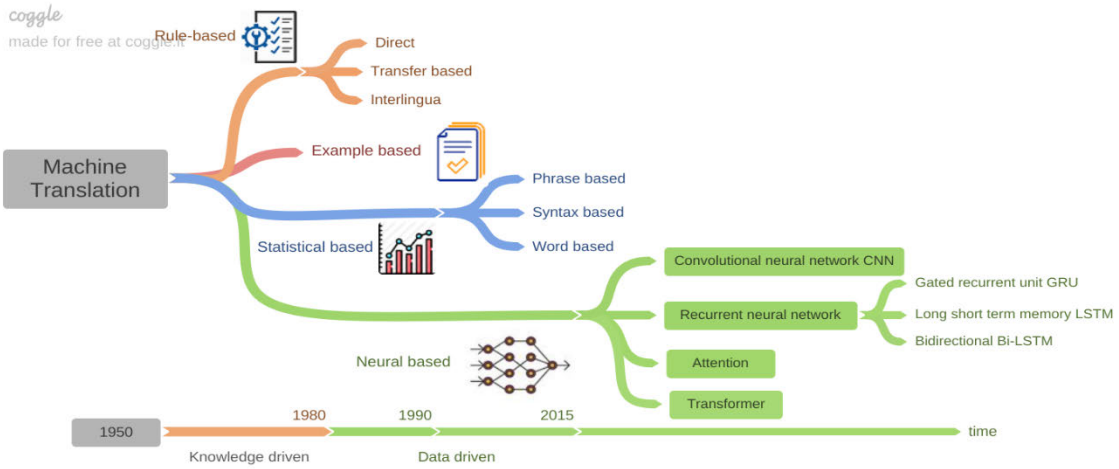

FIGURE 2. Evolution and classification of machine translation approaches.

the source language phrase by just a phrase in the target language [1]. Syntax-based SMT is based on the idea of translating syntactic units, rather than single words or sequence of words [1]. A hierarchical phrase-based model combines the strength of phase-based and syntax-based SMT.

Concerning SMT features, SMT can handle ambiguity by reordering phrase-based translations with their frequency of occurrence on a phrase table [8], [31]. Thus, the translation result generates more fluent and natural than RBMT. In addition, SMT is language independent, easy, cheap, and fast to build. One weakness of SMT is the challenge of translating material that is different from the training corpora as argued [8]. In such cases, the accuracy is poor and to achieve good translation the corpus should be customized for a specific style. Furthermore, by segmenting the source sentence into several phrases and doing phrase replacement, it may ignore the long dependency beyond the length of phrases and thus cause inconsistency in translation results such as incorrect gender agreements [51]. Worth mentioning works were presented as PBSMT systems [15], [52], [53] or as comparative studies [54], [16].

D. NEURAL MACHINE TRANSLATION (NMT)

NMT models based on deep neural network (DNN) have been proposed in early NMT research [51]. A DNN based NMT model employs a neural network system to perform the required machine translation tasks using an encoder-decoder network [55]. The encoder neural network inputs and encodes a source language sentence into a fixed-length vector in each hidden state. Then, given the final hidden state of the encoder, the decoder does the reverse work by transforming the hidden state vector to the target sentence word by word. A translation probability of a source sentence is modelled into the target sentence. Given a source sentences $S = \{s_1, s_2, ..s_n\}$ and a target sentence $T = \{t_1, t_2, ..t_n\}$, the encoder encodes all the words from the source sentence S into a set of hidden states $(h_1, h_2, \ldots h_n)$ and passes the fixed-size vector v, which represents the source sentence, to the decoder. The translation probability with a single neural network is given by following formula [33]:

$$P(T \mid S) = \prod_{i=1}^{n} P(t_i \mid t_{<i}, S) \tag{2}$$

where $t_{<i}$ stands for the sequence preceding the ith target word.

Hence each predicted word t_i is based on the previously predicted word t_{i-1} and the previous hidden states h_{i-1}. However, when the sentences become long the performance deteriorate. This limitation is due to the limited feature representation ability in a fixed-length vector [51]. To overcome this issue and to provide additional word alignment information in translating long sentences, Bahdanau *et al.* [3] introduced the idea of the attention mechanism. Concretely, attention mechanism is an intermediate component between encoder and decoder, which can help to determine the word alignment dynamically. The decoder pays attention to input or to any part of the input sentence. Attention is calculated using each encoder output and the current hidden state, resulting in a vector of the same size as the input sequences using score functions [3]. Figure 3 shows the concept of attention, which can provide additional alignment information rather than just using information in fixed-length vector.

There are mainly three different architectures for constructing NMT.

1) Recurrent neural network (RNN): it has been producing good quality translation result. RNN is composed of encoder and decoder with similar working of sequence-to-sequence learning. Different RNN architecture are experimenting different models i.e., LSTM [56], BiLSTM [3] and GRU [57].

2) Convolution neural network (CNN): it has achieved great results for the word-based MT, but along with RNN [58]. This work applied convolution layer on the bottom of the recurrent layer which hinders the performance.

TABLE 1. NMT-based systems with respective methodologies models and toolkits.

Author name	Year	Methodology	Model	Toolkit
Almahairi *et al.* [35]	2016	Preprocessing, BPE	Bi-GRU encoder, GRU decoder, attention	DL4MT[13]
Durrani *et al.* [27]	2017	Preprocessing, BPE, ensembling	Bidirectional RNN encoder	Nematus[14]
Ameur *et al.* [67]	2017	Preprocessing, NE extraction, transliteration, scoring, selection	GRU/Bi-GRU encoder, attention-based decoder	OpenNMT[15]
Belinkov *et al*. [68]	2017	Attention, POS-tag, Treebank for Arabic, the character-based model is a CNN with a highway network over character	2-layer LSTM encoder-decoder, CNN	seq2seq-attn[16]
Alrajeh [54]	2018	Preprocessing, BPE	RNN encoder-decoder with attention	Marian[17]
Elaraby *et al.* [69]	2018	Additional input constraints, Part-Of-Speech tagging using MADAMIRA	Attention based seq2seq encoder-decoder	NA
Almansor and Al-Ani [70]	2018	Hybrid character model based on encoder and decoder	Encoder-decoder with 4 layers that incorporates both RNN and CNN	NA
Shuo *et al.* [71]	2018	Triangular NMT model, back-translation	Attention-based encoder-decoder with a bidirectional RNN encoder	NA
Oudah *et al.* [16]	2019	Morphology-based and frequency-based tokenization, Sentence length-based system selection, BPE	LSTM with attention	OpenNMT
Shapiro and Duh [72]	2019	Morphological analysis, different embedding types, encoder-decoder model, BPE	Bi-LSTM with attention	OpenNMT
Alkhatib and Shaalan [18]	2020	Hybrid deep learning, preprocessing, multi-feature extraction and selection	CNN followed by Bi-LSTM and CRF	NA
Solyman *et al.* [73]	2020	BPE, fine-tuning	Attention based seq2seq	Fairseq[18], FastText
Saadany and Orasan [74]	2020	Preprocessing, infusing contextual cues at the training	Seq2seq model with LSTM, global attention, 2 Transformers	OpenNMT
Liu *et al.* [75]	2020	Pre-processing, pre-training, back-translation, fine-tuning and decoding	Seq2seq Transformer, with 12 layers of encoder and 12 layers of decoder	Fairseq
Abid [76]	2020	BPE, back-translation, bootstrapping	Transformer-base architecture with 6 encoder and 6 decoder layers	Fairseq
Fan *et al*. [77]	2020	Bridge language mining strategy, combination of dense scaling and language-specific sparse parameters, back-translation, sentencePiece[19]	Seq2seq parallel Transformer, 1.2 Billion parameters	fairscale[20]
Ma *et al*. [78]	2020	Transformer-based multilingual NMT, cross-lingual encoder and decoder initialization, sentencePiece, back-translation	Transformer-big architecture with a 6-layer encoder and decoder	Fairseq
Ji *et al*. [79]	2020	Transfer learning, pivot-based method and multilingual NMT, cross-lingual language model pre-training, zero-shot translation, back-translation, BPE	Transformer-big model	XLM[21]
Zhang *et al.* [80]	2020	Multilingual NMT, pivot-based translation, fin-tuning, BPE	Transformer-base model	NA
Farhan *et al.* [81]	2020	Attention-based NMT model, BPE	Google NMT model, seq2seq model	NA
Ataman *et al.* [57]	2020	Processing morphological inflection, BPE	Bi-LSTM, GRU	OpenNMT
Lin *et al.* [82]	2021	Pre-training, back-translation, BPE, Fine-tuning	Transformer-large architecture with 6-layer encoder and 6-layer decoder,	Fairseq
Stergiadis *et al*. [83]	2021	Back-Translation, fine-tuning parallel data on a mix of synthetic and genuine (1:1 composition), attention, BPE.	Seq2seq transformer-based model,	OpenNMT
Berrichi and Mazroui [84]	2021	Preprocessing, post-processing, semantic and alignment-based segmentation, BPE	RNN encoder-decoder, GRU, attention	NMTpy[22], Nematus

[13] https://github.com/nyu-dl/dl4mt-tutorial
[14] v https://github.com/EdinburghNLP/nematus
[15] https://opennmt.net/
[16] https://github.com/harvardnlp/seq2seq-attn
[17] https://marian-nmt.github.io/
[18] https://github.com/pytorch/fairseq
[19] https://github.com/google/sentencepiece
[20] https://github.com/facebookresearch/fairscale
[21] https://github.com/facebookresearch/XLM
[22] https://github.com/lium-lst/nmtpy

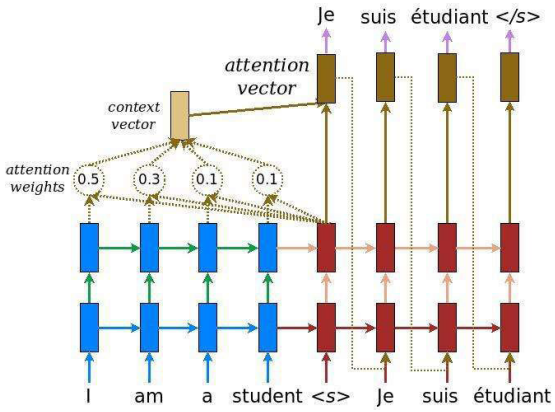

FIGURE 3. The concept of attention mechanism [62].

The bottleneck was handled by implementing the fully convolutional model as suggested by [59]. The performance and accuracy were improved with a number of models; word-based [60], Character-based [23], and recently with attention [61].

3) Self-attention-based Transformer: it is a sequence-to-sequence model [2], which consists of a stack of layers. Each layer first utilizes the self-attention to extract information from the whole sentence, then follows a point-wise feed-forward network to provide non-linearity. The novel idea of self-attention is to extend the mechanism to the processing of input sequences and output sentences as well. In general form, the Transformer attention function uses three vectors: queries(Q), keys (K) and values (V).

The output is a weighted sum of values, where weights are computed by a similarity score between n query vectors and m keys [2]. The attention is defined as follows:

$$Attention(Q, K, V) = softmax(score(Q, K))V \qquad (3)$$

where $score\,(Q, K)$ is an $n \times m$ matrix of similarity scores. A straightforward choice for $score\,(Q, K)$ proposed by Luong *et al.* [62] is the dot product i.e. score(Q, K) = QK^T. The *softmax* function normalizes over the columns of that matrix so that the weights for each query vector sum up to one. There are many variants in the implementation of attention-based models which are classified into two broad categories, global and local attention discussed in detail in this survey [51].

Current state-of-the-art NMT models [63] rely on the Transformer model [2] and multiple attention mechanism [3]. However, the transformer-based language models such as Bidirectional Encoder Representation from Transformers (BERT) [64] expands the function of attention to encompass the main task. It uses self-attention, which is applied to two states within the same sequence, as the foundation for sequence representations rather than an RNN. For Arabic language, two transformer-based language models have been developed so far; notably AraBERT [65] and GigaBERT [66]. Both models aim at solving a masked language-modelling task in order to correctly predict a masked word from its context. Besides, these models aim at resolving a next sentence prediction task especially to decide whether two sentences are consecutive or not.

This work focuses on the recent neural-based AMT approaches. There are different models and toolkits available for developing the NMT system while experimenting with different methodologies as shown in Table 1. The NMT systems, shown below conducted experiments on different corpora and domains and demonstrated reasonable to great results with regard to AMT, as mentioned in the Tables 2 and 3.

IV. CHALLENGES FOR AMT

MT is considered a challenging task as it includes various statistical models, machine learning models and other techniques that aid to develop an MT. Following, we discuss the types of challenges that encounter AMT in two key aspects notably linguistic and technical considerations. Briefly, linguistic challenges are related to the nature of Arabic language such as its morphology richness. Whereas technical challenges are related to the computation process of MT.

A. LINGUISTIC CHALLENGES

Linguistic challenges are specific to each language pair as it is not straightforward to build a general approach applicable to all languages. Especially, the syntactic divergence of Arabic language morphology from English language rise to specific MT issues. Many other morphology-rich languages makes the MT more complex. Languages such as Czech and Russian belonging to Slavic languages,[23] Turkish language, to name but a few, exhibit rich and complex morphology compared to English language [85], but with distinct morphological characteristics: Czech (fusional), Turkish (agglutinative) and Arabic (templatic) [57]. Generally, the below linguistic challenges apply to morphology-rich languages, but with distinct extents.

Specifically, for Arabic language, the work studying the characteristics of Arabic language by Habash [14] has formed a basis for further AMT research through specifying the challenges that should be carefully considered when developing and evaluating any MT system from/to Arabic. We summarize the main linguistic challenges related to AMT as follows:

- *Morphology richness*: Arabic has a rich and complex morphology which is substantially different from English or other western languages. Both pre-processing and post-processing of Arabic source by NLP tools such as segmentation and tokenization has shown to improve the performance of MT by reducing the size of the source vocabulary and improving the quality of word alignments. For instance, a single Arabic token "ولمدرستها" translated to '*and to her school*' in English) is formed by prepending the prefix "و" ('*and*') and "ل" ('to') to the base lexeme "مدرسة" ('*school*), appending the prefix

[23] https://www.britannica.com/topic/Slavic-languages

"ها" ('*her*') and replacing the feminine suffix "ة" of the base lexeme to "ت". This feature of Arabic is challenging and has been addressed by many works [15], [35], [52], [53], [65], [86], [87]. This is often exacerbated by the orthographic ambiguity found in Arabic scripts, such as the inconsistency in spelling certain letters as assumed in [35].

- *Free word order*: Arabic has more freedom in the order of words within a sentence. In fact, Arabic free word order allows the combinations of SVO, VSO, VOS, and OVS within a sentence and thus convey different meanings. Accordingly, the works in [86], [88] addressed the syntactic word reordering.
- *Word sense disambiguation (WSD):* In Arabic it is hard to identify the correct sense of a given word that has multiple meanings (an ambiguous word). Specifically, un-vocalized words have different concepts in different contexts. For instance, the un-vocalized word "ذهب" can be translated as verb such as "go" or as a noun "gold". The works [89], [90], [91] attempt to resolve WSD.
- *Named Entity (NE):* the difficulty in NE is mainly due to the lack of capitalization, the rich lexical variations, and the lack of uniformity in writing Arabic named entities [92]. The proper handling of named entities is crucial to produce a reliable translation result either using meaning-based translation or using phoneme-based transliteration [93], [18].
- *Vocalization:* Arabic can be fully, partially, or un-vocalized at all. While un-vocalized Arabic word cause ambiguities in MT, vocalized Arabic words resolve the ambiguity, however, not optimal for MT training [94]. For instance, depending on its vocalization, the Arabic word "علم" can have these possible translations (flag, science, he knew, it was known, he taught, he was taught). Some works tried to remove diacritics [94] because it reduces the vocabulary size which is often a positive thing for MT [94], while others predicted diacritics [95] and restored diacritics [96] using corpora such as [97].
- *Dialectal variation*: Arabic dialect is pre-dominantly spoken. Despite the increasing use in social media, there is no official standardization. A recent survey of relevant research works is presented by Harrat *et al.* [98] to address MT between Arabic dialects. The authors mentioned that MSA and Arabic dialects use same Arabic script in their writing system, but also, highlighted a considerable number of differences between MSA and Arabic dialects. Arabic dialects vary to different degrees in all levels of the language: phonology, morphology, syntax, and lexicon. On other hand, Arabic speakers mix Arabic varieties in different ways in addition to the use of a non-standard romanization consisting of Latin characters, numeric digits, etc. which contribute to the complexity of dialects cliticization. While MSA possesses a complex case system, Arabic dialects lack case endings. Case endings are diacritical marks attached to the ends of the word to indicate grammatical function [99]. The linguistic differences and commonalties amongst the Arabic dialects and to MSA are studied in detail in [99].

B. TECHNICAL CHALLENGES

Most of the technical challenges are caused by issues related to the nature of the specific language such as having high vocabulary size due to rich morphology. Some other issues are related to computation constraints in sequence models. Hence, Arabic language as well as other morphology-rich languages such as Slavic languages, Turkish language, to name but a few, face almost same technical issues, but with different grades except for Czech language since it is considered as high-resource language [100]. The main technical challenges are encountered while developing, applying, and evaluating MT techniques are following:

- *Low-resource*: Arabic is a low-resource language. MT performs well for high resource languages as compared to low-resources languages since the learning depends on the amount of training data [21]. NMT systems trains millions of data, showing direct proportionality to accuracy [79]. There are many techniques studied for handling low-resource languages; pivot translation [79], triangular NMT [71], back-translation [71], transfer learning [79], fine-tuning [72], multilingual NMT [72], zero-shot translation [33], zero-shot transfer [66].
- *Out-of-domain*: also known as domain mismatch [21]. MT has low quality for out-of-domain text. The translation output misguides by visualizing the fluent output [101], but not the accurate one.
- *Out of vocabulary (OOV) and Vocabulary size*: NMT typically operates with a fixed vocabulary for input and output sequences by using a limited vocabulary of 30K to 50K words [22]. Thus, OOV increase ambiguity of sentences while reordering of in-vocabulary words during MT process. Indeed, NMT systems have a steeper learning curve with respect to the amount of training data, resulting in worse quality in low-resource settings, but better performance in high-resource settings [21].
- *Word alignment*: is the process of aligning input words or phrases to output words or phrases. Arguably, the attention model in NMT has improved the alignment mechanism between source and target words [21]. The performance of attention based NMT is very poor in case of more substantial sentences and it does not provide accurate word alignment but may in fact dramatically diverge [21]. Different techniques have been developed to determine the word alignment dynamically [51].
- *Sentence length*: the translation quality of complex and very long sentences is low as compared to small sentences. Although the introduction of the attention model remedied this problem to some extent, it is still persisting for very long sentences (80 and more tokens) [21].

- *Computational overhead*: there are many issues related to computation overhead such as translation inference runtime. To overcome this limitation, some works focus on computation speed-up, while others focus on using context information such as Copy Mechanism [51] to alleviate unlimited vocabulary size. An important issue is hyperparameter optimization; there are various hyperparameters regarding the NMT architecture and the training regime. An eager attempt to exhaustively search over the entire hyperparameter space is almost always infeasible. Indeed, to find a good hyperparameter configuration is mostly computationally expensive [85].

In next two sections, we thoroughly reviewed the main research papers in each challenge category. Despite there exists many published papers particularly for the morphology problem of Arabic, we could not be exhaustive in citing all works conducted. Nevertheless, we chose to include works that use diverse techniques. Tables 2 and 3 summarize the specific problems related to AMT and the main research efforts to tackle the linguistic and technical challenges, respectively.

C. COMPARISON OF AMT APPROACHES ADDRESSING LINGUISTIC CHALLENGES

We review rule-based RBMT and statistical-based SMT works that address the linguistic challenges marginal, whereas we focus more on neural-based NMT, as shown in Table 2. Some works used hybrid models.

1) COMPLEX MORPHOLOGY

Recently, there are few NMT studies addressing this issue compared to earlier AMT research works. Worth mentioning work is presented by Almahairi *et al.* [35]. The authors developed an attention-based neural machine translation model between Arabic and English. They achieved the highest accuracy using PATB tokenization, with 51.2 and 49.7 BLEU points on NIST 2005 dataset for SMT and NMT, respectively. However, they argued that Arabic lacks a comprehensive parallel corpus, which is the main challenge for the utilization of NMT. Comparative studies were also presented by Alrajeh [54] and Oudah *et al.* [16] to investigate the impact of applying normalization and tokenization on both NMT and SMT systems. The comparison showed a significantly gain on Almahairi *et al.* [35] results.

Belinkov *et al*. [68] investigated several tests on different NMT systems to identify the best possible morphological language-related representations among morphology-rich languages. To make a fair comparison, they trained several NMT models on the intersection of the training data based on same Arabic sentences with different translations. They found out that character-based representations using a CNN performed better at learning word morphology than their word-based counterparts. For instance, for Arabic-English the character-based segmentation showed an improvement of more than 3 BLEU points over the word-based one. On the other hand, the English-Arabic translation direction results of the word-based variant were not improved.

More recently, Martínez *et al.* [102] compared the effect of using linguistic preprocessing to decompose the target words of an Arabic-French factored NMT model. Their model predicted the lemma, and the concatenation of the following factors: POS tag, tense, gender, number, person, and the case information at decoding time. The model performed the training using a small or large parallel training dataset to simulate low-resource and rich resource behaviors, respectively. They applied BPE segmentation for both their Factored and standard NMT architectures. Their evaluation results on several test sets showed that the factored NMT models were far better under low-resource conditions by an improvement of around 3 to 6 BLEU points over the baseline NMT.

2) FREE WORD ORDER

Recently, Alqudsi *et al.* [7] proposed a method to handle the word ordering problem in the context of Arabic-to-English MT. Their proposed method combines rule-based MT with the EM algorithm. They used parallel data from the United Nations (Arabic-English) corpus. They trained their model using 632 sentence pairs and reserved 271 sentence pairs for testing. Their results showed an increase of up to 0.89 BLEU points over their RBMT baseline system.

3) WORD SENSE DISAMBIGUATION (WSD)

Hadni *et al.* [91] proposed a knowledge-based approach using English WordNet and Arabic WordNet. To resolve the issue with ambiguous terms, the authors used MT to select the closest concept for an ambiguous word using the relationships between the ambiguous word and the different concepts in local context. The results show that the proposed approach outperforms other techniques for Arabic WSD, yielding an accuracy of 73.2% when using support vector machine classifier.

4) ARABIC NAMED ENTITY

Recently, Ameur *et al.* [67] proposed a transliteration attempt using an attention-based encoder-decoder for the task of MT between the Arabic and English. The results proved the efficiency of their approach in comparison to some previous research. Lately, Alkhatib and Shaalan [18] applied a hybrid deep learning based on CNN followed by Bi-LSTM and CRF to boost NE transliteration. Their results on the corpora ANERcorp and Kalimat show that their model can be efficient in machine transliteration achieving state-of-the-art results for Arabic-English.

5) ARABIC VOCALIZATION

To restore the diacritics where there is an ambiguity, Alnefaie [103] presented a system that combines morphological analyzers and context similarities. The goal of the morphological analyzers is to generate all word candidates for the diacritics, and the model eliminates word ambiguity through a statistical approach and context similarities. Their

result shows that out of 80 paragraphs their system resolved 57 cases.

6) DIALECTAL VARIATION

For dialectal Arabic, El-taher *et al.* [104] built a translation model as a rule-based approach that relies on transfer rules from the Egyptian dialect to the Modern Standard Arabic using different rules and DA-MSA dictionary, attaining a BLEU score and an average accuracy of 88.7%, 81% respectively. Recently, Farhan *et al.* [81] presented a novel deep learning system that aim to translate dialectal sentences into both supervised and unsupervised settings. The highest BLEU score obtained in the unsupervised setting is 32.14, which is remarkably high compared with the highest BLEU score 48.25 obtained in the supervised setting.

7) GENDER BIAS

In Arabic, as in other languages with grammatical gender, gender-blind single-output MT from English often result without gender agreement. Elaraby *et al.* [69] proposed an English-Arabic NMT to enable producing gender-aware translation. They provided a set of annotation rules to generate the data that marks speaker and listener gender as meta-data input on the source sentence. They trained a NMT model with the labeled data and large set of unlabeled data. Their proposed approach led to an improvement of two BLEU points over the baseline model.

8) ERROR ANALYSIS

Solyman *et al.* [73] proposed an unsupervised method to generate large-scale synthetic training data to overcome the challenge of the scarcity of training data for Arabic. The method is based on confusion function to increase the amount of training set. They applied fine-tuning to improve the performance and get more efficient results in the task of Grammar Error Correction.

Similarly, Saadany and Orasan [74] investigated the challenges involved in translating book reviews from Arabic-English. They analyzed errors that lead to incorrect translation of sentiment polarity and proposed an error typology specific of the translation of Arabic User Generated Content. They addressed this problem by integrating sentiment information in the encoding stage and fine-tuning an NMT model with respect to sentiment polarity using synthetic data.

Some research investigated MT samples with native speakers so they could review the linguistic aspects of MT errors [13], [40]; other research works used neural networks to detect errors [41], [73] or to correct them [42].

D. COMPARISON OF AMT APPROACHES ADDRESSING TECHNICAL CHALLENGES

Similarly, we have classified reviewed research works addressing the technical challenges focusing more on neural-based NMT approaches in Table 3.

1) LOW-RESOURCE

There are some worth mentioning works attempting to handle low-resource languages. For instance, the authors in [71] proposed a novel triangular training architecture (TA-NMT) to improve the translation performance of low-resource pairs. In this architecture, a rich language is taken as the intermediate latent variable, and translation models of the rich language are jointly optimized with a unified bidirectional Expectation-Maximization (EM) algorithm. Their method significantly improves the translation quality of rare languages such as Arabic on MultiUN and IWSLT2012 datasets.

Ji *et al.* [79] proposed a transfer learning approach based on cross-lingual pretraining. First, a universal encoder is trained on several monolingual source languages using a shared feature space. Then, the whole NMT model is trained using parallel data and used in zero-shot translation. The tests on MultiUN (Arabic, Spanish, and Russian) showed that the approach significantly outperforms both pivot-based and multilingual NMT baselines. For the tasks of Spanish and Russian translation from and to Arabic reported an improvement that ranges from 1 to 3 BLEU points over their multilingual NMT baseline model.

Comparatively, multilingual MNMT translation is gaining more interest where a single NMT model is optimized for the translation of multiple language pairs [33]. Multilingual NMT eases model deployment and can enable zero-shot translation i.e. direct translation between a language pair never seen in training. Zhang *et al.* [80] experimented any-to-any multilingual translation on 100 languages where Arabic language is included and no parallel data is available. Their results reveal that zero-shot translation quality still trails behind the pivot-based bilingual NMT translation [80].

Almansor and Al-Ani [70] presented a character-based hybrid NMT model that combines both RNN and CNN networks. They trained their model on a very small portion of the TED parallel corpora containing only 90K sentence pairs, notably IWSLT 2016 Arabic-English. For the case of English-Arabic translation, the improvement in BLEU score exceeded 10 BLEU points. However, they reported noticeable improvements for Arabic-English in comparison to the openNMT word-based NMT model.

Liu *et al.* [75] presented mBART a denoising auto-encoder extended by pre-training BART [105] on several monolingual language corpora. Their model is designed to be fine-tuned to translation tasks without language-specific modifications or initialization schemes. mBART initialization leads to significant gains (up to 12 BLEU points) across low and medium-resource pairs ($<$10M bi-text pairs), without sacrificing performance in high-resource settings. These results further improved with back-translation (BT). For document-level MT, pre-training improved results by up to 5.5 BLEU points. For the unsupervised case, they reported consistent gains and produced the first non-degenerate results for less related language pairs. As far as Arabic

TABLE 2. Comparison of main AMT Approaches addressing Linguistic challenges (m: million, k: thousand).

Challenge	Reference	Year	Language pair	Dataset/Domain	Evaluation metric	Result
Complex morphology	Almahairi *et al.* [35]	2016	Arabic<>English	1.2M sentence pairs from LDC2004T18, LDC2004T17 and LDC2007T08/ News	BLEU	For English-Arabic improvement of 4.98 BLEU points over the baseline. For Arabic-English, they improve by 2 BLEU points
	Belinkov *et al.* [68]	2017	Arabic<>English, French, German, Czech, Hebrew	IWSLT2016/ TED Talks	BLEU	Improvement of 3 BLEU points over the baseline for Arabic-English, no improvement for English-Arabic
	Alrajeh [54]	2018	Arabic-English	NIST MT from 2005, 2006, 2012/ News	BLEU	NMT produces a gain of +13 BLEU points compared to SMT
	Oudah *et al.* [16]	2019	Arabic-English	LDC2004T18, LDC2004T14, LDC2007T08, LDC2010T12, LDC2010T14/ News	BLEU	Improvement of MT quality by 55.64% and 53.54% in BLEU scores for SMT and NMT, respectively
	Martínez et al. [102]	2020	Arabic<>French	4.6M for Arabic and 4.7M for French from News Commentary version 9/ News	BLEU	Improvement of around 3 to 6 BLEU points over the baseline
Free word order	Alqudsi *et al.* [7]	2019	Arabic-English	632 sentence pairs (train) and 271 sentence pairs(test) from UN corpus/ Government	BLEU	Improvement in the BLEU points EM method solves the ambiguity problem
Word Sense Disambiguation	Hadni *et al.* [91]	2016	Arabic-English	153 articles from Essex Arabic Summaries Corpus (EASC)/ General	Precision, recall, F_1	Achieved the highest scores of precision (0,718), recall (0,746) and F_1(0,732)
Arabic Name Entity	Ameur *et al.* [67]	2017	Arabic<>English	35.4M sentences from United Nation, Open Subtitles, News Commentary and MIWSLT2016/ News, government documents	Word Error Rates (WERs) and Character Error Rates (CERs)	Lowest error rate of WER 65.16% and CER 16,35%
	Alkhatib and Shaalan [18]	2020	Arabic-English	55760 words from ANERcorp and Kalimat/ Religion, sport, culture, news, economy	F-measure, precision, recall	Transliteration improvement of 94.2%, 95.3%,92.1% for person, location and organization, respectively
Arabic Vocalization	Alnefaie and Azmi [103]	2017	Arabic-English	80 paragraph/ Sports, politics and social news	Human evaluation	The system was able to eliminate ambiguity in 57 cases from 80 paragraphs

TABLE 2. *(Continued.)* **Comparison of main AMT Approaches addressing Linguistic challenges (m: million, k: thousand).**

Dialectal variation	El-taher *at el.* [104]	2016	Egyptian dialect to MSA	2400 proverbs, 380 youth expressions, 90 expressions	BLEU, Average number of correct sentences	The model attained a BLEU score and an average accuracy of 88.7%, 81% respectively.
	Farhan *et al.* [81]	2020	Egyptian-Jordanian-Saudi dialect to MSA	309K sentences from Open Subtitles 2016	BLEU	Achieved the highest BLEU of 32.14 and 48.25 for unsupervised and supervised setting, respectively.
Gender bias	Elaraby *et al.* [69]	2018	English-Arabic	4M sentences from Open-Subtitles/ General	BLEU	Improvement of the translation quality by 2 BLEU points
Error analysis	Solyman *et al.* [73]	2020	Arabic-English	20M words from, SCUT coupes, QALB 2015/ General	Precision, recall, F_1	Achieved the highest scores of precision (80.23%), recall (63.59%) and F $_1$(70.91%)
	Saadany and Orasan [74]	2020	Arabic-English	230K sentences scraped from Goodreads	Precision, recall, F_1	Transformer with tagging and pre-trained achieved highest scores of precision (0.85), recall (0.79) and F $_1$(0.81)

is concerned, their mBART25 (pretrained on 25 languages) has led to an increase of 10.1 BLEU points for Arabic-English MT.

Similarly, Lin *et al.* [82] proposed multilingual Random Aligned Substitution Pre-training (mRASP); an approach to pre-train a universal multilingual neural machine translation model for many languages, which can be used as a common initial model to fine-tune on arbitrary language pairs. It brings words and phrases with similar meanings across multiple languages closer in the representation space. They pre-trained their model on 32 language pairs jointly with only public datasets. The model is then fine-tuned on downstream language pairs to obtain specialized MT models. Accordingly, the results on English-Arabic pair are improved by 1.8 BLEU points.

To improve MT models without any external sources of data, Abid [76] proposed a NMT model. The author accomplished this by bootstrapping existing parallel sentences and complement this with multilingual training to achieve strong baselines. They created a 4-way benchmark dataset between Egyptian, Levantine, MSA, and English, freely available to the community. The results of the conducted experiments suggest that a multilingual model of dialects and MSA, along with bootstrapping, achieves the best results by 2.56 (9%) BLEU score.

Fan *et al*. [77] introduced M2M-100, a new Many-to-Many MNMT model trained on 7.5 Billion sentences. The model can translate between 100 languages to and from English. The underlying dataset was mined from Common Crawl by language groupings to avoid mining every possible direction. Their results showed that M2M-100 outperforms English-Centric multilingual models trained on data where either the source or target language is English. As far as Arabic is concerned, their multilingual NMT achieved a BLEU score increase of 15.5 points on average over English-Centric baseline when translating directly between Arabic-English directions.

Another model XLM-T, proposed by Ma *et al*. [78], initializes MNMT with a pretrained cross-lingual Transformer encoder, and fine-tunes it using multi-lingual parallel data. For this model, the authors conducted extensive experiments with 10 language pairs from WMT datasets and 94 language pairs from OPUS datasets. The method achieves significant and consistent gains on both large-scale datasets. The overall improvement is 0.9 and 0.4 BLEU points by averaging all 94 English-X and X-English language pairs. As far as Arabic is concerned, their method achieved a BLEU score increase of 1.8 and 0.4 points on average over a Multilingual NMT baseline when translating Arabic-English and English-Arabic, respectively.

2) OUT-OF-DOMAIN

To alleviate the issue with translation mismatch related to out-of-domain, Oudah *et al.* [16] studied the performance of NMT system under morphology-based and frequency-based tokenization schemes and BPE on in-domain data. They evaluated their best performing models on out-of-domain data yielding significant improvements of 37.96% in BLEU score.

3) OOV

Ataman *et al.* [57] proposed a novel NMT decoding method that models word-formation via a hierarchical latent variable that simulates morphological inflection. The model generated words one character at a time by combining two latent variables; the lemmas and the inflectional features. The proposed method is compared to subword and character-level decoding methods on the translation task for three morphology-rich languages. The results show slight improvement of 0.51 BLEU points over the best performing baseline on the task English-Arabic translation. Aqlan *et al*. [106] proposed a romanization system that converts Arabic scripts to subword units to deal with the unknown words problem on the task of MT between Arabic and Chinese. They investigated the effect of their approach on the NMT performance while using various segmentation scenarios. They performed extensive experiments on Arabic-Chinese and Chinese-Arabic translation tasks and showed that their proposed approach can effectively tackle the unknown words problem and improve the translation quality by up to 4 BLEU points.

4) SENTENCE LENGTH

Oudah *et al.* [16] combined two MT systems (SMT and NMT) via a system selection to handle specifically long sentence. The authors used SMT because it performs better for very long sentence, i.e. above 50 tokens. Their system significantly outperforms previous results reported by +4 BLEU points. Recently, Berrichi and Mazroui [84] addressed the problems of OOV words and long sentences. They have developed two techniques for segmenting long sentences into smaller sub-sentences. The first uses a list of 87 English lexical markers accompanied by their Arabic equivalents to serve as junction points between two sub-sentences that can be translated separately. On the other hand, the second technique integrates into the NMT model parallel phrases extracted by an SMT system. Their results on the English-Arabic pair show that the proposed approaches considerably improve the translation quality compared to the basic NMT system by 2.81 BLEU points.

5) WORD ALIGNMENT

Recently, Ellouze *et al.* [17] proposed a hybrid approach to improve the alignment results of the GIZA++ toolkit.[24] Their proposal uses linguistic features like morpho-syntactic tags, syntactic patterns, transliteration and statistical features such as mutual information and harmonic mean. They trained their alignment model using an English-Arabic medical corpus tested on Cambridge dictionary. The authors stated that their results showed an improvement in both the alignment and the translation quality.

Similarly, Berrichi and Mazroui [107] proposed a new alignment process, which is based on morphological preprocessing and incorporation of a bilingual dictionary as an additional source to support some alignment choices. Test results on two corpora namely United Nations parallel corpus and MulTed corpus show that the use of the dictionary has improved the quality of alignment and therefore increase the BLEU score by 5%.

6) COMPUTATIONAL OVERHEAD

Ensembling is a well-known technique in NMT to improve system performance. Durrani *et al.* [27] presented QCRI's MT system as an ensemble system where they trained an NMT system using different genres through fine-tuning, and applying ensemble over eight models. The ensemble system outperforms their very strong PBSMT system. Yet, it is computationally expensive since the decoder needs to apply eight NMT models rather than only one.

To boost the speed of translation, the Transformer [2] was proposed to solve the problem by using CNN together with attention models. Recently, Shapiro and Duh [72] proposed a pipelined approach with a dialect-tuned model using Transformer architecture. The results show an error rate less than 20% even using a small BPE size and large training data.

More recently, Stergiadis *et al*. [83] described and empirically evaluated a multidimensional tagging method for passing sentence-level information to the models. The models are simultaneously fine-tuned on several closely related, yet succinctly different sub-domains. Their human and BLEU evaluation results show that the method can be applied to the problem of multi-domain adaptation and significantly reduce training costs without sacrificing the translation quality on any of the constituent domains. For Arabic-English, the authors reported an improvement of 0.27 BLEU points and 3.09 points over the baselines.

E. TREEBANKS, DATASETS AND CORPORA

In general, MT approaches use custom datasets, large corpora and fully parsed treebanks for training, evaluation and optimization processes. Hence, we present in following major resources used Arabic to English MT. Indeed, most of the datasets came from the Workshop for Machine Translation (WMT)[26] followed by NIST[27] and International Workshop on Spoken Language Translation (IWSLT).[28] Additionally, Table 4 summarizes well-known datasets used for the AMT task.

[24]http://www.fjoch.com/GIZA++.html

[25]https://conferences.unite.un.org/UNCorpus
[26]http://www.statmt.org/wmt20/
[27]https://www.nist.gov/human-language-technology
[28]http://iwslt.org/doku.php

TABLE 3. Comparison of main AMT approaches addressing technical challenges (m: million, k: thousand).

Challenge	Reference	Year	Language	Dataset/Size	Evaluation metric	Result
Low resource	Shuo et al. [71]	2018	English-Arabic	MultiUN and IWSLT2012/ News, government	BLEU	Improvement of 0.7 and 0.9 BLEU point
	Almansor and Al-Ani [70]	2018	English<>Arabic	90K sentences from TED, IWSLT 2016/ General	BLEU	Translation quality reached 18 BLEU score in Arabic-English (BLEU greater than 15%)
	Ji et al. [79].	2020	English as pivot, Arabic, Russian, Spanish	Europarl corpus, MultiUN corpus/ Government	BLEU	Improvement of 1 to 3 BLEU points
	Zhang *et a.* [80]	2020	many-many	55M sentences from OPUS-100 and 2000 sentences for evaluating Arabic/ General	BLEU	Improvement of 10 BLEU points, approaching conventional pivot-based methods
	Farhan et al. [81]	2020	Arabic dialect-English	Multi-parallel corpus of 309K sentences	BLEU	Improvement of MT quality by 32.14%
	Liu et al. [75]	2020	Arabic<>English	2869M tokens from CC25 corpus/ General	BLEU	Improvement of 10.1 and 4.7 BLEU points for Arabic-English and English-Arabic, respectively
	Abid [76]	2020	Arabic dialect<>English	402K sentences and 3.80M tokens for Egyptian,138K sentences and 1.27M tokens for Levantine, and 1.49M sentences and 24.45M tokens for MSA/ News, general	BLEU	Improvement of 2.56 (9%) BLEU score.
	Fan *et al.* [77]	2020	Many to Many	7.5 Billion sentences/ General	BLEU	Improvement of 15.5 BLEU points in average for Arabic<>English
	Ma *et al.* [78]	2020	100 to 100	55M sentences from WMT, OPUS-100/ General	BLEU	Increase of 1.8 and 0.4 BLEU points on average for Arabic-English and English-Arabic, respectively
	Lin *et al.* [82]	2021	English<> Arabic	1.2M sentences from TED, OPUS/ General	BLEU	Improvement of 1.8 and 3.7 BLEU points for English-Arabic and Arabic-English, respectively
OOV	Aqlan *et al.* [106]	2019	Arabic<>Chinese		BLEU	Improvement of 4 BLEU points
	Ataman *et al.* [57]	2020	English-Arabic	TED Talks/ General	BLEU	Improvement of 0.51 BLEU points over the baseline
Out-of-domain	Oudah *et al.* [16]	2019	Arabic-English	1,075 sentence pairs from LDC2010T12 (tuning), 1,056 sentence pairs from LDC2010T14 (in-domain) and 1,535 sentence pairs from LDC2014T02 (out-of-domain)/ News	BLEU	Improvement of 56.18% and 37.96% in BLEU score for in-domain test and out-of-domain test, respectively

TABLE 3. ***(Continued.)*** **Comparison of main AMT approaches addressing technical challenges (m: million, k: thousand).**

Sentence length	Oudah *et al.* [16]	2019	Arabic-English	1.2M sentence pairs from LDC2004T18, LDC2004T14, and LDC2007T08/ News, government	BLEU	Improvement of 4 BLEU points
	Berrichi and Mazroui [84]	2021	Arabic-English	40,539 sentence pairs from UN corpus, 1,570 sentence pairs from Arab-Acquis/ Government	BLEU	Improvement by 2.81 BLEU points over the baseline
Word alignment	Ellouze *et al.* [17]	2018	English-Arabic	Cambridge Dictionary/ General	Error rate	Improvement of the alignment produced by GIZA++ from 56.31% to 87.16%
	Berrichi and Mazroui [107]	2019	English-Arabic	16756 sentences from UN corpus[25] and 1400 sentences from MulTed corpus/ Law, business	Alignment error rate, BLEU	Improvement of word alignment and translation quality to 5%
Computational overhead	Durrani *et al.* [27]	2017	Arabic<>English	240K sentences from TED, 153K sentences from QED, 18.5M sentences from UN 40M sentences from OPUS/ Law, education	BLEU	Improvement of 2 BLEU points
	Shapiro and Duh [72]	2019	Arabic dialect-English	LDC2012T09/ News	Error rate	Error rate less 20%
	Stergiadis *et al.* [83]	2021	Arabic-English	71M sentences fom OPUS, 3M sentences/ Travel	BLEU, Human evaluation	Improvement of 0.27 BLEU points and 3.09 points over the baselines

Penn Arabic Treebanks (PATB) [108] are newswire articles from a variety of news sources. They are available on the Linguistic Data Consortium (LDC) website. PATB provides tokenization, complex POS tags, and syntactic structure, diacritizations, lemma choices and various semantic tags.

MultiUN [109], **UN corpus** [110] are parallel corpora extracted from the official documents of the United Nations (UN). They are freely available in all 6 official languages of the UN (Arabic, Chinese, English, French, Russian, and Spanish), consisting each of around 300 million words per language.

Arabic Gigaword [111] is a corpus released from LDC that contains MSA source texts and corresponding English translations selected from broadcast news data. The corpus consists of LDC2003T12, LDC2006T02, LDC2007T40, and LDC2009T30 with 1,124,609 sentence pairs of Arabic source text and their English translations.

Open Subtitles (OPUS) [112] is a free, multilingual parallel corpus of 90 languages including Arabic-English sentences collected from the translated movies and TV subtitles. This corpus consists of huge data from multiple domains and sources as well as many large datasets such as Tanzil [112] which is a collection of Quran translations.

WIT^3 [113] corpus provides multilingual transcriptions and high-quality translations of diverse Technology, Entertainment, and Design (TED) talks. It has many interesting features like diversity of topics, spoken language transcriptions, and user-generated translations, although the review process ensures a reasonable translation quality.

QED corpus is a dataset implemented by the Qatar Computing Research Institute [114] to prepare data for machine translation tasks. The corpus contains community-generated video subtitles from well-known educational platforms, such as TED and the Khan Academy. It consists of 2.6 million Arabic words and 3.9 million English words.

Human judgment corpus [39] is a parallel corpus of two thousand sentences from the news domain, too small for training a system but potentially useful for MT evaluation.

Annotated Al Jazeera Dialectal speech [115] is a speech corpus that contains dialect-level labels for 57 hours of dialectal Arabic speech (Egyptian, Levantine, North African, and Gulf) from Al Jazeera news channel from June 2014 to January 2015, as well as confidence labeled levels. This corpus also contains 94 hours of dialectal Arabic speech automatically labeled by linking speaker information from the human-labeled set.

Arab-Acquis [116] is a large publicly available dataset for evaluating machine translation between 22 European languages and Arabic. Arab-Acquis consists of over 12,000 sentences from the JRC-Acquis corpus translated twice by professional translators, once from English and once from French, and has over 600,000 words.

Tashkeela [97]: a corpus containing 75.6 million vocalized Arabic words. Recently, Fadel *et al.* [117] prepared a version of this corpus for benchmarking purposes. It is a cleaned version with pre-defined split for training, testing and validation sets resulting in 24.6 million vocalized Arabic words.

Arabic-SQuAD [28] dataset is the Arabic Translation of Stanford Question Answering Dataset (SQuAD) [118], a reading comprehension dataset consisting of 100,000+ questions posed by crowd-workers on a set of 536 Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. This resulted in 48,344 questions on 10,364 paragraphs.

F. AMT EVALUATION

Recently, there are some efforts done to ensure the effectiveness of AMT systems. Indeed, evaluation in AMT is critical and challenging for MT developers to evaluate progress of their work as well as for MT users to select among available MT engines for their language pairs of interest. Table 5 summarizes the evaluation methods proposed recently including results, the used datasets, and tools.

Hussein and Awab [10] addressed the translation of verb-noun collocation from English into Arabic using Google Translate and Bing online translation engines. They evaluated the outputs using human translations and a new proposed metric called verb-noun-collocation-value (NVCOLV). The results showed that Bing scored a verb-noun collocation value of 0.72 with a trend estimation ranging between 0.65 and 0.67. Google scored a verb-noun collocation value of 0.75 (3% higher than Bing) with a trend estimation ranging between 0.63 and 0.85.

The effectiveness of correction techniques of raw MT outputs enhance the output of MT. Specifically, the MT outputs have different linguistic errors [11]–[13] that needs human intervention to rectify it. For this reason, the MT translated texts often need manual, semi-automatic, automatic corrections, known as Post-Editing (PE). Post-editing strategies involved correcting mistakes, inserting omitted words and deleting inserted or repeated words. The error classification task aims to identify and classify actual errors in a translated text that can be aggregated to augment corpora [119].

More recently, a new hybrid MT tool is proposed by Ehab *et al*. [46] as a combination of EMBT and Translation Memory (TM) to translate English medical text to Arabic one. The overall accuracy with a translation memory achieved the highest score of 77.17 and 63.85 for two datasets in the internal medicine domain, which were the highest score using BLEU score.

It can be concluded that the findings of the above works including the MT evaluation studies [1], [25], [43], prove the ineffectiveness of Google Translate when rendering Arabic-English translation. The problem is worse when translating domain-specific topics such as legal texts. This is mainly due to domain mismatch [21], which can be tackled using multi-domain NMT [121].

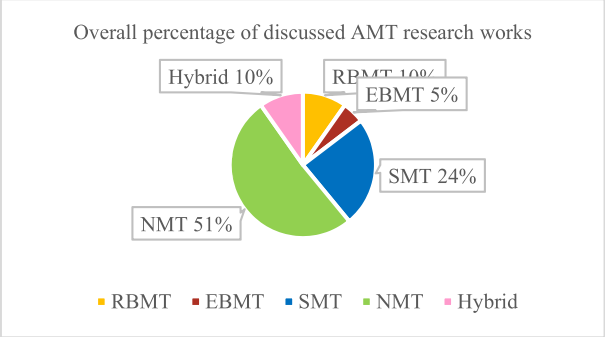

FIGURE 4. Overall percentage of discussed AMT research works.

V. FINDINDS AND DISCUSSION

Throughout this review, we draw many findings, ideas and open issues that could help to reach a certain level of human expectations concerning AMT accuracy and fluency. In total, we reviewed 40 AMT research works and 8 AMT assessment works. Figure 4 show the statistics of 40 AMT works per each challenge category and per MT system type. Note that the statistics reflect only the number of AMT research studies and works discussed in this paper. In this section, we start by listing main observations and then proceed with open challenges.

A. OBSERVATIONS

After a careful analysis, we can make the following main observations:

- Most of the early research works in AMT studied heavily linguistic issues of Arabic, however, few research works addressed technical issues. Whereas, recently, most of the research works are addressing technical more than linguistic issues of Arabic. There are many reasons for the direction of studies. One of the main reasons is the great advance in computational language technology. Nonetheless, the recent research studies based on NMT which constitute the majority, have a strong learning capability that makes explicit linguistic features redundant. However, for Arabic, we have few results [16], [35], [54] indicating that linguistic information is beneficial to NMT models.
- This analysis proves that addressing linguistic challenges of Arabic has a high impact on the accuracy of the Arabic SMT and that NMT is superior to SMT in all cases. In particular, the preprocessing of Arabic did improve the translation quality [16], [35] and NMT is able to perform very well even on a small corpus. The research on Arabic NMT has grown in the last four years

TABLE 4. Most important existing Treebanks, datasets and corpora for AMT (m: million, k: thousand).

Resource	Year	Type	Domain	Size	Availability	Remark
PATB [108]	2003-2016	Treebank	News	12 parts on newswire, size ranges 145K - 1M tokens	https://catalog.ldc.upenn.edu/	(+) Fully parsed treebank, a good option for Arabic syntactic analysis and MT task. (-) not free, restricted domain
MultiUN [109]	2010	Parallel corpus	UN documents	81,41M sentences	http://www.euromatrixplus.net/multi-un/	(+) a very useful resource for MT developers with 21 translation directions
Arabic Gigaword 5th edition [111]	2011	Monolingual corpus	News	1,125K sentences	https://catalog.ldc.upenn.edu/LDC2011T11	(+) distinct resources of Arabic newswire (-) not free, restricted domain
OPUS [112]	2012	Parallel corpus	Movies subtitles	60M sentences	http://opus.nlpl.eu/	(+) free, multiple domain (-) low quality according to [54]
WIT^3 [113]	2012	Parallel corpus	TED Talks	IWSLT 2016: 200K sentences	https://wit3.fbk.eu/	(+) free, multiple reference, ready-to-use version for research purposes, regular update (-) translations are segmented based on sound, hence the correspondence between captions and sentences is weak
QED [114]	2014	Parallel corpus	Education	2,6M words	http://alt.qcri.org/resources/qedcorpus/	(+) open multilingual (parallel) corpus
Human judgment corpus [39]	2014	Parallel corpus	General	2K sentences	http://nlp.qatar.cmu.edu/resources/AL-BLEU/	(+) free, high annotation quality, a good option for MT evaluation, (-) small corpus
Annotated Al Jazeera Dialectal speech [115]	2015	Dialectal corpus	News	47,696 videos	http://alt.qcri.org/resources/aljazeeraSpeechCorpus/	(+) free, automatically labeled, a good option for speech translation task (-) specific domain
QALB corpus [119]	2016	Corpus	News	100K words of English news articles	http://nlp.qatar.cmu.edu/qalb/	(+) free, texts a good option for automatic post-editing tasks
UN corpus [110]	2016	Parallel corpus	UN documents	18,539,207 sentences	https://conferences.unite.un.org/uncorpus	(+) fully aligned sub-corpus, a valuable resource for studying pivoting techniques and multi-source or multi-target approaches
Arab-Acquis [116]	2017	Dataset	Law	12K sentences	https://camel.abudhabi.nyu.edu/arabacquis/	(+) free, very high quality for MT evaluation task (-) small corpus
Tashkeela [97]	2017	Monolingual corpus	Religion	75M words	https://sourceforge.net/projects/tashkeela/	(+) free, a good option for diacritization related tasks (-) restricted domain, needs a clean version [117]
Arabic-SQuAD [28]	2019	Dataset	Wikipedia	48K questions	https://github.com/husseinmozannar/SOQAL	(+) free, a good option for QA (-) small corpus
Qatari heritage expressions [120]	2020	Parallel corpus	General	1000 colloquial Qatari dialect expressions	https://data.world/saraalmulla/qatari-heritage-expressions	(+) free, a good option for dialectal MT evaluation (-) not enough data for MT dialectal training, missing corpus statistics

especially after the release of GNMT model in 2016. In addition, rule-based MT methods are getting less and less attention as standalone method due to their poor performance and high costs [30].

TABLE 5. Comparison of the assessment results of Arabic language on specific MT systems.

Reference	Year	Language pair	Dataset	MT systems	Evaluation metric	Outstanding system
Hussein and Awab [10]	2016	English-Arabic	17 sentences from *Oxford Advanced Learners' Dictionary*	Google Translate and Microsoft Bing	NVCOLV	Google Translate
Al-Rukban and Saudagar [122]	2017	English-Arabic	100 sentences (present, past, conditional, passive, imperative, questions)	Google Translate, Bing Translator and Golden Alwafi.	BLEU, GTM	Golden Alwafi achieves highest BLEU and Google Translator attains highest GTM
El Marouani *et al.* [19]	2018	English-Arabic	1383 sentences NIST 2005 corpus, a small dataset of translated Wikipedia articles.	CMU, QCRI Translate, Google Translate, Microsoft Bing and Columbia	AL-TERp	Google Translate, Microsoft Bing
Almahasees [9]	2018	Arabic-English	An example from *Jordan Petra News agency*	Google Translate and Microsoft Bing	Linguistic Error analysis	Google Translate
Jabak [11]	2019	Arabic-English	8 texts of different lengths from *Thinking Arabic Translation* book	Google Translate and human interference	Lexical and syntactic errors	Human interference
Ehab *et al.* [46]	2019	English-Arabic	259 medical sentences of internal diseases 509 medical sentences	Google Translate and EBTM + Translation memory	BLEU	EBTM + Translation memory
Mahesh *et al.* [123]	2020	Arabic-English	6 sentences	Google Translate, Bing Translator, Systran, PROMT, Babylon, WorldLingo, Yandex and Reverso	BLEU	Google Translate
Alkhawaja *et al.* [6]	2020	English-Arabic	100 passages	Google Translate	Error analysis, BLEU	Google Translate

- The Transformer model is being increasingly used to improve the performance of NMT and speed up the inference process. However, we found only few Transformer-based works [72], [79] addressing dialectal Arabic, for more details see Tables 2 and 3.
- Training NMT with transfer learning [79], fine-tuning [72], multilingual NMT [72], zero-shot translation [33] and zero-shot transfer [66] approaches have shown promising translation results between multiple languages especially for low-resource languages [80] or no resource language [79]. These works showed that these models were able to capture shared representational features across languages, thus offering better transfer capabilities that lead to larger improvement in translation quality. Unfortunately, there are few efforts [79], [80] investigating these techniques for Arabic language. Especially, for MNMT, the current development achieves low accuracy compared with its counterpart which trains an individual model for each language pair. Thus, for Arabic it is worth to boost the MNMT using different approaches and techniques such as [80].
- Training first a teacher model and then distill its knowledge into a student model is considerably greater performance in NMT [124]. Specifically, training NMT as the teacher models on alternating blocks of authentic and synthetic data ("block-regime back-translation" (block-BT) rather than shuffling with mixing authentic and synthetic training data [4] led to a gradually increasing learning curve and thus improving the model performance. However, such novel training procedure is not investigated for AMT yet. Although, we observe reasonable gains from transfer learning in many languages, mostly significant. The only non-significant gain is from Arabic-Russian which does not share the script with the child at all [125] when trained with transfer learning as parent-child model.
- It is observed that NMT is computationally expensive [27], however, it requires least supervision compared to SMT systems where more efforts are required to build the different models. On the other hand, the SMT requires large parallel datasets and lexicons [16], [43] that represent a big barrier for low resource languages, such as Arabic. However, the new NMT Transformer

model architecture computes many steps slightly more computationally efficient than previous NMT models.

- Most of the AMT approaches focus on the translation of news and official texts, whilst few attempts focus on domain specific translation such as medical domain [46]. Specifically, most of the used parallel data available to the researcher was limited to texts produced by international organizations, parliamentary debates or legal texts [126]. Unfortunately, existing single-domain AMT methods do not work well for multiple domains. Thus, multi-domain NMT approaches are more in demand to tackle this limitation.
- Many linguistic challenges such as WSD are not well addressed to-date due to the difficulty in determining the correct meaning that should be chosen for the translation, while some are handled using hybrid solutions such as NE [18]. Likewise, subject embedding makes Arabic as pro-drop language a continual challenge for MT.

On the other hand, we conclude the following secondary observations:

- With the substantial performance improvements brought to MT by neural approaches, a growing interest in translating between pairs of similar languages, language varieties, and dialects has been observed. As for Arabic dialects, we observe many works investigating MT between similar dialects and MSA [72], [98]. Indeed, most contributions are dedicated to translating between dialects, MSA and English. For other languages, no results are available. In this context, all contributions concern mainly English language [98]. In terms of translation direction, most of the contributions translate from dialects to MSA or English, whereas there is very little work that uses the dialect as target language. Likewise, another challenge on developing a NMT model for Arabic dialect to MSA is the absence of the standardized orthographies for all Arabic dialects. It includes morphological differences which are evident in the use of clitics and affixes that do not exist in MSA. However, training NMT models usually require a large amount of annotated data, which is often unavailable in the case of low-resource languages such as Arabic dialects.
- Most contributions for Arabic dialects exploit the proximity between Arabic dialects and MSA when translating between them. Indeed, MT between Arabic dialects is getting much easier than for other languages families. Concerning, the proximity and similarity between dialects and MSA, it is possible to translate multiple dialects into MSA using same hyper dimensional space for common words. This way, it is possible to build a multi-dialect translator by training the model on only one dialect. This reduces the hustle of collecting parallel corpora for every single dialect as studied in Farhan *et al.* [81]. We believe this approach can be generalized to other similar languages where synonyms of different but close languages can have similar vector embeddings. Indeed, all contributions are dedicated to translating between dialects, MSA and English. The plausible techniques that address dialectal Arabic MT challenges are still under study [99].
- Most of the authors evaluated their MT systems using BLEU metric or apply other metric from other languages such as from English, only very few authors, [19], [39] used adapted metrics for Arabic language.
- Most of AMT assessors and users compare the impacts of special Arabic processing on MT using small domain specific datasets [9] on different MT systems. Subsequently, it is hard to assume whether an MT system is better for Arabic or not. From this conclusion, it is apparent that dataset benchmarking is still missing for multi-domain translation evaluation for Arabic language.
- The lesson learned from the assessment works is that Google Translate is being assessed as outstanding MT system among others. However, some unsettled issues such as linguistic errors are still prevalent and will continue as the performance of parsers on Arabic is far below their performance on English. According to [6], [9], [11], they conclude that the most dominant errors in the MT output were mistranslation errors, followed by corruption of the overall meaning of the sentence and then orthographic errors. In terms of adequacy and fluency, the results for English-Arabic translation, Google Translate produces sentences with relatively few errors, and the translated text is fluent to some extent as argued in [6], [10].

Additionally, when comparing our findings and anticipated AMT results to other languages pairs such as Czech-English and vice-versa, we observe followings:

- Although both languages; Arabic and Czech are morphologically-rich [68], have free word order, and share most of the technical challenges, we observe with regard to the latest WMT20 findings [127], a constant improvement in the baseline systems for Czech MT in contrast to AMT.
- Czech MT is constantly improving using better training data, novel training procedures, doubling the encoder depth (to 12 layers), robust training with source-side noising documents [100], etc. While all these techniques are well-experimented and show improvements in the state of the art in Czech- English MT directions, they are not sufficiently investigated for AMT yet or are not priority for AMT.
- The success of NMT depends heavily on the quantity and quality of the training parallel sentences for each language pair. For instance, parallel datasets for Czech-English directions are continuous released in almost each WMT campaign, but not for Arabic-English directions. In addition, long-term efforts to build dataset

such as CzEng 2.0[29] that contain 61.1M authentic sentence pairs, are missing for AMT. Accordingly, all the systems submitted to last WMT20, support Czech language.

- Romanized-based models [106], [128] reduce vocabulary sizes and UNK rates and achieve better or comparable translation results compared with their counterparts in Arabic under various segmentation scenarios. While the 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. Besides, 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. Additionally, for Romanized Arabic, there is no need to translate proper nouns even more English or French words as argued in [98].
- Although such Romanized Arabic can alleviate the limitation of BPE, especially the inconsistent segmentation of inflected words as in the case of Arabic script, however, the resulting longer sequences with potentially suboptimal subword splits may also have a negative influence on translation quality as argued in [128]. Also, for the case of Arabic dialects, where the transliteration step does not have any guidelines or laws, confusion and uncertainty may be caused. Thus, it is difficult to segment Arabic dialects from written text, particularly in cases where both Arabic script and Romanized Arabic are mixed simultaneously.
- Compared to other Western languages, the contribution of AMT is still modest and sometime behind many language pairs. Indeed, MT emerged first for English language, before such inventions are adapted to Arabic language. Subsequently, the adaptation of MT to European languages is to some extent easy due to the closeness of such languages to English in terms of language family, structure and culture. Most translations were done to and from English There were various research conducted on AMT, but this does not result in developed MT system to facilitate the process of MT for Arabic across languages.

B. OPEN CHALLENGES

Although we have witnessed the fast-growing research progresses in AMT, there are still many challenges to be addressed. Based on the extensive recent AMT reviews [29], [30] and recent works on AMT evaluations [13], [40], [41] and NMT conducted research [21], [51], [129] we summarize the major challenges in the following aspects concerning Arabic language:

- In terms of accuracy, MT is an NP-hard problem designed to provide accurate translations. The technology has improved drastically in the past ten years, but there is still a lot of work in development. Therefore, even after post-processing, the meaning of the original text still neither accurate nor fluent. Specifically, NMT still does not perform well in translating very long sentences and out-of-domain data [21].
- In terms of fluency, despite various attempts to enhance word alignment in AMT systems, still it does not fulfil the task and shows divergence. Hence, an efficient alignment algorithm is required after the translation to enhance fluency in target language such as through an induced alignment at the decoding step as demonstrated for English [130].
- In terms of performance, Transformer model has applied innovative structure in its design and therefore brought significant improvement in translation quality and speed [2]. We believe that more refinements in model structure need to be proposed. Additionally, having that RNN based-NMT takes its advantages in modeling sequence order but results in computational inefficiency compared to CNN based-NMT, more future work would consider the trade-off between these two aspects especially under the settings for Arabic language.
- In terms of computation, it is necessary to find a way to speed up training a neural network at both computation and memory levels especially for rich morphological words so that much larger vocabularies for both source and target languages, long sentences and low-frequent words can be used [21].
- Applying Transformer based pre-trained model on Arabic language, such araBERT [65], as additional embedding layer could provide a better performance as demonstrated for English [131]. Additionally, directly applying araBERT as pre-trained model could have similar performance and thus can be more convenient for encoder initialization. Indeed, designing better NMT architectures beyond Transformer shall be very promising to explore for AMT despite of the difficulty.
- Developing post-editing rules can help in handling many problems such as ordering errors, yet ambiguity remains a challenging problem [41], [80]. For instance, idiomatic expressions are hard to handle using current techniques because they are unable to differentiate the expression when it is intended for its literal or idiomatic meaning.
- OOV, rare and unknown words, ambiguous words and spelling mistaken words are hard to handle with current MT techniques. Although current solutions already exist such as entry for Unknown words (UNK) tag [132], entry in the back-off dictionary [132], subword variant BPE [22] and its extreme case, character-level [23], a standard solution is still missing. Hybrid techniques similar in [24] are increasingly adopted as a good strategy to tackle these issues such as achieving open vocabulary NMT before applying BPE. Likewise, such strategies could be investigated for Arabic as well.

[29]https://ufal.mff.cuni.cz/czeng

- NMT systems often do not incorporate any additional linguistic information since they only rely on the raw data. However, recently, the incorporation of arbitrary linguistic features as external linguistic models in NMT named as factored FNMT [133] have shown very promising results for the task of MT between several language pairs. The inclusion of arbitrary linguistic features for AMT may lead to substantial improvements. Though, an appropriate update in the encoder-decoder architecture must be developed to support the inclusion of arbitrary linguistic features of any language.
- Developing a large amount of parallel corpora is required for application of SMT and NMT models and techniques. MT process require an enormous amount of parallel text, even that is not available for specific language pairs since many languages are not rich in lexical resources such as Arabic language [79]. In this case, MNMT is a good option to investigate low-resource NMT issue for Arabic.
- Regarding out-of-domain issue where current NMT produce translations that are fluent, but unrelated to the content of the source sentence, multi-domain NMT [121] and domain-adaptation works such as Factorized Transformer [134] are attempting to resolve this limitation. However, multiple aspects need to be considered for training a multi-domain NMT model in order to prevent the model from two major issues notably overlooking the specificity of each domain [21] and forgetting previously learned knowledge when exposed to the new training examples as reported in [135].
- There are also other secondary open challenge in recent NMT research such as bias present in the training data e.g. gender bias, unsupervised MT, document-level translation, data sparsity, noisy data sources like data with misspellings, unknown words, abbreviations, punctuations, beam size, etc. which are mostly important to achieve an adequate MT. In general, we observe that while many works demonstrate or discuss the existence of bias [136], and also propose bias detection techniques, there is a shortage of works that propose de-biasing approaches for AMT.
- MT without a direct parallel data is defined as unsupervised MT. The amount of monolingual data available for training MT is an important factor for unsupervised MT, but even very challenging [137]. Back-translation, translation with retrieved similar sentences from target mono-lingual data, mining sentences from Wikipedia and use them as weakly supervised translation pairs are attempts to overcome the lack of parallel data [75].
- Similarly, most current NMT systems translates each sentence independently of other sentences since it seems too hard for sequence-to-sequence models to learn long-range document translation directly as argued by [75]. Document-level MT for Arabic still not investigated to date.
- Data sparsity is an important challenge for token-level based systems, especially for languages with rich morphology [85]. Data sparsity is challenging to obtain good word-level vector representations for rich languages such as Arabic.
- Most parallel corpora are typically gathered using web mining services e.g. Common Crawl.[30] The data is collected without enough guarantees about quality and less control about mined web sites. Thus, the data is noisy and from different domains. Since noisy training data has been recognized as a challenge for NMT training. An essential step in using such data is filtering or discounting noisy sentence pairs [127].
- In NMT, the size of all searched possible translation for each input word is termed as beam size. It is not directly proportional to the accuracy of the system as after a point it starts degrading. Optimal beam size setting does not consistently improve translation quality and may result in worse translations [21].

VI. CONCLUSION

In this paper, a review of different MT techniques along with the research challenges conducted on recent AMT is presented. This serves the developers with resources required for modelling different NMT types as well as corpora, domains, toolkits, techniques, models, features and their evaluation measures. A comparison of research works on different AMT is also performed. To-date, most of the researchers implement the SMT and NMT approaches in their work studies while the hybrid systems are gaining more attention due to their improved performance in situations where both fail to achieve a satisfactory level of accuracy and fluency [7], [46]. In spite of that, several AMT approaches need to be improved further to produce accurate and fluent translations. Although we have witnessed the fast-growing research progresses in AMT, there are still many challenges to be addressed. Based on the extensive recent AMT reviews [29], [30] and recent works on AMT evaluations [13], [40], [41] and NMT conducted research [21], [51], [129] we list some potential directions in the following aspects concerning Arabic language:

- For Arabic language, there is a need to train several Transformer NMT models and perform cross-domain testing and evaluation to gain some insight into model robustness against domain changes. The authors [138] consider domain robustness an unsolved problem and encourage further research.
- Additionally, better architectures, ensemble techniques, etc. for developing multiple models requires merging multiple alternative independent models and then combining their outputs. The ensemble technique can be applied using checkpoints averaging [4].
- Another area for future work is to extend the analysis to other word representations, deeper networks,

[30] https://commoncrawl.org/

and more semantically oriented tasks such as semantic role-labeling or semantic parsing, as suggested by [68].
- As per the analysis, NMT and hybrid approaches perform better as compared to other techniques, thus it could be considered for future use.

ACKNOWLEDGMENT

The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] N. T. Alsohybe, N. A. Dahan, and F. M. Ba-Alwi, "Machine-translation history and evolution: Survey for Arabic-English translations," *Current J. Appl. Sci. Technol.*, vol. 23, no. 4, pp. 1–19, Sep. 2017.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process.*, 2017, pp. 6000–6010.
[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[4] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtský, "Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals," *Nature Commun.*, vol. 11, no. 1, Dec. 2020, Art. no. 4381, doi: 10.1038/s41467-020-18073-9.
[5] H. Hassan *et al.*, "Achieving human parity on automatic Chinese to English news translation," 2018, *arXiv:1803.05567*.
[6] L. Alkhawaja, H. Ibrahim, F. Ghnaim, and S. Awwad, "Neural machine translation: Fine-grained evaluation of Google translate output for English-to-Arabic translation," *Int. J. English Linguistics*, vol. 43, vol. 10, no. 4, 2020, doi: 10.5539/ijel.v10n4.
[7] A. Alqudsi, N. Omar, and K. Shaker, "A hybrid rules and statistical method for Arabic to English machine translation," in *Proc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Riyadh, Saudi Arabia, May 2019, pp. 1–7, doi: 10.1109/CAIS.2019.8769545.
[8] A. Alqudsi, N. Omar, and K. Shaker, "Arabic machine translation: A survey," *Artif. Intell. Rev.*, vol. 42, no. 4, pp. 549–572, Dec. 2014.
[9] Z. M. Almahasees, "Assessment of Google and Microsoft bing translation of journalistic texts," *Int. J. Lang., Literature Linguistics*, vol. 4, no. 3, pp. 231–235, Sep. 2018.
[10] S. Hussein and S. Awab, "Evaluation of Google and Bing online translations of verb-noun collocations from English into Arabic," *J. Mod. Lang.*, vol. 25, pp. 35–59, Dec. 2016.
[11] O. O. Jabak, "Assessment of Arabic-English translation produced by Google translate," *Int. J. Linguistics, Literature Transl.*, vol. 2, no. 4, p. 10, 2019, doi: 10.32996/ijllt.2019.2.4.24.
[12] M. H. Al-khresheh and S. A. Almaaytah, "English proverbs into Arabic through machine translation," *Int. J. Appl. Linguistics English Literature*, vol. 7, no. 5, p. 158, Sep. 2018.
[13] M. El Marouani, T. Boudaa, and N. Enneya, "Statistical error analysis of machine translation: The case of Arabic," *Computación y Sistemas*, vol. 24, no. 3, pp. 1053–1061, Sep. 2020.
[14] N. Habash, *Introduction to Arabic Natural Language Processing*. San Rafael, CA, USA: Morgan & Claypool, 2010.
[15] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proc. Hum. Lang. Technol. Conf. NAACL*, New York, NY, USA, 2006, pp. 49–52.
[16] M. Oudah, A. Almahairi, and N. Habash, "The impact of preprocessing on Arabic-English statistical and neural machine translation," in *Proc. Mach. Transl. Summit XVII (MT Summit)* vol. 1. Dublin, Ireland, 2019, pp. 214–221.
[17] M. Ellouze, W. Neifar, and L. Belguith, "Word alignment applied on English-Arabic parallel corpus," in *Proc. LPKM*, 2018, pp. 1–9.
[18] M. Alkhatib and K. Shaalan, "Boosting Arabic named entity recognition transliteration with deep learning," in *Proc. 33rd Int. Florida Artif. Intell.*, FL, USA, 2020, pp. 484–488.
[19] M. E. Marouani, T. Boudaa, and N. Enneya, "Incorporation of linguistic features in machine translation evaluation of Arabic," in *Proc. BDCA*, 2018, pp. 500–511.
[20] A. Hatem and N. Omar, "Syntactic reordering for Arabic-English phrase-based machine translation," in *Proc. Int. Conf. Bio-Sci. Bio-Technol.*, in Communications in Computer and Information Science, vol. 118, 2010, pp. 198–206.
[21] P. Koehn and R. Knowles, "Six challenges for neural machine translation," 2017, *arXiv:1706.03872*.
[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1715–1725.
[23] M. R. Costa-Jussà and J. A. R. Fonollosa, "Character-based neural machine translation," 2016, *arXiv:1603.00810*.
[24] 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*, 2016, pp. 1054–1063.
[25] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Evaluating Arabic to English machine translation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 11, pp. 68–73, 2014.
[26] H. Alotaibi, "Arabic-English parallel corpus: A new resource for translation training and language teaching," *Arab World English J.*, vol. 8, no. 3, pp. 319–337, Sep. 2017.
[27] N. Durrani, F. Dalvi, H. Sajjad, and S. Vogel, "QCRI machine translation systems for IWSLT 16," 2017, *arXiv:1701.03924*.
[28] H. Mozannar, K. E. Hajal, E. Maamary, and H. Hajj, "Neural Arabic question answering," 2019, *arXiv:1906.05394*.
[29] H. M. Elsherif and T. R. Soomro, "Perspectives of Arabic machine translation," *J. Eng. Sci. Technol.*, vol. 12, no. 9, pp. 2315–2332, 2017.
[30] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic machine translation: A survey of the latest trends and challenges," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100305.
[31] M. Alkhatib and K. Shaalan, "The key challenges for Arabic machine translation," in *Intelligent Natural Language Processing: Trends and Applications*, vol. 740. Springer, 2017, pp. 139–156.
[32] M. Okpor, "Machine translation approaches: Issues and challenges," *Int. J. Comput. Sci. Issues*, vol. 11, no. 5, p. 159, 2014.
[33] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, Oct. 2017.
[34] J. Oladosu, A. Esan, I. Adeyanju, B. Adegoke, O. Olaniyan, and B. Omodunbi, "Approaches to machine translation: A review," *J. Eng. Technol.*, vol. 1, no. 1, pp. 120–126, 2016.
[35] A. Almahairi, K. Cho, N. Habash, and A. Courville, "First result on Arabic neural machine translation," 2016, *arXiv:1606.02680*.
[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, 2002, pp. 311–318.
[37] C. Boxing and K. Roland, "AMBER: A modified BLEU, enhanced ranking metric," in *Proc. 6th Workshop Stat. Mach. Transl.*, Edinburgh, U.K., 2011, pp. 71–77.
[38] A. B. Sai, A. K. Mohankumar, and M. M. Khapra, "A survey of evaluation metrics used for NLG systems," 2020, *arXiv:2008.12009*.
[39] H. Bouamor, H. Alshikhabobakr, B. Mohit, and K. Oflazer, "A human judgement corpus and a metric for Arabic MT evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 207–213.
[40] Z. AlMahasees, "Diachronic evaluation of Google translate, Microsoft translator and Sakhr in English-Arabic translation," M.S. thesis, School Humanities, Univ. Western Australia, Perth, WA, Australia, 2020.
[41] N. Madi and H. Al-Khalifa, "Error detection for Arabic text using neural sequence labeling," *Appl. Sci.*, vol. 10, no. 15, p. 5279, Jul. 2020.
[42] D. Watson, N. Zalmout, and N. Habash, "Utilizing character and word embeddings for text normalization with sequence-to-sequence models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 837–843.
[43] L. Han, "Machine translation evaluation resources and methods: A survey," 2016, *arXiv:1605.04515*.
[44] Y. Salem, A. Hensman, and B. Nolan, "Implementing Arabic-to-English machine translation using the role and reference grammar linguistic model," in *Proc. 8th Annu. Int. Conf. Inf. Technol. Telecommun. (ITT)*, Dublin, Ireland, 2008, pp. 103–110.
[45] R. A. Dam and A. Guessoum, "Building a neural network-based English-to-Arabic transfer module from an unrestricted domain," in *Proc. Int. Conf. Mach. Web Intell.*, Oct. 2010, pp. 94–101.
[46] R. Ehab, E. Amer, and M. Gadallah, "English-Arabic hybrid machine translation system using EBMT and translation memory," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 195–203, 2019, doi: 10.14569/IJACSA.2019.0100126.

[47] A. B. Phillips, V. Cavalli-Sforza, and R. D. Brown, "Improving example based machine translation through morphological generalization and adaptation," in *Proc. 9th Mach. Transl. Summit (MT Summit IX)*, 2007, pp. 369–375.

[48] S. M. Kadhem and Y. R. Nasir, "English to Arabic example-based machine translation system," *J. Comput., Commun., Control Syst. Eng.*, vol. 15, no. 3, pp. 1–17, 2015.

[49] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Comput. Linguistics*, vol. 30, no. 4, pp. 417–449, Dec. 2004.

[50] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[51] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," 2020, *arXiv:2002.07526*.

[52] Y.-S. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan, "Language model based Arabic word segmentation," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Sapporo, Japan, 2003, pp. 399–406.

[53] A. Hatem and N. Omar, "Syntactic reordering for Arabic-English phrase-based machine translation," in *Database Theory and Application, Bio-Science and Bio-Technology*. Athens, Greece: Assoc. Comput. Linguistics, 2010.

[54] A. Alrajeh, "A recipe for Arabic-English neural machine translation," 2018, *arXiv:1808.06116*.

[55] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 3104–3112.

[56] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 371–383, Dec. 2016.

[57] D. Ataman, W. Aziz, and A. Birch, "A latent morphology model for open-vocabulary neural machine translation," 2019, *arXiv:1910.13890*.

[58] F. Meng, Z. Lu, M. Wang, H. Li, W. Jiang, and Q. Liu, "Encoding source language with convolutional neural network for machine translation," 2015, *arXiv:1503.01838*.

[59] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, "A convolutional encoder model for neural machine translation," 2016, *arXiv:1611.02344*.

[60] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," 2016, *arXiv:1610.10099*.

[61] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1243–1252.

[62] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[63] F. Stahlberg, "Neural machine translation: A review," *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, Oct. 2020.

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[65] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[66] W. Lan, Y. Chen, W. Xu, and A. Ritter, "GigaBERT: Zero-shot transfer learning from English to Arabic," 2020, *arXiv:2004.14519*.

[67] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic machine transliteration using an attention-based encoder-decoder model," in *Proc. 3rd Int. Conf. Arabic Comput. Linguistics*, Dubai, UAE, 2017, pp. 1–11.

[68] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass, "What do neural machine translation models learn about morphology?" 2017, *arXiv:1704.03471*.

[69] M. Elaraby, A. Y. Tawfik, M. Khaled, H. Hassan, and A. Osama, "Gender aware spoken language translation applied to English-Arabic," in *Proc. 2nd Int. Conf. Natural Lang. Speech Process.*, Apr. 2018, pp. 1–6.

[70] E. Almansor and A. Al-Ani, "A hybrid neural machine translation technique for translating low resource languages," in *Proc. Int. Conf. Mach. Learn. Data Mining Pattern Recognit.* Cham, Switzerland: Springer, 2018, pp. 347–356.

[71] S. Ren, W. Chen, S. Liu, M. Li, M. Zhou, and S. Ma, "Triangular architecture for rare language translation," 2018, *arXiv:1805.04813*.

[72] P. Shapiro and K. Duh, "Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation," in *Proc. 6th Workshop NLP Similar Lang., Varieties Dialects*, Ann Arbor, MI, USA, 2019, pp. 214–222.

[73] A. Solyman, W. Zhenyu, T. Qian, A. A. M. Elhag, M. Toseef, and Z. Aleibeid, "Synthetic data with neural machine translation for automatic correction in Arabic grammar," *Egyptian Informat. J.*, vol. 22, no. 3, pp. 303–315, Sep. 2021, doi: 10.1016/j.eij.2020.12.001.

[74] H. Saadany and C. Orasan, "Is it great or terrible? Preserving sentiment in neural machine translation of Arabic reviews," 2020, *arXiv:2010.13814*.

[75] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020.

[76] W. Abid, "The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6030–6043.

[77] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond English-centric multilingual machine translation," 2020, *arXiv:2010.11125*.

[78] S. Ma, J. Yang, H. Huang, Z. Chi, L. Dong, D. Zhang, H. H. Awadalla, A. Muzio, A. Eriguchi, S. Singhal, X. Song, A. Menezes, and F. Wei, "XLM-T: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders," 2020, *arXiv:2012.15547*.

[79] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo, "Cross-lingual pre-training based transfer for zero-shot neural machine translation," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 115–122.

[80] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," 2020, *arXiv:2004.11867*.

[81] W. Farhan, B. Talafha, A. Abuammar, R. Jaikat, M. Al-Ayyoub, A. B. Tarakji, and A. Toma, "Unsupervised dialectal neural machine translation," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102181, doi: 10.1016/j.ipm.2019.102181.

[82] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, and L. Li, "Pre-training multilingual neural machine translation by leveraging alignment information," 2020, *arXiv:2010.03142*.

[83] E. Stergiadis, S. Kumar, F. Kovalev, and P. Levin, "Multi-domain adaptation in neural machine translation through multidimensional tagging," 2021, *arXiv:2102.10160*.

[84] S. Berrichi and A. Mazroui, "Addressing limited vocabulary and long sentences constraints in English–Arabic neural machine translation," *Arabian J. Sci. Eng.*, vol. 46, no. 9, pp. 8245–8259, Sep. 2021. [Online]. Available: https://0-doi.org.mylibrary.qu.edu.qa/10.1007/s13369-020-05328-2

[85] A. Akdemir, T. Shibuya, and T. Güngör, "Subword contextual embeddings for languages with rich morphology," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 994–1001, doi: 10.1109/ICMLA51294.2020.00161.

[86] E. Othman and M. J. A. Aziz, "English to Arabic machine translation based on reordring algorithm," *J. Comput. Sci.*, vol. 7, no. 1, pp. 120–128, Jan. 2011.

[87] 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.

[88] B. Chen, M. Cettolo, and M. Federico, "Reordering rules for phrase-based statistical machine translation," in *Proc. IWSLT*, 2006, pp. 182–189.

[89] A. Farag and A. Nürnberger, "Arabic/English word translation disambiguation using parallel corpora and matching schemes," in *Proc. 12th Eur. Mach. Transl. Conf. (EAMT)*, 2008, pp. 6–11.

[90] M. E. B. Menai and W. Alsaeedan, "Genetic algorithm for Arabic word sense disambiguation," in *Proc. 13th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, Aug. 2012, pp. 195–200.

[91] M. Hadni, S. E. Alaoui, and A. Lachkar, "Word sense disambiguation for Arabic text categorization," *Int. Arab J. Inf. Technol.*, vol. 13, no. 1, pp. 215–222, 2016.

[92] K. Shaalan, "A survey of Arabic named entity recognition and classification," *Comput. Linguistics*, vol. 40, no. 2, pp. 469–510, 2014.

[93] 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 (HLT)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 57–60.

[94] M. Diab, M. Ghoneim, and N. Habash, "Arabic diacritization in the context of statistical machine translation," in *Proc. MT-Summit*, 2007, pp. 1–7.

[95] K. Darwish, H. Mubarak, and A. Abdelali, "Arabic diacritization: Stats, rules, and hacks," in *Proc. 3rd Arabic Natural Lang. Process. Workshop*, 2017, pp. 9–17.

[96] H. Mubarak, A. Abdelali, H. Sajjad, Y. Samih, and K. Darwish, "Highly effective Arabic diacritization using sequence to sequence modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association Computational Linguistics, Jun. 2019, pp. 2390–2395.
[97] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems," *Data Brief*, vol. 11, pp. 147–151, Apr. 2017.
[98] S. Harrat, K. Meftouh, and K. Smaili, "Machine translation for Arabic dialects (survey)," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 262–273, Mar. 2019, doi: 10.1016/j.ipm.2017.08.003.
[99] M. J. Althobaiti, "Automatic Arabic dialect identification systems for written texts: A survey," 2020, *arXiv:2009.12622*.
[100] M. Popel, "CUNI English-Czech and English-polish systems in WMT20: Robust document-level training," in *Proc. 5th Conf. Mach. Transl. (WMT)*, 2020, pp. 269–273.
[101] M. A. Farajian, M. Turchi, M. Negri, N. Bertoldi, and M. Federico, "Neural vs. phrase-based machine translation in a multi-domain scenario," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 280–284.
[102] M. García-Martínez, W. Aransa, F. Bougares, and L. Barrault, "Addressing data sparsity for neural machine translation between morphologically rich languages," *Mach. Transl.*, vol. 34, no. 1, pp. 1–20, Apr. 2020, doi: 10.1007/s10590-019-09242-9.
[103] R. Alnefaie and A. M. Azmi, "Automatic minimal diacritization of Arabic texts," in *Proc. 3rd Int. Conf. Arabic Comput. Linguistics*, Dubai, UAE, 2017, pp. 169–174.
[104] F. E.-Z. El-Taher, A. A. Hammouda, and S. Abdel-Mageid, "Automation of understanding textual contents in social networks," in *Proc. Int. Conf. Sel. Topics Mobile Wireless Netw. (MoWNeT)*, Apr. 2016, pp. 1–7.
[105] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[106] F. Aqlan, X. Fan, A. Alqwbani, and A. Al-Mansoub, "Arabic–Chinese neural machine translation: Romanized Arabic as subword unit for Arabic-sourced translation," *IEEE Access*, vol. 7, pp. 133122–133135, 2019, doi: 10.1109/ACCESS.2019.2941161.
[107] S. Berrichi and A. Mazroui, "Guiding word alignment with prior knowledge to improve English-Arabic machine translation," in *Proc. 4th Int. Conf. Big Data Internet Things*. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1–5, doi: 10.1145/3372938.3372957.
[108] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The Penn Arabic treebank: Building a large-scale annotated Arabic corpus," in *Proc. NEMLAR Conf. Arabic Lang. Resour. Tools*, vol. 27, Cairo, Egypt, 2004, pp. 466–467.
[109] A. Eisele and Y. Chen, "MultiUN: A multilingual corpus from United Nation documents," in *Proc. LREC*, 2010, pp. 1–5.
[110] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The United Nations parallel corpus," in *Proc. LREC*, 2016, pp. 3530–3534.
[111] R. Parker, D. Graff, K. Chen, and M. K. Kong, "Arabic Gigaword," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep., 2011.
[112] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, Istanbul, Turkey, 2012, pp. 2214–2218.
[113] M. Cettolo, C. Girardi, and M. Federico, "WIT 3: Web inventory of transcribed and translated talks," in *Proc. 16th Conf. Eur. Assoc. Mach. Transl. (EAMT)*, Trento, Italy, 2012, pp. 261–268.
[114] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "AMARA corpus: Building parallel language resources for the educational domain," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, Reykjavik, Iceland, 2014, pp. 1044–1054.
[115] S. Wray and A. Ali, "Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.
[116] N. Habash, N. Zalmout, D. Taji, H. Hoang, and M. Alzate, "A parallel corpus for evaluating machine translation between Arabic and European languages," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Valencia, Spain, 2017, pp. 235–241.
[117] A. Fadel, I. Tuffaha, B. Al-Jawarneh, and M. Al-Ayyoub, "Neural Arabic text diacritization: State of the art results and a novel approach for machine translation," 2019, *arXiv:1911.03531*.
[118] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016, *arXiv:1606.05250*.
[119] W. Zaghouani, N. Habash, O. Obeid, B. Mohit, and K. Oflazer, "Building an Arabic machine translation post-edited corpus: Guidelines and annotation," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Portorož, Slovenia, 2016, pp. 1869–1876.
[120] S. Al-Mulla and W. Zaghouani, "Building a corpus of Qatari Arabic expressions," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools*, Marseille, France, 2020, pp. 24–31.
[121] W. Wang, Y. Tian, J. Ngiam, Y. Yang, I. Caswell, and Z. Parekh, "Learning a multi-domain curriculum for neural machine translation," 2019, *arXiv:1908.10940*.
[122] A. Al-Rukban and A. K. J. Saudagar, "Evaluation of English to Arabic machine translation systems using BLEU and GTM," in *Proc. 9th Int. Conf. Educ. Technol. Comput.* New York, NY, USA: Association for Computing Machinery, 2017, pp. 228–232.
[123] V. Mahesh and A. Milam, "A comparison of free online machine language translators," *J. Manage. Sci. Bus. Intell.*, vol. 5, no. 1, pp. 26–31, 2020.
[124] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 1317–1327.
[125] T. Kocmi and O. Bojar, "Trivial transfer learning for low-resource neural machine translation," 2018, *arXiv:1809.00357*.
[126] N.-E. Hira, S. A. Rauf, K. Kiani, A. Zafar, and R. Nawaz, "Exploring transfer learning and domain data selection for the biomedical translation," in *Proc. 4th Conf. Mach. Transl.*, vol. 3, Florence, Italy, 2019, pp. 156–163.
[127] P. Koehn, V. Chaudhary, A. El-Kishky, N. Goyal, P.-J. Chen, and F. Guzman, "Findings of the WMT 2020 shared task on parallel corpus filtering and alignment," *Proc. 5th Conf. Mach. Transl. (WMT)*, 2020, pp. 726–742.
[128] C. Amrhein and R. Sennrich, "On romanization for model transfer between scripts in neural machine translation," 2020, *arXiv:2009.14824*.
[129] R. Sennrich and B. Zhang, "Revisiting low-resource neural machine translation: A case study," 2019, *arXiv:1905.11901*.
[130] Y. Chen, Y. Liu, G. Chen, X. Jiang, and Q. Liu, "Accurate word alignment induction from neural machine translation," 2020, *arXiv:2004.14837*.
[131] S. Clinchant, K. W. Jung, and V. Nikoulina, "On the use of BERT for neural machine translation," 2019, *arXiv:1909.12744*.
[132] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," 2014, *arXiv:1410.8206*.
[133] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," 2016, *arXiv:1606.02892*.
[134] A. Bapna, N. Arivazhagan, and O. Firat, "Simple, scalable adaptation for neural machine translation," 2019, *arXiv:1909.08478*.
[135] Y. Deng, H. Yu, H. Yu, X. Duan, and W. Luo, "Factorized transformer for multi-domain neural machine translation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 4221–4230.
[136] E. Vanmassenhove, D. Shterionov, and M. Gwilliam, "Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation," 2021, *arXiv:2102.00287*.
[137] A. Fraser, "Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT," in *Proc. 5th Conf. Mach. Transl. (WMT)*, 2020, pp. 765–771.
[138] M. Müller, A. Rios, and R. Sennrich, "Domain robustness in neural machine translation," 2019, *arXiv:1911.03109*.